PACLIC 39 (2025)

Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation

Vietnam Institute for Advanced Study in Mathematics
Hanoi, Vietnam
5-7 December 2025

Emmanuele Chersoni, Jong-Bok Kim (editors)

The papers are published by the Institute for the Study of Language and Information (ISLI) at Kyung Hee University, Seoul, Korea

@2025 PACLIC 39 (2025) Organizing Committee and PACLIC Steering Committee
All rights reserved. Except as otherwise expressly permitted under copyright law, no part of this publication may be reproduced, digitized, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or recording without the permission of the publisher. Copyright of contributed papers reserved by respective authors.

ISBN 979-8-89176-357-9

Acknowledgement
PACLIC 39 is hosted by the Vietnam Institute for Advanced Study in Mathematics, Hanoi, Vietnam.
The conference is held under the auspices of the PACLIC Steering Committee.

Foreword

Welcome to PACLIC 39! The PACLIC conference emphasizes the synergy of theoretical frameworks and processing of natural language, providing a forum for researchers from different fields to share and discuss progress in scientific studies, development and application of the topics related to the study of languages. For the 2025 edition, PACLIC has been held in Vietnam, in the beautiful capital city of Hanoi. After the 2020 edition had to be moved online because of the pandemic, we are deeply grateful to have now the chance to meet in person with the vibrant research community of Vietnamese linguistics and language technologies.

For this edition, we received a total of 158 submissions and we accepted 76 of them for publication (a 48.1% acceptance rate). In parallel with the main conference, we had a satellite event: the workshop of Language and Food in Asia, for which 6 abstract were selected for presentation.

We would like to thank the keynotes, Marco Marelli (University of Milan-Bicocca), Mark Liberman (University of Pennsylvania), Derek Wong (University of Macau) and Shalom Lappin (University of Gothenburg) who graciously accepted our invitation and enriched the conference program with their contribution. A special thanks goes to Chu-Ren Huang: without his vision, this conference series would not be the same, and thus we consider his invited talk as a symbolic culmination of almost forty years of commitment to the PACLIC community.

We would also like to thank all the members of the Program Committee for reviewing the submissions and providing constructive feedback to the authors. Finally, we thank the Vietnam Institute for Advanced Study in Mathematics, for hosting and sponsoring this edition of the conference.

Emmanuele Chersoni
Le Minh Nguyen
Rachel E. O. Roxas
Shirley N. Dita
39th PACLIC Program Committee Chairs
(on behalf of the Organizing Committee)

Organizing Committee

Honorary conference chairs
Chu-Ren Huang (The Hong Kong Polytechnic University, China)
Yasunari Harada (Waseda University, Japan)

General chairs
Jong-Bok Kim (Kyung Hee University, South Korea)
Nguyen T.M. Huyen (VNU University of Science, Vietnam)

Program co-chairs
Emmanuele Chersoni (The Hong Kong Polytechnic University)
Le Minh Nguyen (Japan Advanced Institute of Science and Technology, Japan)
Rachel Edita Oñate Roxas (University of the Philippines Los Baños, Philippines)
Shirley N. Dita (De La Salle University, Philippines)

Program Committee

Kathleen Ahrens, Taichi Aida, Xiaoyi Bao, Philippe Blache, Alessandro Bondielli, Auriane Boudin, James Britton, Dominique Brunato, Lia Calinescu, Maria Cassese, Adel Chaouch-Orozco, Jing Chen, Cristiano Chesi, Won Ik Cho, Marco Ciapparelli, Yan Cong, Max Diaz, Luca Dini, Yo Ehara, Zhaoxin Feng, Irene Fioravanti, Maiwenn Fleig, Michael Flor, Zhe Gao, Jinghang Gu, Linh Ha My, Shu-Kai Hsieh, Hai Hu Bernard Jap, Elisabetta Jezek, Maria Dolores Jimenez-Lopez, Ian Joo, Anisia Katinskaia, Jong-Bok Kim, Iksoo Kwon, Ryan Ka Yau Lai, Chaak-Ming Lau, Huong Thanh Le, Anh Cuong Le, Hoang-Quynh Le, Phuong Le-Hong, Yuxi Li, Ke Liang, Ming Liu, Hongchao Liu, Lu Lu, Jianfei Ma, Matthew Ma, Raphael Merx, Alessio Miaschi, Timothee Mickus, Martina Miliani, Mohammad Momenian, Natchanan Natpratan, The Quyen Ngo, Thi Minh Huyen Nguyen, Le-Minh Nguyen, Kiet Van Nguyen, Minh-Tien Nguyen, Ha-Thanh Nguyen, Ngan Nguyen, Ha Nguyen Tien, Stefano Occhipinti, Soyoung Oh, Byung-Doh Oh, Makoto Okada, Masanori Oya, Irene Pagliai, Alexander Panchenko, Ludovica Pannitto, Gabor Parti, Andrea Pedrotti, Eric Pelzl, Bo Peng, Yingying Peng, Massimo Piccardi, Beatrice Portelli, Pranav A, Mattia Proietti, Laurent Prévot, Giovanni Puccetti, Le Qiu, Lisa Raithel, Giulia Rambelli, Lavinia Salicchi, Teeranoot Siriwittayakorn, Nguyen Truong Son, Huacheng Song, Francesca Strik Lievers, Qi Su, Kazuhiro Takeuchi, Xuemei Tang, Ran Tao, Marina Tiuleneva, Vu Tran, Vinh Van Nguyen, Mingyu Wan, Wenbo Wang, Vincent Wang, Yu Wang, Nizhuan Wang, Bruce Xiao Wang, Yu Wang, Xueyi Wen, Yuhan Xia, Qihui Xu, Han Xu, Yiheng Yang, Guanqun Yang, Shaoyun Yu, Winnie Huiheng Zeng, Yu Zhai, Kaile Zhang, Xiaojing Zhao, Beth Zhong, He Zhou, Michael Zock.

Improving Interpretability of Lexical Semantic Change with Neurobiological Features

Kohei Oda[1] **Hiroya Takamura**[2] **Kiyoaki Shirai**[1] **Natthawut Kertkeidkachorn**[1]
[1]Japan Advanced Institute of Science and Technology
[2]National Institute of Advanced Industrial Science and Technology
[1]{s2420017,kshirai,natt}@jaist.ac.jp
[2]takamura.hiroya@aist.go.jp

Abstract

Lexical Semantic Change (LSC) is the phenomenon in which the meaning of a word change over time. Most studies on LSC focus on improving the performance of estimating the degree of LSC, however, it is often difficult to interpret how the meaning of a word change. Enhancing the interpretability of LSC is a significant challenge as it could lead to novel insights in this field. To tackle this challenge, we propose a method to map the semantic space of contextualized embeddings of words obtained by a pre-trained language model to a neurobiological feature space. In the neurobiological feature space, each dimension corresponds to a primitive feature of words, and its value represents the intensity of that feature. This enables humans to interpret LSC systematically. When employed for the estimation of the degree of LSC, our method demonstrates superior performance in comparison to the majority of the previous methods. In addition, given the high interpretability of the proposed method, several analyses on LSC are carried out. The results demonstrate that our method not only discovers interesting types of LSC that have been overlooked in previous studies but also effectively searches for words with specific types of LSC.[1]

1 Introduction

The meanings of words change over time. For example, according to the Oxford English Dictionary (OED)[2], the word *gay* acquired the meaning of *homosexual* around 1934, in addition to its earlier meaning of *cheerful*. This phenomenon is called Lexical Semantic Change (LSC) and actively studied in recent years (Tahmasebi et al., 2019, 2021; Periti and Montanelli, 2024). Many studies in this field represent the meanings of words as vectors using embedding models, such as static word embeddings (Mikolov et al., 2013) and BERT (Devlin et al., 2019), and learn separate spaces for different time periods (Kim et al., 2014; Hamilton et al., 2016; Bamler and Mandt, 2017) or handle multiple time periods within the same space (Hu et al., 2019; Giulianelli et al., 2020; Martinc et al., 2020b). While these techniques are useful for estimating the degree of LSC, they are inappropriate for humans to interpret LSC.

Several methods have been proposed to improve the interpretability of LSC, including a method presenting neighboring words in a vector space (Gonen et al., 2020), obtaining representative co-occurrence words (Montariol et al., 2021), predicting substitutions (Card, 2023), assigning predefined word senses (Tang et al., 2023), and generating definition sentences of word meanings (Giulianelli et al., 2023; Fedorova et al., 2024). These methods help humans interpret LSC through natural language, e.g., by showing indicative words or definition sentences. However, explanations of LSC based on words and sentences are ambiguous and lack a systematic explanatory framework.

Motivated by the above, we propose a method to improve the interpretability of LSC by using neurobiological features proposed by Binder et al. (2016), which we call *Binder features* in this paper. There are 65 Binder features such as *Vision*, *Audition*, and *Happy*. The values of these 65 Binder features have been estimated for 535 English words and are open to the public.[3] Based on previous studies (Utsumi, 2018, 2020; Turton et al., 2021), we use the public dataset above to train a regression model that maps the BERT semantic space to the Binder space for the quantitative and multi-perspective interpretation of LSC.

First, the potential of the Binder features in an-

[1]Our code is available at: https://github.com/iehok/LSC_with_Binder.
[2]https://www.oed.com/
[3]https://www.neuro.mcw.edu/index.php/resources/brain-based-semantic-representations/

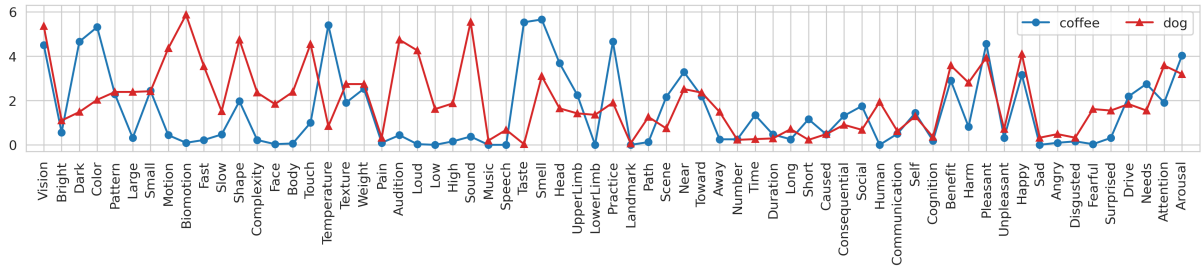


Figure 1: Binder feature values for "coffee" and "dog"

alyzing LSC is evaluated by applying our method to a task aimed at estimating the degree of LSC. Second, utilizing the high interpretability of our method, we analyze types of LSC. The integration of our method with Sparse PCA (Principal Component Analysis) enables us to identify interesting types of LSC that have not been found in previous studies. Finally, we apply our method to detect amelioration and pejoration (Traugott, 2017), and successfully identify ameliorative and pejorative words in a real corpus.

The contributions of our paper are summarized as follows:

- We introduce neurobiological features into the field of lexical semantic change, thereby improving the interpretability.
- We discover several interesting types of LSC that have not been noted in previous studies by combining our method with Sparse PCA.
- We propose a method that can easily detect specific types of LSC, amelioration and pejoration, using our approach.

2 Related Work

2.1 Lexical Semantic Change

LSC is mainly studied in the fields of linguistics and natural language processing (NLP). Even when being constrained to NLP, numerous methods are proposed such as a method that utilizes mutual information (Gulordava and Baroni, 2011; Hamilton et al., 2016; Schlechtweg et al., 2019), Bayesian models (Emms and Jayapal, 2016; Frermann and Lapata, 2016; Inoue et al., 2022), and static word embeddings (Kulkarni et al., 2015; Takamura et al., 2017; Del Tredici et al., 2019).

Recently, with the emergence of pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) that can generate representations of the meanings of words in a context, methods using these models have been actively studied (Kutuzov and Giulianelli, 2020; Martinc et al., 2020a; Liu et al., 2021b). Hu et al. (2019) propose a method to identify how the meaning of a word changes by calculating the distribution of word senses over time, where example sentences in the OED are used to assign senses to words in the corpus. Giulianelli et al. (2020) propose a method to calculate the distribution of usage types (pseudo senses) without using a dictionary. This method uses k-means clustering on a set of contextualized embeddings from all time periods to assign usage types to words in the corpus. Additionally, the degree of LSC between two different time periods is estimated using either the Jensen-Shannon divergence (JSD) between usage type distributions or the average pairwise distance (APD) between sets of contextualized embeddings from these time periods. In this study, we model LSC based on Giulianelli et al. (2020) coupled with the Binder features to improve the interpretability.

2.2 Interpretable Word Embeddings

Interpretable representations, i.e., methods of assigning roles (interpretations) to each dimension of an embedding, have been extensively studied (Panigrahi et al., 2019; Şenel et al., 2020; Aloui et al., 2020). However, these methods often suffer from a lack of clarity of a role for each dimension or coarse granularity of roles.

Binder et al. (2016) propose interpretable word vectors by defining 65 features based on neurobiological perspectives and manually assign these strengths (0 to 6) to 535 words, including 434 nouns, 62 verbs, and 39 adjectives. Figure 1 shows the Binder features and their corresponding values for "coffee" and "dog." The values of *Taste* and *Smell* are relatively high for "coffee," while the values of *Biomotion* and *Sound* are high for "dog." Additionally, the *Vision* feature has high values for both words.

Binder features are actively studied in the fields of cognitive linguistics and NLP. Utsumi (2020); Chersoni et al. (2021); Flechas Manrique et al. (2023) investigate what kind of information is encoded in static word embeddings, such as SGNS (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), by mapping these word embedding spaces to the Binder feature space. Turton et al. (2020) assign the Binder values to words other than the original 535 words by the aforementioned mapping from word embeddings. Turton et al. (2021) demonstrate that contextualized word embeddings generated from Transformer (Vaswani et al., 2017) based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) can derive the Binder values in the same way as the mapping of static word embeddings.

3 Mapping BERT Space to Binder Space

To enhance the interpretability of LSC, we first convert the semantic space of contextualized word embeddings derived from BERT to that of the Binder features. Specifically, a regression model is trained, which maps the BERT space (768 dimensions) to the Binder space (65 dimensions). The regression model, designated as ψ, is formalized as follows,

$$\mathbf{b}_w = \psi(\mathbf{r}_w), \tag{1}$$

where $\mathbf{r}_w$ and $\mathbf{b}_w$ are BERT and Binder vectors, respectively.

3.1 Word Embeddings on BERT Space

Let $\mathcal{C}$ be the corpus used for training, and let $\mathcal{C}_w$ be the set of (s, i), a pair of a sentence s in $\mathcal{C}$ that contains the target word w and its position i in s. The representation of w in the entire $\mathcal{C}$ is defined as follows.

$$\mathbf{r}_w = \frac{1}{|\mathcal{C}_w|} \sum_{(s,i)\in\mathcal{C}_w} \phi(s, i). \tag{2}$$

The function $\phi(s, i)$ denotes the hidden state of the final layer for the i-th token of the BERT model when s is given as an input. In this study, `bert-base-uncased`[4] is used as the BERT model.

The Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) is used as $\mathcal{C}$. CCOHA is an English corpus covering the period from 1820 to 2020, divided into ten-year segments. It consists of five genres: TV/Movies, Fiction, Magazine, Newspaper, and Non-fiction.

[4] https://huggingface.co/google-bert/bert-base-uncased

	LT	MLP
1910-2010	.571	.645
1960-2010	**.569**	.689

Table 1: Average MSE for 10 trials

3.2 Training of Regression Model

Two architectures of the regression model are applied: a simple linear transformation (LT) and a multilayer perceptron (MLP). The MLP consists of four hidden layers (300, 200, 100, 50 dimensions), following Turton et al. (2021). The output of each layer is activated by ReLU. To match the scale of the Binder value, in both the LT and the MLP, the final output is activated by Sigmoid and subsequently multiplied by 6 to convert values within the range of 0 to 6. The regression models are trained using 535 words associated with the Binder feature vectors (Binder et al., 2016). The loss function is set to the mean squared error (MSE) between predicted and ground-truth values of all Binder features of the target words.

3.3 Settings

We conduct experiments using two different periods of CCOHA: 1910-2010 and 1960-2010. The period 1910-2010 follows the setting in Giulianelli et al. (2020), while the period 1960-2010 is set to the most recent half of it, as the Binder dataset (Binder et al., 2016) was created in 2016. The performance of the trained regression model is evaluated using k-fold cross-validation, where k is set to 10. The batch size, the learning rate, and the number of epochs are set to 16, 1e-3 and 100, respectively. The quality of the regression model is evaluated by the MSE on the test set. The MSE is measured at each epoch, and the minimum MSE is recorded.

3.4 Results

Table 1 shows the average MSE for 10 trials of the cross-validation. LT significantly outperforms MLP, while the time period of the training corpus has a minimal influence on the results. This may be because the words in the Binder dataset are well-known and common, which leads to a relatively stable representation over time. This finding partially agrees with the results obtained by Hamilton et al. (2016).

4 Lexical Semantic Change Detection

This section proposes and evaluates a method to detect LSC using the Binder feature vectors.

4.1 Task Definition

SemEval-2020 Task 1 (Subtask 2) (Schlechtweg et al., 2020) is a task that aims to predict the degree of LSC of a word w. Specifically, the goal is to predict an LSC score representing how drastically the meaning of w changes between $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$, which are corpora of two different periods t_1 and t_2. The dataset consists of 37 English target words with manually assigned LSC scores. $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$ are parts of CCOHA from 1810 to 1860 and 1960 to 2010, respectively. Evaluation is performed by measuring Spearman's rank correlation coefficient between the predicted and ground-truth LSC scores.

4.2 Predicting Degree of LSC

To predict the degree of LSC of a word w from t_1 to t_2, the set of contextualized embeddings of w in the corpus $\mathcal{C}^t_w$ is calculated for each time period:

$$\mathcal{U}^t_w = \bigcup_{(s,i)\in\mathcal{C}^t_w} \{\, \psi(\phi(s,i)) \,\}, \tag{3}$$

where ψ is the regression model, either LT or MLP explained in Section 3. Then, following Giulianelli et al. (2020), the degree of LSC between $\mathcal{U}^{t_1}_w$ and $\mathcal{U}^{t_2}_w$ is measured by the average pairwise distance (APD):

$$\mathrm{APD}(\mathcal{U}^{t_1}_w, \mathcal{U}^{t_2}_w) = \frac{1}{|\mathcal{U}^{t_1}_w| \cdot |\mathcal{U}^{t_2}_w|} \sum_{\mathbf{u}_i \in \mathcal{U}^{t_1}_w} \sum_{\mathbf{u}_j \in \mathcal{U}^{t_2}_w} d(\mathbf{u}_i, \mathbf{u}_j), \tag{4}$$

where d is a distance function. We compared the following three distance functions: Euclidean distance, cosine distance, and Spearman distance. Spearman distance is defined as $(1 - \mathrm{sc}(\mathbf{u}_i, \mathbf{u}_j))$, where $\mathrm{sc}(\mathbf{u}_i, \mathbf{u}_j)$ is Spearman's rank correlation coefficient between sets of values of dimensions in two vectors.

4.3 Results

Table 2 shows a comparison between the baseline BERT space and our methods based on different regression models. The performance of LSC detection is slightly improved by mapping to the Binder space using the linear regression model. The architecture of the regression model significantly impacts the performance of LSC detection, while the period of the corpus used for training the regression model has a relatively small impact; this tendency is similar to that in Table 1. Among the three distance functions, the cosine distance is relatively stable and performs well.

Model	Euclid	Cosine	Spearman
BERT space	.616	.645	.618
LT-1910-2010	.633	.644	**.647**
LT-1960-2010	**.635**	**.667**	.634
MLP-1910-2010	.499	.587	.562
MLP-1960-2010	.483	.442	.540

Table 2: Spearman's rank correlation coefficient for SemEval-2020 Task 1. "BERT space" means a method that calculates APD in the BERT space without mapping it to the Binder space.

Model	EK	Score
SSCS (Tang et al., 2023)	✓	.589
XL-LEXEME (Cassotti et al., 2023)	✓	.757
SDML (Aida and Bollegala, 2024)	✓	.774
NLPCR (Rother et al., 2020)		.436
APD (Laicher et al., 2021)		.571
ScaledJSD (Card, 2023)		.547
SSCD (Aida and Bollegala, 2023)		.383
LT-1960-2010 (ours)		**.667**

Table 3: Spearman's rank correlation coefficient for previous methods on SemEval-2020 Task 1. "EK" (external knowledge) means methods that are fine-tuned with WiC corpora (Raganato et al., 2020; Martelli et al., 2021; Liu et al., 2021a) or methods using the information of dictionaries such as WordNet (Miller, 1994) and BabelNet (Navigli and Ponzetto, 2010).

Table 3 shows a comparison of our method with other existing methods. Although our method is simple, it achieves the best performance compared to other methods that do not use any external knowledge.

5 Analysis of LSC Types

This section describes an analysis of LSC types using our method. Given the high interpretability of neurobiological features, our goal is to identify the types of semantic changes of words between two different periods t_1 and t_2.

5.1 Target Words

The target words used for this analysis are collected from WordNet (Miller, 1994) lemmas meeting the following three conditions: (i) included in the vocabulary of the BERT tokenizer, (ii) having two

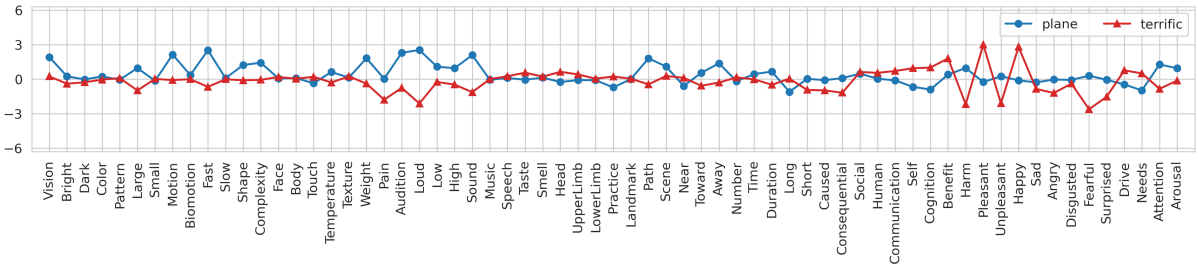


Figure 2: LSC vectors for *plane* and *terrific*

or more senses in WordNet, and (iii) four or more characters long. Condition (i) is set because our method is incapable of handling words that are divided into subwords. Condition (ii) is set because words whose meanings have changed are likely to have newly added senses, thereby resulting in polysemy. Condition (iii) is set because short words are more likely to become subwords within other words. Based on these three conditions, a total of 8,570 target words are chosen.

5.2 Corpora

The CCOHA 1910s (from 1910 to 1920) and 2000s (from 2000 to 2010) are used as the corpora $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$. This corresponds to the first and last decades of the period from 1910 to 2010 used for training the regression model (Section 3).

5.3 Methods

To analyze how the meaning of words changed between two periods t_1 and t_2, the LSC vector of the word w, denoted as $\mathbf{v}_{\mathrm{lsc}}(w)$, is computed as follows:

$$\mathbf{v}_{\mathrm{lsc}}(w) = \frac{1}{|\mathcal{U}_w^{t_2}|} \sum_{\mathbf{u}_i \in \mathcal{U}_w^{t_2}} \mathbf{u}_i - \frac{1}{|\mathcal{U}_w^{t_1}|} \sum_{\mathbf{u}_i \in \mathcal{U}_w^{t_1}} \mathbf{u}_i. \quad (5)$$

This vector represents the semantic changes of all Binder features. A positive value in a dimension of the LSC vector means that the meaning of the corresponding Binder feature is newly acquired from t_1 to t_2, while a negative value implies a loss of the meaning.

After calculating the LSC vectors for all target words, Sparse PCA is applied to the LSC vectors of the 500 target words with the largest norms, supposing that the meanings of words with small norms are not significantly changed. Unlike conventional PCA, Sparse PCA enhances interpretability by setting many elements in the eigenvectors to zero, and the eigenvectors do not need to be orthogonal to each other. Since the number of principal components should be predetermined, it is set to 10 in this experiment. The analysis of different numbers of principal components remains a subject for future work.

It is hypothesized that each principal component (PC) of Sparse PCA represents a type of LSC. For each PC, the top three Binder features with the highest values in the eigenvector are extracted to provide a clear interpretation of the LSC type. Subsequently, we check the words in descending or ascending order of their values of the principal component and verify whether they are representative words. The validity of the chosen representative words is evaluated by the following procedures. First, following Giulianelli et al. (2020), usage types (pseudo senses) are assigned to the target words in example sentences by conducting k-means clustering on a set of contextualized embeddings. Second, the five examples closest to the center of each cluster are examined to confirm whether they are correctly divided according to their meanings. Finally, the change in the distribution of usage types from t_1 to t_2 is checked to investigate whether it supports the LSC type augmented by the related Binder features.

This method is similar to the analysis by applying PCA in BERT space (Aida and Bollegala, 2025), but enhances the interpretability of LSC types. First, not only words with large or small principal components but also the values of eigenvectors can be used for analyzing LSC types. Second, since Sparse PCA assigns a zero to many elements, it is easier to find relevant (non-zero) Binder features for each LSC type.

5.4 Results

Figure 2 shows the LSC vectors for *plane* and *terrific*. According to the OED, the word *plane* acquired the meaning of *airplane* around 1908, in addition to its existing meaning of *a flat geometri-*

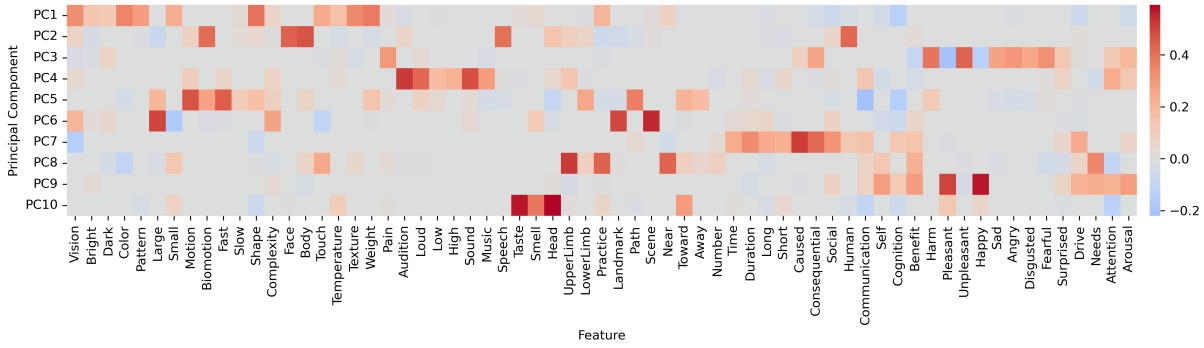


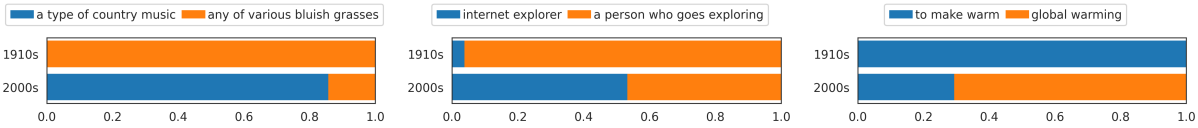
Figure 3: Eigenvectors obtained by Sparse PCA. The number of principal components is set to 10. The horizontal and vertical axes represent the 65 Binder features and the 10 principal components, respectively.

PC	LSC Type Label	Top 3 Binder Features	Representative Words
1	Artifact	Shape, Weight, Color	↑ console, plastic, vogue ↓ overall, album, bluegrass
2	Human	Body, Face, Human	↑ coach, shooter, racer ↓ yahoo, explorer, major
3	Negative Meaning	Unpleasant, Harm, Fearful	↑ serial, aids, parkinson ↓ offence, terrific, crook
4	Sound	Audition, Sound, Loud	↑ bluegrass, plane, blues ↓ instrumentation, click, booming
5	Transportation	Motion, Fast, Path	↑ pickup, sedan, plane ↓ steamed, omnibus, coach
6	Place	Scene, Large, Landmark	↑ facility, resort, berkeley ↓ manila, chihuahua, bologna
7	Social	Caused, Consequential, Duration	↑ warming, launch, summit ↓ briefs, console, offensive
8	Familiar Thing	UpperLimb, Practice, Near	↑ topical, sink, blackberry ↓ album, shooter, warming
9	Positive Meaning	Happy, Pleasant, Benefit	↑ bonding, outgoing, terrific ↓ intelligence, utility, console
10	Food	Head, Taste, Smell	↑ bologna, bourbon, steamed ↓ alcoholic, blackberry, bluegrass

Table 4: The result of analysis of Sparse PCA. "LSC Type Label" is a manually assigned label for the LSC type. The symbols ↑ and ↓ indicate words with relatively large and small principal component values, respectively, suggesting the words have acquired or lost their meanings of the features.

cal surface. As illustrated in Figure 2, the values of the Binder features such as *Motion*, *Audition*, and *Path* exhibit a substantial increase. Additionally, according to the OED, *terrific* acquired the meaning of *amazing* around 1871, in addition to its existing meaning of *causing terror*. The values of the Binder features *Pleasant* and *Happy* have increased significantly, while the values of the Binder features *Harm*, *Unpleasant*, and *Fearful* have decreased significantly. This indicates that the major meaning of *terrific* has shifted from a negative to a positive meaning.

Figure 3 shows the eigenvectors obtained by Sparse PCA. Many elements in the eigenvectors are zero, making them relatively easy to interpret. In addition, by examining the absolute values in the eigenvectors, it is possible to identify Binder features that are deeply related to or not related to LSC. For example, the absolute values of *Vision* at the first, sixth, and seventh PCs are relatively high, indicating *Vision* is likely to be deeply related to LSC. On the other hand, the absolute values of *Number* are nearly zero in all PCs, suggesting that *Number* does not contribute to LSC.



(a) bluegrass (b) explorer (c) warming

Figure 4: Distributions of the usage types for *bluegrass*, *explorer*, and *warming*

Table 4 shows the types of LSC corresponding to PCs. Many interesting types of LSC, which have not been noted in previous studies (Traugott, 2017; Campbell, 2020), are discovered by our method. Figure 4 shows the distributions of the usage types for some illustrative examples of words: *bluegrass*, *explorer*, and *warming*. For *bluegrass* in PC4, the meaning has changed from *any of various bluish grasses* to *a type of country music*, shifting to a meaning related to sounds. For *explorer* in PC2, the meaning related to humans has declined due to its increased use in the collocation *internet explorer*. For *warming* in PC7, the meaning has changed to a meaning related to social due to its increased use in the collocation *global warming*. The distributions for other words are shown in the Appendix A.

6 Analysis of Amelioration and Pejoration

The process of mapping the BERT space to the Binder space not only improves the interpretability of LSC, as described in Section 5, but also facilitates the search for words corresponding to specific types of LSC. This section presents a case study to search for words that went through amelioration or pejoration, where amelioration means acquiring positive sentiment and pejoration means acquiring negative sentiment (Traugott, 2017). In addition, we evaluate the ability of our method to identify specific words that acquire a positive or negative meaning over time.

6.1 Known Words of Amelioration and Pejoration

Several pieces of literature have already reported examples of amelioration and pejoration. From these references, the sets of known words of amelioration and pejoration, $\mathcal{W}_{\text{ame}}$ and $\mathcal{W}_{\text{pej}}$ respectively, are extracted. Table 5 shows $\mathcal{W}_{\text{ame}}$ and $\mathcal{W}_{\text{pej}}$ with their references. Although both sets are small, they are used as ground truth to examine whether our method successfully identifies these words as amelioration or pejoration.

$\mathcal{W}_{\text{ame}}$	**hysteria** (Cook and Stevenson, 2010)
	brilliant, **fabulous**, **fantastic**, **spectacular** (Altakhaineh, 2018)
	terrific (de Wit, 2021)
$\mathcal{W}_{\text{pej}}$	**dynamic**, **synthesis** (Cook and Stevenson, 2010)
	abuse, **addiction**, **harassment**, **prejudice**, **trauma** (Haslam, 2016)
	terrible (Altakhaineh, 2018)
	awful (de Wit, 2021)

Table 5: Sets of known words of amelioration $\mathcal{W}_{\text{ame}}$ and pejoration $\mathcal{W}_{\text{pej}}$

6.2 Methods

First, we select the Binder features that are related to positive or negative meanings. Referring to Binder et al. (2016), the features related to positive meanings $\mathcal{I}_{\text{pos}}$ are defined as *Pleasant* and *Happy*, while the features related to negative meanings $\mathcal{I}_{\text{neg}}$ are defined as *Pain*, *Harm*, *Unpleasant*, *Sad*, *Angry*, *Disgusted*, and *Fearful*. Indeed, some of these features indicate that the LSC type of PC9 and PC3 in Table 4 are amelioration and pejoration, respectively. Furthermore, *Happy*, *Sad*, *Angry*, *Disgusted*, and *Fearful* are derived from the basic emotions proposed by Ekman (1992), which are closely related to emotion analysis (Plaza-del Arco et al., 2024) in the field of NLP.

Next, for each target word collected in Section 5.1, a score indicating the degree of positive or negative lexical semantic change (called LSC score in this paper) is calculated as follows:

$$\text{LSCS}(w, x) = \max_{i \in \mathcal{I}_x} \mathbf{v}_{\text{lsc}}(w)[i], \qquad (6)$$

where $\mathcal{I}_x$ is either $\mathcal{I}_{\text{pos}}$ or $\mathcal{I}_{\text{neg}}$. That is, the maximum value of the positive (or negative) features in the LSC vector is employed as the LSC score. Our motivation behind this definition is that a word should be recognized as amelioration or pejoration if one of the features in $\mathcal{I}_{\text{pos}}$ or $\mathcal{I}_{\text{neg}}$ increases significantly.

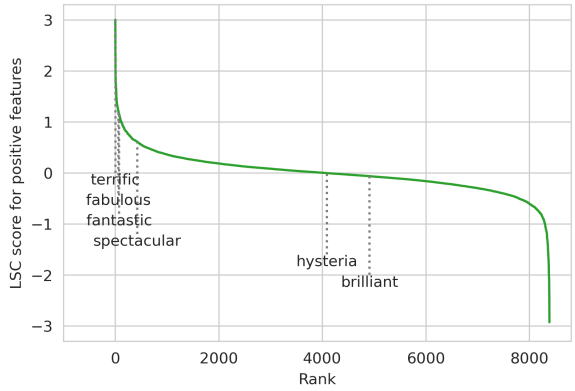


Figure 5: LSC scores for positive features

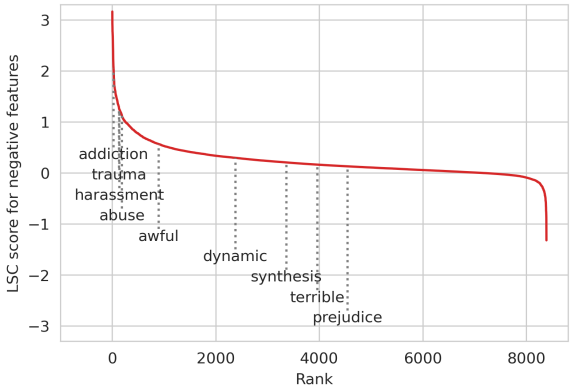


Figure 6: LSC scores for negative features

Finally, we sort all the target words in order of their LSC scores and verify whether the words in $\mathcal{W}_{\text{ame}}$ or $\mathcal{W}_{\text{pej}}$ are highly ranked.

While previous methods (Cook and Stevenson, 2010; Goworek and Dubossarsky, 2024) are specialized for detecting amelioration and pejoration, our approach can extend to identify words of other LSC types discovered in Section 5.

6.3 Results

Figures 5 and 6 show the LSC scores for positive features $\text{LSCS}(w, \text{pos})$ and negative features $\text{LSCS}(w, \text{neg})$, respectively. Words changing in a positive direction (i.e., the LSC score is greater than zero) account for about half of the total, while words changing in a negative direction account for about 75%. This indicates that words tend to change in a negative direction more than in a positive direction.

Figure 5 shows that the rank of most known words of amelioration in Table 5 are relatively high. In particular, *terrific* is ranked first. The OED and de Wit (2021) denote that *terrific* began to be used with a positive meaning in addition to a negative one in the late 19th century, and today it is mainly used with a positive meaning. On the other hand, the LSC scores for *hysteria* and *brilliant* are nearly zero. For *hysteria*, no positive meaning similar to those shown by Cook and Stevenson (2010) are found in the OED and examples in the CCOHA. For *brilliant*, according to the OED, this word originally meant *shining* and acquired the metaphorical meaning of *splendid* around 1739. This semantic shift was not captured because the LSC score is measured between periods of the 1910s and 2000s.

Figure 6 indicates that the meaning of all the known words of pejoration in Table 5 are shifted in a negative direction. The words *abuse*, *addiction*, *harassment*, and *trauma*, which are suggested by Haslam (2016), are ranked relatively high. According to Haslam (2016), as the meanings of these words expand, people become more sensitive to their negative connotations. On the other hand, the ranks of some words in $\mathcal{W}_{\text{pej}}$ are low. For *prejudice*, the results are similar to those of Vylomova et al. (2019), and unlike other words in Haslam (2016), its meaning has not drastically shifted in a negative direction. For *dynamic* and *synthesis*, no negative meaning similar to those shown by Cook and Stevenson (2010) is found in the OED and examples in the CCOHA. For *terrible*, since this word has only negative meaning, it is unlikely that its meaning will change in a more negative direction.

To sum up, these results demonstrate the effectiveness of our method in the detection of amelioration and pejoration.

7 Conclusion

This study proposed a novel method to improve the interpretability of LSC by mapping the semantic space of the pre-trained language model to the neurobiological space. In the experiments designed to estimate the degree of LSC, our method demonstrated better performance than the baseline methods that did not map the semantic spaces. By leveraging the high interpretability of our method, we discovered interesting types of LSC that had not been identified previously. Additionally, in the detection of amelioration and pejoration, our method assigned appropriate LSC scores for words, which evaluated how their meanings changed positively or negatively. In the future, we plan to apply our method to detect words of other types of LSC.

Limitations

In this study, we analyzed several LSC types from the perspective of the Binder features. On the other hand, according to Traugott (2017), there are different types of LSC, such as metaphorization, metonymization, narrowing, and generalization. The method proposed in this paper might struggle to capture these LSC types because there is no clear correlation between the Binder features and them. Therefore, it is necessary to extend the current method or adopt new methods of representation (e.g., representing the meaning of a word in a sentence with box embeddings (Oda et al., 2024)).

In addition, it is necessary to increase the number of target words. In our method, words that are not included in the vocabulary of the tokenizer of pre-trained language models are outside the scope of the analysis, resulting in failure to capture the LSC of those words. Even when a word is split into multiple subwords, contextualized embeddings should be obtained, for example, by taking an average vector of the contextualized embeddings of these subwords (Montariol et al., 2021).

References

Taichi Aida and Danushka Bollegala. 2023. Swap and predict – predicting the semantic changes in words across corpora by context swapping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7753–7772, Singapore. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2024. A semantic distance metric learning approach for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7570–7584, Bangkok, Thailand. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2025. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1413–1437, Abu Dhabi, UAE. Association for Computational Linguistics.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Cindy Aloui, Carlos Ramisch, Alexis Nasr, and Lucie Barque. 2020. SLICE: Supersense-based lightweight interpretable contextual embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abdel Rahman Mitib Altakhaineh. 2018. The semantic change of positive vs. negative adjectives in modern english. *Lingua Posnaniensis*, 60(2):25–38.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 380–389. JMLR.org.

Jeffrey R. Binder, Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai and. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4):130–174. PMID: 27310469.

Lyle Campbell. 2020. *Historical Linguistics*. Edinburgh University Press, Edinburgh.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ilse de Wit. 2021. *A Terrific Paper: A Corpus Study of Amelioration and Pejoration in Adjectives Related to Fear*. Ph.D. thesis, Stockholm University, Faculty of Humanities, Department of English.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Martin Emms and Arun Kumar Jayapal. 2016. Dynamic generative model for diachronic sense emergence detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1362–1373, Osaka, Japan. The COLING 2016 Organizing Committee.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing interpretability using human similarity judgements to prune word embeddings. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 169–179, Singapore. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Roksana Goworek and Haim Dubossarsky. 2024. Toward sentiment aware semantic change analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 350–357, St. Julian's, Malta. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Nick Haslam. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27(1):1–17.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2022. Infinite SCAN: An infinite model of diachronic semantic change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1605–1616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021a. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021b. Statistically significant detection of semantic shifts using contextual word embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Kohei Oda, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2024. Learning contextualized box embeddings with prototypical networks. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change. *Preprint*, arXiv:1811.06278.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Elizabeth Closs Traugott. 2017. Semantic change. *Oxford Research Encyclopedia of Linguistics*.

Jacob Turton, Robert Elliott Smith, and David Vinson. 2021. Deriving contextualised semantic features from BERT (and other transformer model) embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 248–262, Online. Association for Computational Linguistics.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 1–8, Marseille, France. European Language Resources Association.

Akira Utsumi. 2018. A neurobiologically motivated analysis of distributional semantic models. *Preprint*, arXiv:1802.01830.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

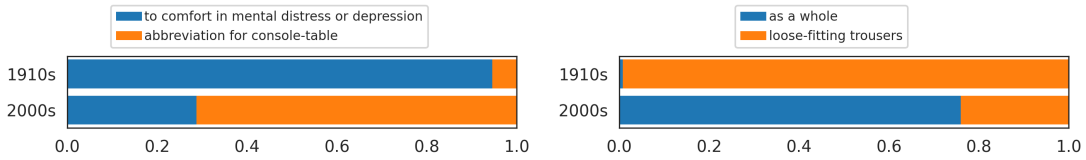
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Lütfi Kerem Şenel, İhsan Utlu, Furkan Şahinuç, Haldun M. Ozaktas, and Aykut Koç. 2020. Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering*, 27(6):721–746.

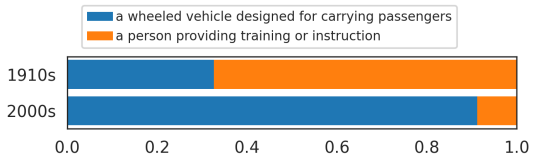
A Distributions of the usage types

The distributions of the usage types for several representative words in Table 4 are shown in Figures 7 and 8.

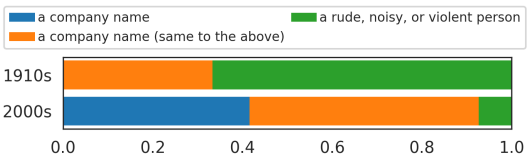


(a) console [PC1, Artifact, ↑]

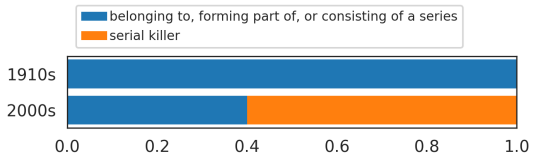
(b) overall [PC1, Artifact, ↓]



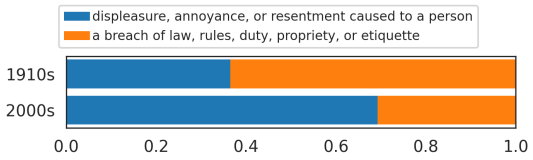
(c) coach [PC2, Human, ↑]



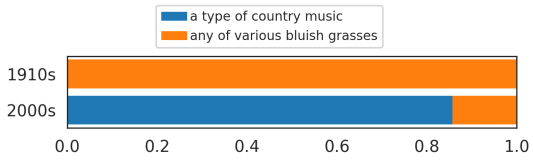
(d) yahoo [PC2, Human, ↓]



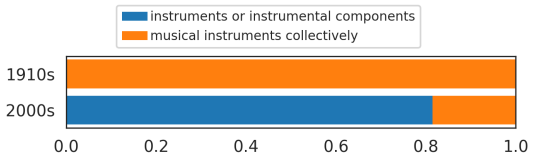
(e) serial [PC3, Negative Meaning, ↑]



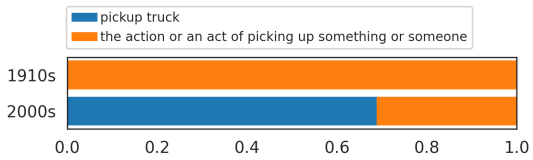
(f) offence [PC3, Negative Meaning, ↓]



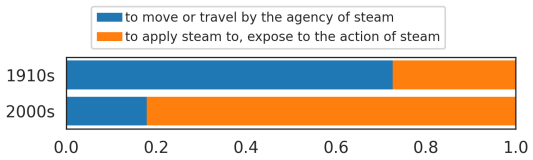
(g) bluegrass [PC4, Sound, ↑]



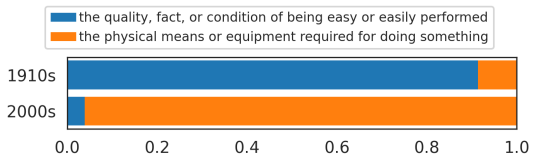
(h) instrumentation [PC4, Sound, ↓]



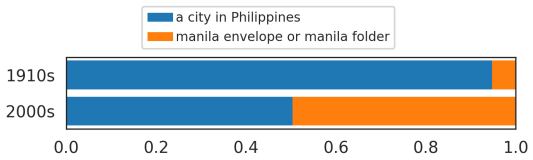
(i) pickup [PC5, Transportation, ↑]



(j) steamed [PC5, Transportation, ↓]

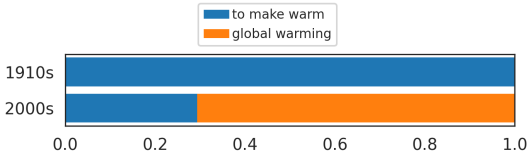


(k) facility [PC6, Place, ↑]

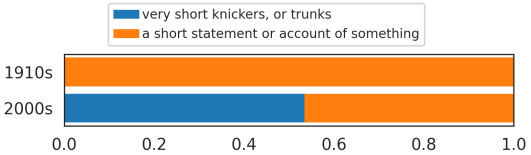


(l) manila [PC6, Place, ↓]

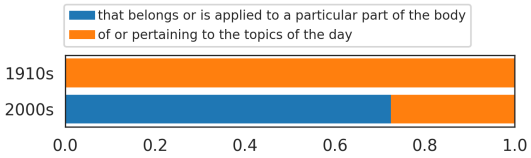
Figure 7: Distributions of usage types of representative words. These words are excerpted from Table 4. An ID of a principal component, an LSC type label, and an arrow indicating the direction of semantic change of each word are in parentheses.



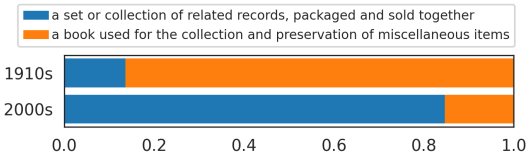
(m) warming [PC7, Social, ↑]



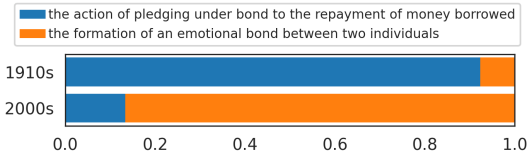
(n) briefs [PC7, Social, ↓]



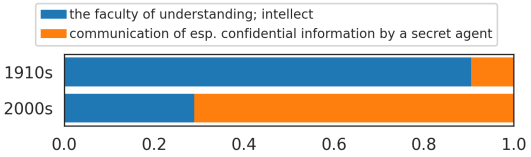
(o) topical [PC8, Familiar Thing, ↑]



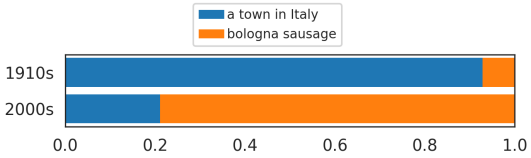
(p) album [PC8, Familiar Thing, ↓]



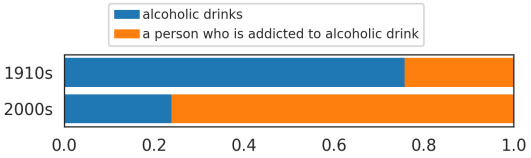
(q) bonding [PC9, Positive Meaning, ↑]



(r) intelligence [PC9, Positive Meaning, ↓]



(s) bologna [PC10, Food, ↑]



(t) alcoholic [PC10, Food, ↓]

Figure 8: Distributions of usage types of representative words (cont.)

Same Spelling, Different Functions *Mah*: Evaluating Language Models' Understanding of Singlish Particles

Chan Young Jung
Department of Linguistics, Korea University
laspebro@korea.ac.kr

Abstract

Singlish discourse particles exhibit tone-sensitive polysemy, where identical orthographic forms—distinguished only by prosodic cues in speech—serve distinct pragmatic functions. This poses a fundamental challenge for unimodal language models that must infer particle meanings solely from text. We thus investigate whether contextual information enables language models to predict appropriate particles in a cloze-style task, and whether increased data exposure—through domain-specific pretraining or in-context prompting—improves performance. To enable fair evaluation, we organize particles into semantic groups that minimize intra-group functional overlap. We test three BERT variants—including a Singlish domain-specific SingBERT model—and GPT-4.1 under zero-shot, definition-prompted, and few-shot conditions. Results demonstrate that domain-specific pretraining yields consistent performance gains over general English models (56.2% vs 30.1%), yet absolute performance remains modest across all approaches. GPT-4.1 shows variable performance across semantic groups and prompting strategies (23.8%–66.4%). These findings reveal that contextual cues only partially compensate for the absence prosodic information, highlighting fundamental limitations of text-only approaches for contact languages with substrate-derived pragmatic systems and the need for prosody-aware computational methods.

1 Introduction

Colloquial Singapore English (hereafter Singlish) is an English-based contact language that draws substrate influences from Singapore's multilingual landscape, including Malay, Tamil, and Sinitic varieties such as Hokkien and Cantonese (Deterding, 2007; Leimgruber, 2011; Chow and Bond, 2022; Ningsih and Rahman, 2023). A hallmark characteristic of Singlish is its extensive use of pragmatic discourse particles—such as *lah*, *leh*, *hor*, and *sia*—which, while removable without affecting grammaticality, encode important propositional content (Ler, 2006; Chow, 2021). Crucially, these particles are seldom monosemous; their meanings and discourse functions are jointly determined by contextual and prosodic cues (Lim, 2007; Wong, 2014; Soh et al., 2022). This tone-sensitivity potentially undermines the ability of unimodal language models to process Singlish, as they operate solely on orthographic input without prosodic notation.

Intra-particle polysemy is illustrated in the following sentences from the English subset of the National University of Singapore SMS corpus (Chen and Kan, 2015), a collection of over 55K messages in Singapore English:

(1) U typing the outline into the google doc *hor*? (#15340)

(2) Drive carefully when u come back *hor*... Raining heavily... (#15123)

In (1), *hor* functions as a confirmation-seeking question marker, converting the proposition into an interrogative while presuming its truth value. In contrast, in (2) *hor* adds precautionary force to an imperative, emphasizing the warning nature of the utterance. These functional distinctions are distinguished through rising versus falling tonal contours (Gupta, 1992; Kim, 2014; Lee, 2018; Chow, 2021; Liu et al., 2022; Chow et al., 2024). These examples illustrate that while prosodic cues disambiguate particle functions in speech, particle meanings in written Singlish must be inferred from context. This raises the question of whether contextual information alone enables models to predict the appropriate particle.

Since Singlish particles are embedded within English lexical items and syntactic structures, two possibilities arise: on the one hand, English-based models might benefit from cross-lingual transfer

from high-resource English data, as demonstrated by Armstrong et al. (2022) on Jamaican Patois. Conversely, tone-conditioned discourse particles—which are absent from standard English varieties—may fall outside the latent knowledge models acquire through pretraining. We investigate whether increased data exposure, either in the form of language-specific training data or particle definitions and usage examples within a prompt, can compensate for limited exposure to Singlish during initial pretraining.

This work addresses the computational challenge of Singlish particle disambiguation through the evaluation of contemporary language models using a novel semantic grouping approach. We extract particle-containing sentences from the NUS-SMS corpus and develop a manually annotated subset. Drawing on extensive prior literature, we identify the pragmatic functions of 10 common Singlish particles and organize them into semantically coherent groups that minimize functional overlap within groups, while enabling fair comparison of model performance across different functional categories. We evaluate three masked language models with varying training data exposure—from general English pretraining (BERT-base-uncased) to multilingual training (BERT-base-multilingual-uncased) to domain-specific training on Singlish and Malaysian English texts (SingBERT)—alongside a generative model (GPT-4.1) tested under different data exposure conditions: zero-shot, few-shot, and definition-prompted settings.

Our contributions are threefold: (1) We develop a semantic grouping strategy for Singlish particles that minimizes intra-group functional overlap, enabling fair evaluation of model performance on distinct pragmatic functions; (2) We evaluate how different training data exposures affect particle function recognition across masked language models and generative models; (3) We provide empirical evidence that even domain-specific pretraining achieves only modest performance on tone-sensitive particles, demonstrating fundamental limitations of text-only approaches and the need for prosody-aware methods for contact languages.

2 Related Work

2.1 NLP for Creoles

Natural language processing research on creole languages has historically been sparse, despite creoles being spoken by hundreds of millions of people worldwide (Lent et al., 2021). This neglect stems from societal stigmatization rooted in colonial histories, their predominantly oral nature, and exclusion from major multilingual datasets and language family classifications (Lent et al., 2022b, 2024).

Creoles present computational challenges that distinguish them from typical low-resource scenarios. Unlike languages with clear genealogical lineages, creoles emerge from complex contact situations involving multiple substrate and superstrate languages. This mixed ancestry undermines standard transfer learning assumptions: Lent et al. (2022a) demonstrated that straightforward transfer from ancestor languages to creoles often fails to achieve expected performance gains, as lexical items may derive from one language while syntactic structures reflect another.

Recent efforts have expanded to include named entity recognition (Adelani et al., 2021), sentiment analysis (Muhammad et al., 2022), and comprehensive multilingual evaluation frameworks (Lent et al., 2024). However, semantic disambiguation challenges—particularly for substrate-derived features like Singlish discourse particles—remain largely unaddressed. Addressing these challenges requires approaches that account for the complex interplay between superstrate lexical foundations and substrate pragmatic systems, as we examine in the context of Singlish particle processing.

2.2 Computational Approaches to Singlish Particle Disambiguation

Computational approaches to Singlish discourse particles have emerged from diverse methodological directions, with early work focusing on syntactic representation rather than semantic disambiguation. Wang et al. (2017) conducted foundational work by creating a Universal Dependencies treebank for Singlish and training neural dependency parsers with neural stacking to integrate English syntactic knowledge. While achieving significant parsing improvements, their approach treated particles uniformly within grammatical frameworks rather than addressing their polysemous functions.

Rule-based approaches have attempted representation within formal grammar frameworks. Chow (2021) and Chow and Bond (2022) developed HPSG-based grammars representing sentence-final particles as heads selecting sentences as complements, organizing particles into hierarchical types based on positional constraints. Although structurally thorough, these approaches focus on syn-

tactic distribution rather than semantic disambiguation.

Neural generation approaches have addressed particles within broader paraphrasing frameworks. Liu et al. (2022) integrated particle processing into Singlish-to-English translation through "semantic level rewriting," demonstrating that particles like *lah* (mood marker) and *leh* (tentative request marker) require clause-level understanding rather than word-level replacement. However, their approach did not specifically target disambiguation of particle functions based on prosodic or contextual cues.

Recent work has begun to explicitly address particle semantics. Chow et al. (2024) created SingDict, an open-source dictionary including particles with tonal annotations, while Foo and Ng (2024) specifically tackled disambiguation for three particles (*lah*, *meh*, *hor*) using task-driven representations with SingBERT, subtracting vector embeddings to isolate particle representations and performing unsupervised clustering to identify pragmatic functions. Current computational approaches to processing Singlish have also expanded to include style transfer (Liang et al., 2025), content moderation (Foo and Khoo, 2025), and multimodal understanding (He et al., 2025), reflecting growing recognition of Singlish's computational importance.

Our work differs from prior approaches in three key respects. First, rather than treating all particles uniformly (as in syntactic approaches) or focusing on individual particles in isolation (as in clustering-based methods), we explicitly address how tone-dependent polysemy creates overlapping pragmatic functions across particles. Second, we investigate whether varying levels of data exposure enable models to learn contextual patterns that compensate for missing prosodic information. Third, while previous work has primarily focused on syntactic parsing or translation, we directly evaluate models' capacity for particle prediction in authentic conversational Singlish, demonstrating that substrate-derived, tone-sensitive pragmatic features fundamentally limit models' processing of contact languages.

3 Methods

3.1 Dataset Construction and Particle Selection

We extract particle-containing sentences from the English subset of the National University of Singapore SMS corpus (Chen and Kan, 2015), which contains over 55,000 short, informal text messages from Singaporeans, primarily students at the National University of Singapore. The corpus provides naturalistic data with some chronologically ordered conversations, allowing for opportunistic retrieval of conversational context.

To identify target particles for analysis, we compiled definitions and functional descriptions from extensive prior research on Singlish discourse particles (Chow, 2021; Gupta, 1992; Kim, 2014; Khoo, 2012; Lee, 2018; Leimgruber, 2016; Leimgruber et al., 2021; Lim, 2007; Liu et al., 2022; Platt and Ho, 1989; Soh et al., 2022; Wong, 2014). We used regular expressions to extract sentences containing these particles, focusing specifically on substrate-derived particles (thus excluding *one* and *what*, which share orthographic forms with Standard English words despite having distinct pragmatic functions in Singlish).

Through manual inspection and frequency analysis, we identified 10 particles that were both frequent in the corpus and whose meanings could be reliably verified in the literature. Four particles exhibited tonal polysemy—distinct pragmatic functions associated with different tonal variants. Table 1 presents our final particle inventory with their tonal variants, pragmatic functions, syntactic environments, and number of appearances in our dataset.

We applied basic preprocessing including deduplication, removal of anonymization artifacts, filtering sentences with fewer than three words to ensure sufficient context, and exclusion of purely Mandarin or other substrate language content. For particles exhibiting tonal polysemy, we manually classified each instance into the closest functional variant based on contextual and syntactic cues, following the definitions established in our literature review.

3.2 Semantic Grouping Strategy

Rather than attempting simultaneous classification across all particle functions, which would unfairly penalize models due to substantial semantic overlap and an excessively wide range of candidate labels,

Particle	Tone	Pragmatic Function	DEC	IMP	INT	Count
ah	rising	confirms understanding/acknowledgement	+	+		425
	low	tag question/echo-question marker			+	367
bah	low	hedge; uncertainty/lack of commitment	+	+	+	316
hor	rising	question marker; speaker believes proposition true			+	107
	low falling	warning/disclaimer marker	+	+		50
lah	low	persuades acceptance of proposition	+	+		511
	falling	presents solutions; conveys annoyance	+	+		88
leh	mid-level	persuades action/belief acceptance	+	+		146
	low	marks new information/counters assumptions	+			672
liao	falling	past tense/perfective aspect marker	+	+	+	624
lor	mid-level	marks obviousness/resignation; agreements	+	+		957
mah	high	marks information as obvious	+			254
meh	high	question marker; skepticism/doubt			+	181
sia	rising	reduces distance; surprise/admiration	+		+	212
Total						**4,910**

Table 1: Singlish sentence-final particles with tonal variants, pragmatic functions, syntactic environments, and number of occurrences in our NUS-SMS subset. DEC = declarative, IMP = imperative, INT = interrogative. A "+" indicates the particle can occur in the sentence type. Particles with multiple rows show tone-sensitive polysemy.

Group	Particles	Count
1	low *ah*, rising *hor*, high *meh*	655
2	low *leh*, falling *liao*, high *mah*, rising *sia*	1,762
3	falling *lah*, mid-level *leh*, low falling *hor*, low *bah*	600
4	low *lah*, rising *ah*, mid-level *lor*	1,893

Table 2: Semantic grouping of Singlish particles to minimize functional overlap within groups.

Paradigm	Model	Data Exposure
BERT-based	BERT-base-uncased	English
	BERT-base-multilingual-uncased	Top 102 languages (Wikipedia)
	SingBERT	Singlish/Manglish (subreddits, forums)
GPT-4.1	Zero-shot	Baseline
	Few-shot	NUS-SMS sentences with particles
	Definition-prompted	Particle definitions

Table 3: Model configurations with their respective data exposure. BERT models are pretrained on the indicated corpora; GPT-4.1 variants receive information through in-context prompting.

we organized particles into four semantic groups that minimize functional overlap within groups while allowing fair evaluation across distinct pragmatic domains. Table 2 summarizes our semantic grouping with the distribution of instances across groups.

Our grouping strategy prioritizes syntactic and functional similarity while maintaining relatively balanced instance counts across groups. Group 1 comprises question markers that occur exclusively in interrogative contexts. Group 2 includes particles that function as non-imperative markers. Groups 3 and 4 organize the remaining particles using complementary pragmatic functions while avoiding intra-group semantic overlap.

3.3 Model Selection and Implementation

Table 3 summarizes our experimental setup across two paradigms: BERT-based models and GPT-4.1, selected to capture different aspects of particle understanding and exposure to training data.

Our evaluation task requires models to predict the correct particle for masked positions in authentic Singlish sentences. Within each semantic group, models select from 3–4 candidate particles (e.g., Group 1 candidates are low *ah*, rising *hor*, and high *meh*). The models predict which particle should appear in context, and we compare this against the actual particle found in the corpus.

BERT-based models. For masked language models, we implement a probabilistic scoring approach to handle particles that tokenize into multiple subwords. Given a sentence with [MASK] in the particle position, we expand the mask to accommodate the number of subwords in each candidate particle. For each candidate, we compute the cumulative log-probability by iteratively predicting each subword position, conditioning subsequent predictions on previously selected tokens (greedy left-to-

right decoding). The candidate with the highest cumulative log-probability is selected as the model's prediction. We use BERT-base-uncased, BERT-base-multilingual-uncased, and SingBERT in their released forms with default tokenizers, without additional fine-tuning on our task.

GPT-4.1. For the generative model, we use structured prompts that present the cloze sentence and explicitly list the candidate particles for the given semantic group. We set `temperature=0.0` to ensure deterministic predictions and `max_tokens=5` to enforce concise responses containing only the predicted particle. For the few-shot condition, we provide 3 example sentences per particle drawn from the NUS-SMS corpus; these examples were held out from the test set to prevent data leakage. For the definition-prompted condition, we include functional descriptions of each candidate particle based on our literature review. Complete prompt templates and particle definitions are provided in Appendix A.

3.4 Evaluation Metrics

We report accuracy as our primary metric: the proportion of correct predictions out of all test instances. Given the substantial frequency imbalance across particles within groups (Table 1), we report both micro-averaged and macro-averaged metrics. Micro-averaged accuracy weights predictions by instance count, reflecting overall performance as weighted by the natural distribution of particles in conversational Singlish. Macro-averaged metrics compute unweighted averages across particles, revealing whether models perform consistently across all particle types regardless of frequency. For detailed group-level analysis (Tables 4–7), we report micro-averaged accuracy alongside macro-averaged precision and F1 scores to assess per-particle performance.

4 Results and Discussion

Figure 1 presents micro-averaged accuracy across all groups and models, while Figure 2 shows overall performance averaged across semantic groups. Detailed results for each semantic group are shown in Tables 4–7.

4.1 Domain-Specific Training Effects

SingBERT consistently outperforms both general English and multilingual BERT models across all semantic groups, achieving an overall micro-averaged accuracy of 56.2% compared to 30.1% for BERT-base and 34.1% for BERT-multi. However, these modest absolute performance levels highlight the fundamental difficulty of the particle disambiguation task when prosodic information is unavailable. The performance advantage is most pronounced in Group 2 (non-imperative markers), where SingBERT achieves 62.4% accuracy compared to 15.0% and 18.9% for the baseline models respectively, demonstrating the critical importance of domain-specific exposure to Singlish linguistic patterns.

Notably, the performance gap narrows in Group 3, where BERT-multi outperforms SingBERT (40.2% vs 36.7%). This suggests that cross-lingual transfer from substrate languages may benefit certain pragmatic functions that bridge multiple linguistic systems within the multilingual architecture.

Model	Acc	Prec	F1
BERT-base	52.4	39.4	32.5
BERT-multi	51.9	35.9	30.8
SingBERT	65.5	63.4	59.1
GPT-4.1 (0-shot)	63.5	58.7	58.4
GPT-4.1 (def)	66.4	62.5	58.2
GPT-4.1 (few)	65.8	45.0	43.2

Table 4: Group 1 results: Question markers (low *ah*, rising *hor*, high *meh*). All models show relatively strong performance, with GPT-4.1 definition-prompted achieving highest accuracy. For all groups (Tables 4–7), accuracy is micro-averaged while precision and F1 are macro-averaged.

Model	Acc	Prec	F1
BERT-base	15.0	20.7	11.9
BERT-multi	18.9	20.7	17.5
SingBERT	62.4	58.9	52.3
GPT-4.1 (0-shot)	52.1	19.2	17.9
GPT-4.1 (def)	54.8	29.2	27.8
GPT-4.1 (few)	55.9	25.1	25.1

Table 5: Group 2 results: Non-imperative markers (low *leh*, falling *liao*, high *mah*, rising *sia*). Largest performance gap between SingBERT and other models, highlighting domain expertise importance.

4.2 Generative Model Performance

GPT-4.1 demonstrates variable performance across prompting strategies and semantic groups. Zero-shot micro-averaged performance ranges from 23.8% (Group 3) to 63.5% (Group 1), indicating substantial but uneven knowledge of Singlish particle functions. Definition prompting consistently improves performance across all groups, with the

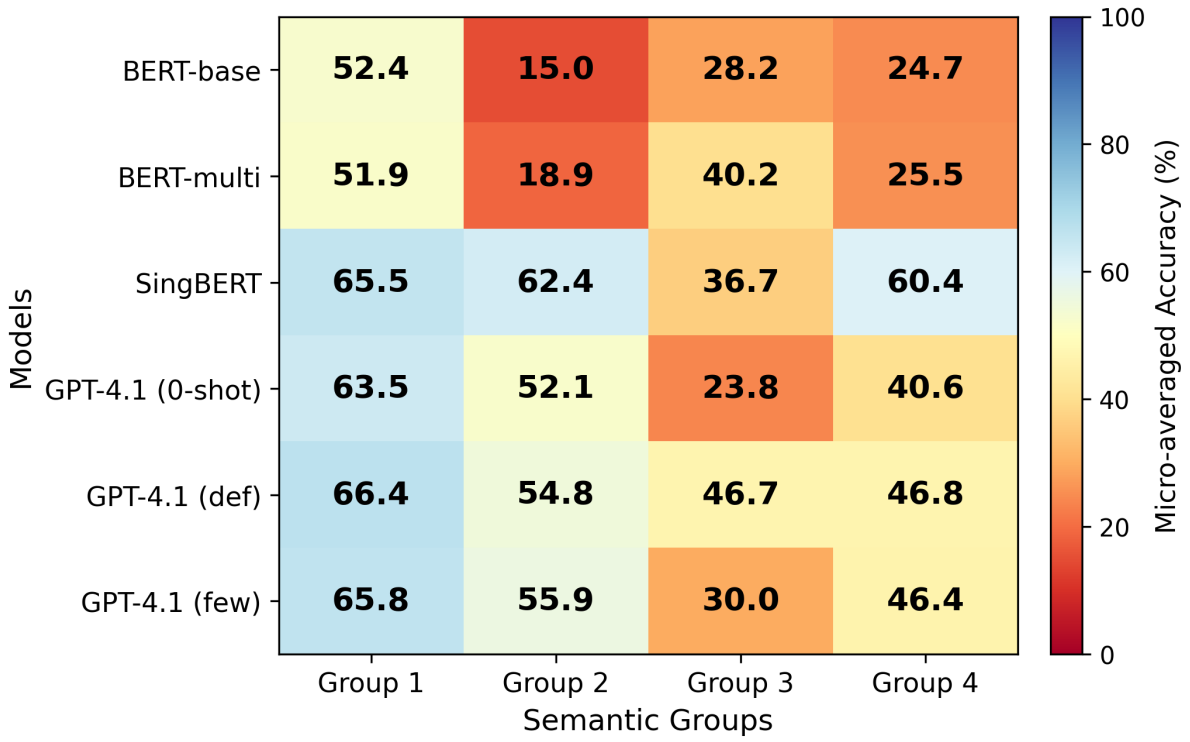


Figure 1: Model performance across semantic groups showing micro-averaged accuracy percentages. SingBERT demonstrates consistent performance across groups, while other models show variable effectiveness depending on pragmatic domain.

Model	Acc	Prec	F1
BERT-base	28.2	21.2	20.8
BERT-multi	40.2	26.3	25.3
SingBERT	36.7	46.3	37.4
GPT-4.1 (0-shot)	23.8	20.7	10.6
GPT-4.1 (def)	46.7	44.3	31.9
GPT-4.1 (few)	30.0	40.1	23.7

Table 6: Group 3 results: Mixed pragmatic functions (falling *lah*, mid-level *leh*, low falling *hor*, low *bah*). Most challenging group across all models, with definition prompting showing greatest improvement.

Model	Acc	Prec	F1
BERT-base	24.7	39.1	19.7
BERT-multi	25.5	36.3	19.8
SingBERT	60.4	57.6	55.0
GPT-4.1 (0-shot)	40.6	28.3	23.8
GPT-4.1 (def)	46.8	47.6	45.1
GPT-4.1 (few)	46.4	49.6	46.1

Table 7: Group 4 results: Persuasive/confirmatory markers (low *lah*, rising *ah*, mid-level *lor*). SingBERT shows strong performance, with consistent improvement across GPT-4.1 prompting strategies.

most substantial gains in Group 3 (46.7% vs 23.8% zero-shot). This improvement pattern indicates that GPT-4.1 possesses latent knowledge about Singlish pragmatics that can be activated through appropriate metalinguistic scaffolding.

Few-shot prompting shows mixed results, sometimes underperforming definition prompting (Group 3: 30.0% vs 46.7%), suggesting that metalinguistic guidance may be more effective than exemplar-based learning for this task.

4.3 Semantic Group Analysis

Group 1 (question markers) shows the most consistent performance across models, with all models except BERT-base achieving above 50% accuracy. This suggests that interrogative particles may be more learnable due to their clearer syntactic constraints.

Group 2 exhibits the largest performance disparity between domain-specific and general models, highlighting the importance of exposure to Singlish-specific pragmatic patterns. The poor performance of general English models (15.0% and 18.9%) indicates that these particles encode discourse functions not readily transferable from standard English patterns.

Group 3 proves most challenging across all models, with no model achieving above 47% accuracy.

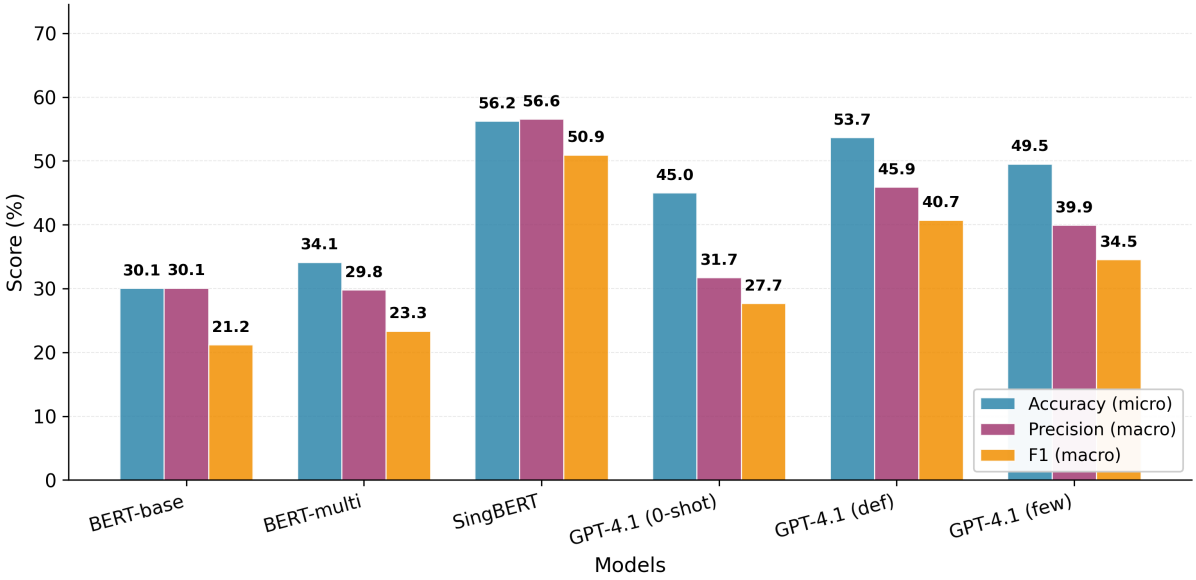


Figure 2: Overall performance comparison showing micro-averaged accuracy, macro-averaged precision, and macro-averaged F1 scores, each averaged across the four semantic groups.

This group contains particles with the most complex pragmatic functions and highest degree of contextual sensitivity, suggesting that current approaches struggle with highly context-dependent semantic disambiguation.

Group 4 shows strong performance for SingBERT (60.4%) and moderate but consistent improvement for GPT-4.1 across prompting strategies, indicating that persuasive and confirmatory functions may be more accessible to computational models.

4.4 Implications for Contact Language Processing

Our findings reveal several key implications for computational approaches to contact languages. The consistent benefits of domain-specific pretraining underscore the necessity of specialized training data for creole language processing. However, the universally modest absolute performance levels—even SingBERT achieves only 56.2% overall accuracy—point to fundamental limitations in current text-only approaches when pragmatic meaning is prosodically encoded. The variable effectiveness of different prompting strategies suggests that large generative models possess relevant but unevenly accessible knowledge about contact language features, with definition prompting (53.7%) substantially outperforming zero-shot approaches (45.0%) while few-shot prompting shows inconsistent benefits (49.5%).

5 Conclusion

This study demonstrates that tone-sensitive pragmatic phenomena in Singlish discourse particles expose fundamental limitations of contemporary language models operating solely on orthographic input. Even with domain-specific pretraining, performance remains far from human-level, underscoring the difficulty of capturing prosodically encoded meaning from text alone. These findings highlight that pragmatic interpretation in contact languages cannot be reduced to surface form recognition: it requires sensitivity to prosody, stance, and interactional context.

Our semantic grouping framework provides a principled evaluation methodology that mitigates functional overlap between particles, offering an approach generalizable to other contact varieties with complex pragmatic systems. The framework reveals systematic performance patterns: interrogative markers (Group 1) achieve relatively consistent results across models, while non-imperative markers (Group 2) and context-dependent functions (Group 3) prove substantially more challenging, particularly for models lacking Singlish-specific training. The variable success of definition prompting—with gains of over 20 percentage points in Group 3—further indicates that large generative models contain latent knowledge of such systems, but that this knowledge requires explicit scaffolding to be reliably accessed.

Taken together, these results argue for expanding

computational approaches to under-resourced contact languages beyond text-only evaluation toward multimodal, prosody-aware methods that recognize the interplay of substrate-derived pragmatic systems with lexifier structures. Beyond Singlish, this work illustrates how contact languages can serve as critical testing grounds for theories of meaning in NLP, revealing where current models succeed, where they fail, and what linguistic knowledge remains inaccessible through distributional learning alone.

Limitations

While our evaluation provides valuable insights into computational particle disambiguation, certain limitations must be acknowledged. First, our manual annotation of tonal variants introduces potential subjectivity, particularly for ambiguous cases where contextual cues are insufficient to determine intended prosodic realization. Given the complex sociolinguistic nature of Singlish and the inherent difficulty of defining "native speaker" status in a contact language context, some degree of interpretive judgment is unavoidable. We mitigated this through extensive consultation of established literature and consistent application of documented functional criteria.

Second, the NUS-SMS corpus, while representing a specific demographic (primarily university students), constitutes one of the few available naturalistic Singlish corpora with substantial particle usage. Despite its demographic constraints, this corpus provides valuable authentic data. Our semantic grouping strategy, while involving theoretical judgment about functional similarity, is grounded in established pragmatic distinctions documented in extensive prior literature on Singlish discourse particles.

Finally, our evaluation framework focuses on Singlish and textual particle prediction, which may limit generalizability to other contact languages or multimodal contexts. Future work incorporating prosodic information and expanding to additional creole varieties could enhance our understanding of computational challenges across diverse contact language phenomena.

Acknowledgments

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). We acknowledge the creators of the NUS-SMS corpus for making their data available for research purposes.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. Jampatoisnli: A jamaican patois natural language inference dataset. ArXiv preprint arXiv:2212.03419.

Tao Chen and Min-Yen Kan. 2015. The national university of singapore sms corpus. ScholarBank@NUS Repository, https://doi.org/10.25540/WVM0-4RNX.

Siew Yeng Chow. 2021. *A computational grammar of Singlish using HPSG*. Ph.D. thesis, Nanyang Technological University.

Siew Yeng Chow and Francis Bond. 2022. Singlish where got rules one? constructing a computational grammar for singlish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5243–5250.

Siew Yeng Chow, Chang-Uk Shin, and Francis Bond. 2024. This word mean what: Constructing a singlish dictionary with chatgpt. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 41–50.

David Deterding. 2007. *Singapore English*. Edinburgh University Press.

Jessica Foo and Shaun Khoo. 2025. Lionguard: A contextualized moderation classifier to tackle localized unsafe content. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 707–731.

Linus Tze En Foo and Lynnette Hui Xian Ng. 2024. Disentangling singlish discourse particles with task-driven representation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pages 1–6.

Anthea Fraser Gupta. 1992. The pragmatic particles of singapore colloquial english. *Journal of Pragmatics*, 18(1):31–57.

Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy Chen,

and Aiti Aw. 2025. Meralion-audiollm: Advancing speech and language understanding for singapore. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 22–30.

Velda Yuan Ling Khoo. 2012. The sia particle in colloquial singapore english.

Chong-Hyuck Kim. 2014. Discourse particles: Focusing on colloquial singapore english. *The New Studies of English Language and Literature*, 59:225–248.

Si Kai Lee. 2018. *A Nanosyntactic approach to sentence-final particles in Singlish: A cartographic perspective*. Ph.D. thesis, National University of Singapore.

Jakob RE Leimgruber. 2011. Singapore english. *Language and Linguistics Compass*, 5(1):47–62.

Jakob RE Leimgruber. 2016. Bah in singapore english. *World Englishes*, 35(1):78–97.

Jakob RE Leimgruber, Jun Jie Lim, Wilkinson Daniel Wong Gonzales, and Mie Hiramoto. 2021. Ethnic and gender variation in the use of colloquial singapore english discourse particles. *English Language & Linguistics*, 25(3):601–620.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. *arXiv preprint arXiv:2109.06074*.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022a. Ancestor-to-creole transfer is not a walk in the park. *arXiv preprint arXiv:2206.04371*.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. What a creole wants, what a creole needs. *arXiv preprint arXiv:2206.00437*.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, and 1 others. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Vivien Soon Lay Ler. 2006. A relevance-theoretic approach to discourse particles in singapore english. In *Approaches to discourse particles*, pages 149–166. Brill.

Jinggui Liang, Dung Vo, Yap Hong Xian, Hai Leong Chieu, Kian Ming A Chai, Jing Jiang, and Lizi Liao. 2025. Colloquial singaporean english style transfer with fine-grained explainable control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26962–26983.

Lisa Lim. 2007. Mergers and acquisitions: on the ages and origins of singapore english particles 1. *World Englishes*, 26(4):446–473.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, and 1 others. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*.

Nourma Silvia Ningsih and Fadhlur Rahman. 2023. Exploring the unique morphological and syntactic features of singlish (singapore english). *Journal of English in Academic and Professional Communication*, 9(2):72–80.

John T Platt and Mian Lian Ho. 1989. Discourse particles in singaporean english: Substratum influences and universals. *World Englishes*, 8(2):215–221.

Ying Qi Soh, Junwen Lee, and Ying-Ying Tan. 2022. Ethnicity and tone production on singlish particles. *Languages*, 7(3):243.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial singaporean english. *arXiv preprint arXiv:1705.06463*.

Jock O Wong. 2014. *The culture of Singapore English*. Cambridge University Press.

A Prompt Templates

A.1 Zero-shot Prompt

```
You are given a Singlish sentence with
a missing word marked as [MASK]. Fill
in the [MASK] with exactly one of the
following particles: {candidates}. Do
not output anything else.

Sentence: {cloze_text}
```

A.2 Definition-prompted Template

```
You are given a Singlish sentence with
a missing word marked as [MASK]. Fill
in the [MASK] with exactly one of the
following particles: {candidates}. Use
the following definitions to guide your
choice:

[Particle definitions inserted here
based on semantic group]

Do not output anything else.

Sentence: {cloze_text}
```

A.3 Few-shot Template

You are given a Singlish sentence with a missing word marked as [MASK]. Fill in the [MASK] with exactly one of the following particles: {candidates}. Below are example usages for each particle:

[Examples for each particle inserted here]

Do not output anything else.

Sentence: {cloze_text}

A.4 Particle Definitions by Semantic Group

A.4.1 Group 1: Question Markers

ah: question marker (tag question or echo-question), elicits affirmation or confirmation
hor: question marker; indicates that the speaker believes the proposition to be true
meh: question marker; indicates skepticism or doubt

A.4.2 Group 2: Non-imperative Markers

leh: marks new information or re-asserts old information, possibly to counter the addressee's assumptions; used for declaratives
liao: past tense/perfective aspect marker; used for declaratives, imperatives, interrogatives
mah: marks information as obvious; used for declaratives
sia: reduces the distance between interlocutors and marks coarseness, surprise, or admiration; used for declaratives and interrogatives

A.4.3 Group 3: Mixed Pragmatic Functions

lah: presents answers or solutions to questions or situations; conveys annoyance and unfriendliness towards the addressee; used for declaratives and imperatives
leh: persuades the addressee to take action or accept a belief; used for declaratives and imperatives
hor: indicates a warning or disclaimer; used for declaratives and imperatives
bah: hedge; marks uncertainty and lack of commitment about a proposition; used for declaratives, imperatives, interrogatives

A.4.4 Group 4: Persuasive/Confirmatory Markers

lah: persuades the addressee to accept a proposition the speaker believes to be true; used for declaratives and imperatives
ah: confirms the addressee's acknowledgement or understanding of a proposition; used for declaratives and imperatives
lor: marks obviousness and resignation; often used for agreements; used for declaratives and imperatives

A.5 Few-shot Examples by Semantic Group

A.5.1 Group 1 Examples

ah:
1. K.:)you are the only girl waiting in reception ah?
2. Oh all have to come ah?
3. K. Did you call me just now ah?

hor:
1. Hey, u haven't upload the latest copy hor?
2. U typing the outline into the google doc hor?
3. Dear so we waiting u at orchard hor? Head of train k

meh:
1. Can meh? Thgt some will clash... Really ah, i dun mind...
2. Now got tv 2 watch meh? U no work today?
3. Huh... U serious of poning ah... Deepavali not nxt wk meh?

A.5.2 Group 2 Examples

leh:
1. haha but no money leh... Later got to go for tuition...
2. Tmr v crowded leh, weekday go la...
3. Huh... I mean e orientation in e first wk leh... Not majors...

liao:
1. Juz now havent woke up so a bit blur blur... Dad went out liao...
2. not goin 4 any camps... My faculty camp oso over liao...
3. oredi on my way to e class liao...

mah:
1. C movie is juz last minute decision mah. Juz watch 2 lar...
2. U must key in the amount on top first mah
3. Lol because if im not there and you kena caught, it will be very awkward mah lol.

sia:
1. guess wad sia? i won preview tickets to this korean show!
2. Haha. Good what. Can earn another 1k plus. Rich sia u.
3. U so serious till hallucinate?! Serious sia! u better stop training

A.5.3 Group 3 Examples

lah:
1. Then give mine to the person who doesnt have it lah.
2. Okayokay but what's done is done lah.
3. Borrow the one at home lah. Also, camera i think no choice...

leh:
1. Let me know asap leh
2. What u all buying? Help me to buy leh. I go join u all now.
3. It's ok. I'm already at your place. Open the door for me leh.

hor:
1. U dun say so early hor... U c already then say...
2. Drive carefully when u come back hor... Raining heavily...
3. u so naughty!!!! dun sleep so late hor. hug you tight tight.

bah:
1. Shld be ard 4 to 5 bah. What time e thing starts ar..
2. Don't know leh. Maybe his office bah.
3. Okie... scarly u arrive first arh lol... I think he shld be going bah

A.5.4 Group 4 Examples

lah:
1. Ur haircut not bad lah, quite nice and dun really look gong.
2. i think no need lah..i go borrows from steve

3. Ohh.. I heard australia is good lah. Haha. I don’t intend to travel.

ah:
1. dun forget u still owe me a treat ah..haha
2. 630. Today ah! Later on... Dont be late... And dont gelek
3. fetch me at 6 ah.. arts there, e place u always pick me up one..

lor:
1. Anything lor up 2 u... Dun buy anything too expensive...
2. ya lor. as my friends doing agency job then many of them got more tutor than student.
3. No. They bound to tease at us, so just let them tease lor.

Evaluating Syntactic Generalization in Transformer-Based Models Using Korean Honorific Agreement

Nayoung Kwon[♠] **Seongmin Mun**[◊]
[♠]University of Oregon [◊]Kyungpook National University
nkwon@uoregon.edu seongminmun@knu.ac.kr

Abstract

This paper examines whether Transformer-based language models can generalize over abstract syntactic structures in low-resource languages. Focusing on Korean subject–verb honorific agreement, we evaluate KoBERT and KoGPT-2 using classification and attention analyses before and after fine-tuning. Results show that KoGPT-2, especially after fine-tuning, outperforms KoBERT in handling structurally complex constructions. Attention analyses reveal that KoGPT-2 shows more syntactic alignment than KoBERT, although inconsistently. Both models remain susceptible to interference from honorific attractors. These findings highlight key differences between autoregressive and masked LMs in syntactic generalization and show that attention may reflect syntactic structure but not reliably indicate grammatical competence.

1 Introduction

Transformer-based language models have achieved strong performance across a wide range of natural language processing tasks. Although growing evidence suggests that these models implicitly encode aspects of syntactic knowledge (Clark et al., 2019; Hewitt & Manning, 2019; Lin, Tan, & Frank, 2019; Wilcox, Futrell, & Levy, 2024), it remains unclear to what extent they can track syntax-sensitive dependencies and generalize over abstract syntactic structures independently of lexical content. Most work has focused on high-resource languages like English, while low-resource languages such as Korean—making up less than 1% of web content—pose unique challenges due to limited data. This study uses Korean to evaluate the syntactic knowledge encoded in Transformer-based models, focusing specifically on subject–verb honorific agreement.

Although sentences appear linear on the surface, linguistic theory has shown language is fundamentally hierarchical. This raises the question of whether language models can capture such structural information beyond surface patterns. A common strategy for probing syntactic generalization involves agreement phenomena, where two linguistic elements must match in morphosyntactic features. For instance, Goldberg (2019) showed BERT (Devlin et al., 2019) performs well on subject–verb agreement in English, suggesting that Transformer-based models can track syntactic dependencies. Bacon and Regier (2019) replicated these findings with accuracy above 90% across 26 languages. Their results further showed that model performance declines in the presence of agreement attractors and longer dependencies. More recently, Lasri, Lenci, and Poibeau (2022) demonstrated that Transformer-based models' generalization abilities are not fully lexically independent, particularly when processing sentences with attractors. Chaves and Richter (2021) similarly observed that while BERT encodes rich syntactic representations, it often relies on shallow heuristics (see also Wu & Dredze, 2020), in contrast to GPT-2 (Radford, 2019), which exhibited more informed behavior (Chang & Bergen, 2024). Taken together, these findings highlight the need for further investigation into the syntactic generalization capacities of language models, especially in typologically diverse and low-resource languages.

To this end, the present study evaluates the syntactic generalization abilities of KoBERT (Jeon, Lee, & Park, 2019) (92 million parameters) and KoGPT-2 (Jeon, 2021) (125 million parameters), using subject–verb honorific agreement in Korean. By evaluating model behavior before and after fine-tuning, we examine whether syntactic knowledge can emerge in the absence of explicit

syntactic modules, and how such knowledge interacts with attention mechanisms. This study contributes to our understanding of how neural language models engage with syntactic structure in a low-resource language, with implications for multilingual NLP and model interpretability.

2 Subject–Verb Honorific Agreement in Korean

Korean is an agglutinative SOV language in which each morpheme typically encodes a distinct grammatical function. For example, the honorific marker *-si-* attaches to a verb to signal respect for the subject (1). While the use of *-si-* is optional when the subject is honorifiable (2), it becomes ungrammatical with an unhonorifiable subject (3) (Sohn, 2001). Such violations elicit a P600 response (Kwon & Sturt, 2024), similar to effects reported for number and person agreement violations in English (Osterhout & Mobley, 1995) and Spanish (Barber & Carreiras, 2003). This suggests that Korean honorific agreement is processed as syntactic information, akin to other agreement phenomena.

(1) Honorifiable subject with *-si-*
emenim-i wu-si-ess-ta
mother-nom cry-HON-past-decl
'Mother cried.'

(2) Honorifiable subject without *-si-*
emenim-i wul-ess-ta
mother-nom cry-past-decl
'Mother cried.'

(3) Unhonorifiable subject with *-si-*
*kkoma-ka wu-si-ess-ta
kid-nom cry-HON-past-decl
'The kid cried.'

That is, subject–verb honorific agreement in Korean is conditioned by the syntactic accessibility of the subject (Yoon, 2009). While the use of the honorific marker *-si-* reflects social and pragmatic features such as respect, its grammaticality depends on whether the subject structurally licenses agreement. Thus, as an optional but structurally constrained phenomenon, it provides a unique testing ground for determining whether language models can acquire abstract syntactic dependencies without categorical surface cues.

Thus, this study uses subject–verb honorific agreement to test whether Transformer-based language models can generalize over abstract syntactic structures in Korean, a low-resource language. We also evaluate whether fine-tuning on honorific agreement data enhances models' sensitivity to syntactic dependencies. To this end, we constructed sentences with two syntactic configurations illustrated in English in (4) (Chomsky, 1981; Kwon & Polinsky, 2006): in NP1 control, the subject of the embedded verb (underlined) is the main clause subject, NP1; in NP2 control, it is a direct or indirect object, NP2.

(4) NP1-NOM NP2-to $\underline{\text{go}}_{\text{.emb}}$ $\text{told}_{\text{main}}$
NP1 control: 'NP1$_i$ told NP2 that he$_i$ went.'
NP2 control: 'NP1 told NP2$_i$ PRO$_i$ to go.'

To test whether KoBERT and KoGPT-2 have learned syntactic representations, we use subject–verb honorific agreement as a diagnostic via a classification task and self-attention analysis.

3 Experiment

3.1 Datasets

The dataset included the sentence types illustrated in (4), along with matched ungrammatical counterparts in which the honorific verb appears with a structurally licit but non-honorifiable subject. We also varied the honorific features of structurally illicit potential subjects to test for interference. This yielded all four NP1–NP2 feature combinations (H–H, H–NH, NH–H, NH–NH) across both NP1- and NP2-control types. Since the embedded verb is always honorific, NP1-control sentences require NP1 to be honorific, and NP2-control sentences require NP2—regardless of the other noun's features. This design isolates structural understanding from lexical honorific effects. For clarity, sample sentences are shown in English in (5) and (6), although the study was run in Korean. Note that asterisks (*) indicate grammatical violations in Korean (Sohn, 2001), as summarized in Table 1.

(5) NP1 control: The *kid$_i$/teacher$_i$ told the teacher/kid that __$_i$ closed (honorific) the door.

(6) NP2 control: The kid/teacher told the teacher$_i$/*kid$_i$ __$_i$ to close (honorific) the door.

Honorific features		Control type	
NP1	NP2	NP1	NP2
H	H	✓	✓
	NH	✓	✗
NH	H	✗	✓
	NH	✗	✗

Table 1: Grammatical acceptability of dataset

Reflecting naturally occurring distributional patterns in Korean, the dataset contained 1,336 NP1-control sentences and 5,304 NP2-control sentences, which were used for training and evaluation. We split the data into training (90%) and test (10%) sets.

3.2 Classification task analysis

Following previous studies (e.g., McCormick, 2019; Vázquez, 2020; Wu, 2019), we implemented a binary classification task to evaluate whether the models could distinguish grammatically acceptable from unacceptable sentences based on honorific agreement. For KoBERT, we used standard embedding techniques: token, position, and segment embeddings (Devlin et al., 2019). Each sentence was tokenized using the KoBERT tokenizer, truncated or padded to a maximum length of 256 tokens. Tokens were indexed based on the KoBERT vocabulary. Segment embeddings were assigned as binary indicators (0 or 1) to distinguish between sentence segments. Grammaticality labels were stored separately for use in supervised training. Training parameters were set as follows: batch size = 16, number of epochs = 30, random seed = 42, maximum sequence length = 256, epsilon = 8e–8, and learning rate = 0.0001. We fine-tuned KoBERT (Jeon, Lee, and Park 2019) on our dataset using the BertForSequenceClassification class from Huggingface's Transformers library (Wolf, 2019). Following training, we evaluated the model's performance on previously unseen sentences from the test set.

The KoGPT-2 training procedure followed a similar setup, with notable differences in input handling and model architecture. Unlike BERT, GPT-2 uses bytepair encoding (BPE) instead of WordPiece tokenization. In terms of training objectives, BERT is trained using masked language modeling and next-sentence prediction, whereas GPT-2 is trained using a causal (left-to-right) language modeling objective. Additionally, BERT processes input bidirectionally, while GPT-2 processes text unidirectionally, from left to right, without relying on special [CLS] or [SEP] tokens. We used KoGPT-2-base-v2 (Jeon, 2021), implemented using the GPT2ForSequenceClassification and PreTrainedTokenizerFast classes from the Transformers library (Wolf, 2019).

We fine-tuned both models for 30 epochs on the subject–verb honorific agreement dataset and evaluated their classification accuracy and attention alignment before and after training. This setup allowed us to assess their ability to generalize over syntax-sensitive dependencies and examine whether fine-tuning enhances syntactic sensitivity in low-resource settings.

3.3 Attention patterns analysis

Prior work has shown that attention maps in models like BERT can capture meaningful linguistic patterns (see Clark et al., 2019; DeRose, Wang, & Berger 2020; Vig, 2019; Park et al., 2019; for different views, see Jain &Wallace, 2019; Mohankumar et al., 2020; Serrano & Smith, 2019; Thorne et al., 2019). Thus, to investigate how the models represent syntactic dependencies in Korean, we conducted an attention analysis focusing on self-attention weights originating from the critical first verb (VERB1) in each sentence. This verb corresponds to the embedded predicate, where honorific agreement is morphologically marked by *-si-*. The goal of the analysis was to determine whether the model allocates greater attention to the syntactically appropriate subject—either the main clause subject (NP1) in NP1-control sentences or a structurally lower NP (NP2) in NP2-control sentences—when computing the contextual representation of VERB1. We extracted attention weights from VERB1 to the two potential subjects, NP1 and NP2. These values reflect the extent to which VERB1 attends to information provided by each NP, serving as an indirect measure of which constituents the model treats as syntactically or semantically relevant.

Because KoBERT and KoGPT-2 differ in architecture and implementation, raw attention weights are not directly comparable across models. To allow for meaningful comparison, we analyzed normalized attention ratios, which capture the relative distribution of attention across potential antecedents (NP1 vs. NP2). Attention weights were extracted during the models' evaluation of sentence grammaticality at both Epoch 1 (pre-trained) and

Epoch 30 (fine-tuned). For each trial, we computed the attention ratio for NP1 as the proportion of attention allocated to NP1 relative to the total attention directed to NP1 and NP2 (i.e., NP1 Ratio = NP1 Attention / [NP1 Attention + NP2 Attention] × 100), and likewise for NP2. This normalized, directional metric allows us to assess attention patterns independently of differences in absolute attention scale. To quantify relative attentional preference, we then calculated an attention difference score for each trial by subtracting the NP2 ratio from the NP1 ratio (i.e., NP1 Ratio – NP2 Ratio). Accordingly, positive values indicate a preference for NP1, while negative values reflect greater attention to NP2.

To explore how attention patterns relate to classification outcomes, we selected 400 sentences that were correctly classified by both models and 56 that were misclassified by both. These subsets formed the basis of our attention analysis. Transformer models compute attention matrices over 12 layers and 12 heads based on tokenized inputs. Because tokenization significantly affects the model's interpretation of sentence structure, it plays a critical role in attention analysis. In this study, we used a syllable-based tokenizer, which occasionally led to token splitting inconsistencies due to whitespace handling. To enhance interpretability while preserving the basic clause structure [NP1 NP2 VERB1 VERB2], we adopted a modified version of the method proposed by Mun and Shin (2025), as illustrated in Appendix A.

3.4 Statistical analyses[1]

Classification accuracy was analyzed using generalized linear mixed-effects models (GLMER) with a binomial link, implemented in R 4.2.1 (R Core Team, 2024) via the lme4 package (Bates et al., 2015, v1.1-31). P-values were computed using *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017, version 3.1-3). Attention data were analyzed using linear mixed-effects regression (LMER). Classification accuracy models included four fixed effects: Model (KoBERT vs. KoGPT-2), Control Type (NP1 vs. NP2), and the Honorific features of NP1 and NP2 (H vs. NH), along with all interactions. Attention models included Model, Control Type, and Classification Accuracy (correct vs. incorrect) as fixed effects. To evaluate the impact of fine-tuning on syntactic sensitivity, we compared Epoch 1 (pre-trained) and Epoch 30 (fine-tuned) performance using models with Model, Epoch, and Control Type as predictors. All predictors were sum-coded. Random intercepts for sentence were included to account for trial-level variability. Random intercepts for sentence were included, and maximal random-effects structures (Barr et al., 2013) were simplified for convergence. Holm-adjusted pairwise comparisons were conducted using the *emmeans* package (Lenth, 2024, v1.8.4-1).

3.5 Results and Discussion

Classification Accuracy Before and After Fine-tuning (Epoch 1 vs. Epoch 30). Appendix B presents the grammatical acceptability classification accuracies for KoBERT and KoGPT-2. Corresponding statistical results are in Appendix C. The analysis showed main effects of Model and Control Type: KoGPT-2 (M = 89.7%) significantly outperformed KoBERT (M = 81.1%), and NP2-control sentences (M = 87.6%) were classified more accurately than NP1-control sentences (M = 76.8%). Honorific features of NP1 and NP2 also had main effects, indicating sensitivity to subject–verb honorific agreement. A significant Control Type × NP1 × NP2 interaction revealed that classification accuracy varied by agreement configuration. When agreement was disrupted by a feature-matching attractor—e.g., NP1-control with non-honorific NP1 and honorific NP2—accuracy dropped sharply (KoBERT: 45.3%; KoGPT-2: 52.3%). In contrast, when the attractor did not linearly intervene—as in NP2-control with honorific NP1 and non-honorific NP2—KoGPT-2 maintained high accuracy (92.4%), correctly identifying these sentences as ungrammatical, whereas KoBERT's accuracy was substantially lower (76.7%). KoGPT-2 significantly outperformed KoBERT across conditions (ps $< .03$), except in NP1-control sentences with honorifiable NP1 and non-honorifiable NP2, where KoBERT (81.3%) outperformed KoGPT-2 (69.5%) ($p < .0001$).

These findings suggest that language models' performance declines with increasing agreement distance and feature-matching attractors that intervene in subject–verb dependencies (Bacon &

[1] All data and analysis code are available at the following link: https://osf.io/fw82j/?view_only=56b9d1dc865e453086dfbc8957fee340

Regier, 2019; Ryu & Lewis, 2021; Lakretz et al., 2022), paralleling patterns in human sentence processing (e.g., Kwon & Sturt 2016, 2019). The results were also consistent with previous studies showing that autoregressive models like GPT-2 are more robust to such interference and better at maintaining long-distance syntactic dependencies, whereas masked language models like BERT are more susceptible to feature-based interference from structurally irrelevant attractors (Chaves & Richter, 2021; Lasri, Lenci, & Poibeau, 2022).

We next examined whether fine-tuning improves the models' sensitivity to syntactic structure. Appendix D presents the classification accuracy of fine-tuned KoBERT and KoGPT-2, by control type and honorific features of NP1 and NP2. Appendix E presents the corresponding statistics, and Figure 1 visualizes aggregated results by model, epoch, and control type.

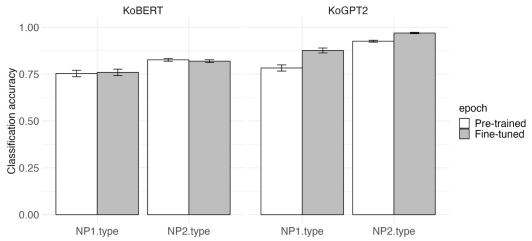


Figure 1. Classification accuracy aggregated across honorific conditions, grouped by model, control type, and training state (pre-trained vs. fine-tuned: Epoch 1 vs. Epoch 30)

Analysis revealed significant main effects of Model, Control type, and Epoch, indicating overall higher accuracy for KoGPT-2 (92.4%) than KoBERT (80.9%), and for NP2-control (88.5%) over NP1-control (79.3%) sentences. Accuracy also improved after fine-tuning (88.0%) compared to the pre-trained state (85.4%). These main effects were qualified by significant interactions, including Model × Control type and Epoch × Model and a three-way Model × Epoch × Control type interaction. Holm-adjusted pairwise comparisons showed that KoGPT-2 exhibited substantial gains after fine-tuning for both NP1-control ($z = -8.21$, $p < .001$) and NP2-control sentences ($z = -9.95$, $p < .001$). In contrast, KoBERT showed no significant improvement for NP1-control ($z = 1.20$, $p = .23$) and even declined in NP2-control accuracy ($z = 4.01$, $p < .001$).

During pre-training, both KoBERT and KoGPT-2 exhibited high error rates when a feature-matching attractor linearly intervened in subject–verb honorific agreement (i.e., NP1-control sentences with a non-honorifiable NP1 and an honorifiable NP2). After fine-tuning, both models showed improved classification accuracy for this construction (KoBERT: 45.3% → 61.8%; KoGPT-2: 52.4% → 72.8%). To evaluate whether fine-tuning reduced attractor interference, we conducted a follow-up analysis. Focusing on NP1-control sentences, we compared conditions in which the honorific features of NP1 and NP2 differed (H–NH and NH–H) to those in which they matched (H–H and NH–NH). If models adhered strictly to syntactic structure, performance should be unaffected by the features of an illicit subject. However, Welch's two-sample *t*-tests revealed that mismatched conditions significantly reduced classification accuracy for both models: H–NH vs. H–H (KoBERT: $t(1089.3) = 5.60$, $p < .001$; KoGPT-2: $t(679.16) = 10.87$, $p < .001$), and NH–H vs. NH–NH (KoBERT: $t(1075.9) = -6.57$, $p < .001$; KoGPT-2: $t(635.37) = -14.25$, $p < .001$). These findings suggest that although fine-tuning substantially improved KoGPT-2's overall accuracy and structural sensitivity, both models remain vulnerable to honorific feature interference in structurally complex NP1-control configurations.

Overall, the results suggest that KoGPT-2 benefits substantially from fine-tuning, showing improved classification accuracy across both control types. In contrast, KoBERT appears less responsive to fine-tuning and even declined in NP2-control performance. This divergence may reflect architectural differences in how the two models encode syntactic dependencies, with KoGPT-2's autoregressive design better supporting structural sensitivity. Nonetheless, both models remain vulnerable to honorific feature interference, especially in NP1-control sentences where structurally irrelevant NPs disrupt subject–verb agreement.

To further examine how syntactic information is internally represented, we next analyze the models' attention patterns during the classification task.

Attention Patterns Before and After Fine-Tuning (Epoch 1 vs. Epoch 30). The attention analysis included 456 trials from the classification task that were either correctly (n = 400) or incorrectly (n = 56) classified by both models. This set comprised 104 NP1-control and 352 NP2-control sentences. By comparing correct and incorrect classifications, we aimed to explore how attention patterns relate to model performance and to identify structurally

relevant attention cues that may support accurate classification.

Because absolute attention weights vary between GPT-2 and BERT due to differences in architecture and implementation, we focused on relative attention distributions. Specifically, we analyzed preference attention ratios, which quantify how attention is distributed between NP1 and NP2, independent of model scale (see Section 3.3). In our analysis, positive values indicate a preference for NP1, while negative values reflect greater attention to NP2. Accordingly, under this metric, a structurally aligned model should yield more positive scores for NP1-control sentences and more negative scores for NP2-control sentences.

Figure 2 presents the attention difference scores by model and classification accuracy for pre-trained KoBERT and KoGPT-2. The statistical analysis results are presented in Appendix F.

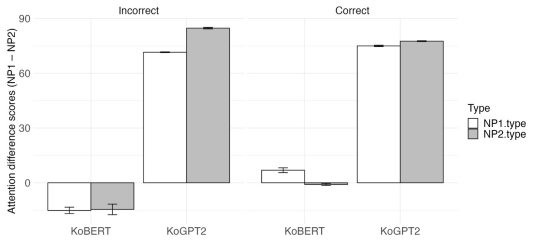


Figure 2. Attention difference scores by model and classification accuracy for KoBERT and KoGPT-2 in their pre-trained states.

The analysis revealed significant main effects of Model and Classification Accuracy: KoGPT-2 had higher attention difference scores (M = 76.8, SE = 0.55) than KoBERT (M = −1.6, SE = 1.48), suggesting that KoGPT-2 allocated more attention to NP1 than KoBERT. Correct trials also yielded higher scores (M = 38.7, SE = 2.84) than incorrect ones (M = 0.29, SE = 8.89). These main effects were moderated by significant interactions: Model × Classification accuracy and Model × Control type. Holm-corrected pairwise comparisons indicated that KoBERT allocated more attention to NP2 in incorrect trials than in correct ones ($t(192) = -4.45$, $p < .0001$), whereas KoGPT-2 did not show a comparable difference ($t(192) = 0.45$, n.s.). In addition, KoGPT-2 allocated more attention to NP1 in NP2-control than NP1-control sentences ($t(192) = -1.97$, $p = .05$). In contrast, KoBERT did not exhibit such an asymmetry ($t(192) = 0.90$, n.s.).

These findings suggest that, in the pre-trained state, attention allocation patterns do not consistently reflect syntactic roles, despite decent classification accuracy. KoGPT-2 directed significantly more attention to NP1 in NP2-control sentences, where NP2 is the syntactically appropriate subject. This indicates a misalignment between attention and the underlying grammatical dependencies. KoBERT's attention favored NP2 in incorrect trials, likely reflecting interference rather than accurate subject identification. Overall, these results highlight that attention distributions in pre-trained models may not consistently indicate syntactic understanding. This aligns with prior work demonstrating weak links between attention weights and model performance or reasoning processes (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019).

Having established baseline attention patterns in the pre-trained models, we next examined how fine-tuning affected attention allocation in KoBERT and KoGPT-2. Using the same statistical models, we evaluated whether fine-tuning improved the alignment between attention and syntactic roles. Figure 3 presents the attention difference scores (NP1 – NP2) by model, classification accuracy, and control type for the fine-tuned models. The corresponding statistical analysis results are shown in Appendix G.

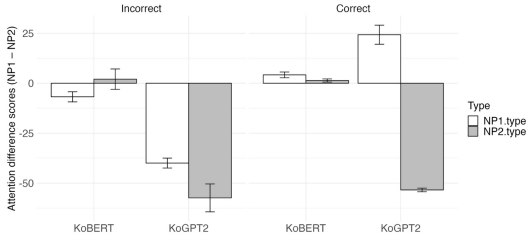


Figure 3. Attention difference scores by model and classification accuracy for fine-tuned KoBERT and KoGPT-2

The analysis revealed significant main effects of Model and Classification accuracy. KoGPT-2 showed lower attention difference scores (M = −41.9, SE = 3.65) than KoBERT (M = 0.95, SE = 1.91), and correctly classified trials yielded higher scores (M = −19.9, SE = 2.72) than incorrectly classified ones (M = −24.3, SE = 6.11). A significant main effect of Control type also emerged, with NP1-control sentences eliciting higher scores (M = −1.69, SE = 5.56) than NP2-control sentences (M = −26.1, SE = 2.66). These main effects were qualified by significant interactions: Model × Control type, Model × Classification accuracy, Control type ×

Classification accuracy, and a three-way interaction of Model × Control type × Classification accuracy. Holm-corrected comparisons showed that KoGPT-2 allocated significantly more attention to NP1—the correct subject—in correctly classified NP1-control sentences than in incorrect ones (t = −6.40, p < .0001), indicating greater syntactic alignment. No such effect was observed for KoGPT-2 in NP2-control sentences (t = −0.27, p = .789) or for KoBERT in either condition (ps > .78).

These results after fine-tuning shed light on the interpretability of attention in Transformer models, particularly regarding their sensitivity to syntactic roles. The significant three-way interaction among model type, control type, and classification accuracy suggests that attention allocation is not uniformly predictive of classification performance but is modulated by both structural configuration and model architecture. Notably, KoGPT-2 showed a strong link between attention and classification accuracy in NP1-control sentences after fine-tuning, suggesting improved syntactic alignment. This effect was absent in NP2-control sentences, likely due to the small number of classification errors (n = 6), which limited statistical power. In contrast, KoBERT's attention was unaffected by classification accuracy following fine-tuning, aligning with its limited performance gains during training. This suggests that KoBERT's attention patterns may be less sensitive to syntactic structure or more weakly coupled with task success, echoing previous findings that attention weights do not always reflect meaningful model behavior (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019).

Taken together, these findings highlight the importance of evaluating attention behavior in relation to both model architecture and syntactic structure. They also caution against interpreting attention weights as direct indicators of linguistic competence: while attention can reflect syntactic alignment in some cases, this is not a reliable property across models or sentence types. KoBERT's lack of attention modulation despite high classification accuracy raises questions about the interpretability of attention in masked language models. We return to this issue in the general discussion.

4 General Discussion and Conclusion

This study aimed to investigate whether Transformer-based language models can encode abstract syntactic structures, even in the absence of an explicit layer dedicated to syntactic representation. We addressed this question using Korean, a low-resource language that differs typologically from widely studied English, by examining model behavior through the lens of subject–verb honorific agreement. Although honorifics may appear socio-pragmatic in nature, the use of the honorific suffix *-si-* requires a licensing subject with matching features in a structurally appropriate position. Accordingly, -si- honorific agreement provides a unique testing ground for evaluating whether language models can acquire abstract syntactic dependencies in the absence of categorical surface cues. To this end, we employed both a grammaticality classification task and a self-attention analysis to evaluate the performance of two Korean-language models, KoBERT and KoGPT-2, in both their pre-trained states and after fine-tuning (i.e., before and after exposure to honorific agreement patterns in Korean). The training and classification datasets included sentences featuring subject–verb honorific agreement that varied in control type and in the honorific features of the potential subject NPs.

Our results offer three key contributions to the study of syntactic generalization in neural language models. First, although both models achieved relatively strong classification performance—suggesting some degree of syntactic understanding consistent with prior work (Clark et al., 2019; Hewitt & Manning, 2019; Lin, Tan, & Frank, 2019)—KoGPT-2 consistently outperformed KoBERT and responded more robustly to fine-tuning (Chaves & Richter, 2021). Even in their pre-trained states, KoGPT-2 generally outperformed KoBERT. While KoBERT showed only modest gains—or even declines—in performance after fine-tuning, KoGPT-2 improved significantly across both NP1- and NP2-control conditions. Both models struggled with attractor interference, but KoBERT appeared especially susceptible. In the syntactically complex NP1-control condition—where the attractor NP2 linearly intervenes between the subject (NP1) and the verb—KoBERT's pre-trained accuracy was only 45.3%, compared to 52.4% for KoGPT-2. After fine-tuning, KoGPT-2's performance rose to 72.8%, while

KoBERT reached only 61.8%. These findings support prior claims that autoregressive models like GPT-2 are better equipped to track hierarchical dependencies than masked language models like BERT (Chaves & Richter, 2021; Lasri, Lenci, & Poibeau, 2022). In contrast, BERT may rely more heavily on shallow heuristics, rendering it more susceptible to lexical interference (McCoy, Pavlick, & Linzen, 2019). Importantly, our use of a low-resource language extends these insights beyond high-resource settings yielding critical evidence about models' capacity to generalize over abstract syntactic structure.

Second, our attention analyses contribute empirical evidence to the ongoing debate about the interpretability of attention mechanisms in Transformer-based models. Specifically, our results support previous claims that attention weights do not reliably correspond to linguistic explanations (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019; Zhao et al., 2024) even though alignment does emerge in some cases. For KoGPT-2, fine-tuning led to greater alignment between attention allocation and syntactic roles, but only under specific conditions. In correctly classified NP1-control sentences, KoGPT-2 consistently directed greater attention to NP1, the structurally appropriate subject. In contrast, in misclassified NP1-control trials, attention shifted toward NP2, suggesting that syntactically guided attention supports successful classification. However, this relationship is not guaranteed. For instance, KoBERT showed no such modulation: its attention patterns remained largely insensitive to syntactic structure or classification outcome. Notably, KoBERT's classification accuracy plateaued across training (~81%) and even declined in certain conditions after finetuning—mirroring its lack of syntactic alignment in attention, despite relatively strong overall performance. These results reinforce growing concerns that attention weights, while useful in some contexts, may not consistently reflect underlying grammatical knowledge or task-relevant reasoning. They also underscore the importance of jointly evaluating both model performance and internal interpretability when assessing syntactic generalization in neural language models (Jain and Wallace 2019).

Third, our results underscore the utility of Korean honorific agreement as a rigorous diagnostic for evaluating syntactic sensitivity in language models. Unlike number agreement in English, Korean subject–verb honorific agreement is governed by structural licensing conditions that are not always transparent at the surface level. The optionality and morphosyntactic specificity of -si-honorific agreement allow researchers to probe whether models can move beyond surface-level lexical heuristics and encode deeper grammatical generalizations. On the other hand, the fact that both KoBERT and KoGPT-2 remained susceptible to interference from honorific attractors even after finetuning suggests that fully abstracting over syntactic structure—particularly in constructions involving long-distance dependencies and feature checking—remains a significant challenge for current Transformer architectures.

Taken together, our findings demonstrate that Transformer-based language models can acquire sensitivity to morphosyntactic dependencies, even when these dependencies are tied to socio-pragmatic cues such as honorifics. While both models performed reasonably well in their pre-trained states, fine-tuning—especially for KoGPT-2—led to notable improvements in classification accuracy and more syntactically aligned attention patterns in structurally complex sentences. However, these gains were not uniform. KoGPT-2's attention aligned with syntactic roles only under certain conditions, and KoBERT showed little evidence of syntactically guided attention despite moderate classification accuracy.

These results underscore the limits of using attention as indicators of grammatical knowledge. Attention may sometimes reflect syntactic reasoning, but it does not reliably track structural representations across models or constructions (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019). This study underscores the value of cross-linguistic research, especially with low-resource languages like Korean, whose rich morphology and flexible word order can reveal model limitations obscured in English-centric evaluations (cf. Chang & Bergen, 2024; Wu & Dredze, 2020).

At the same time, it is important to acknowledge certain limitations of our study. In particular, our analysis did not fully address how implementation details might have influenced the results. Factors such as tokenization (e.g., the proportion of [UNK] tokens), differences in model size (KoBERT and KoGPT-2 differ not only in inference type but also in parameter scale), and the characteristics of the

pretraining corpora (e.g., written vs. spoken language styles) could all have contributed to the observed outcomes. To minimize potential confounds, all test sentences were lexically matched across the experimental conditions, with the only differences arising from the honorific features that served as the manipulation. This design should therefore reduce the likelihood that other sentence components drove the observed effects. Nevertheless, we cannot fully exclude this possibility, particularly given differences in the models' pretraining corpora.

Future research should explore whether incorporating explicit syntactic supervision or inductive biases—such as training on treebank-annotated corpora or employing structure-aware architectures—enhances models' ability to generalize robustly and yields more interpretable internal representations, especially across typologically diverse languages.

References

Bacon, G., & Regier, T. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3861–3871).

Barber, H., & Carreiras, M. (2003). Integrating gender and number information in Spanish word pairs: An ERP study. Cortex, 39(3), 465–482.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255–278.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software. https://arxiv.org/abs/1406.5823

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. Computational Linguistics, 50(1), 293–350.

Chaves, R. P., & Richter, S. N. (2021). Look at that! BERT can be easily distracted from paying attention to morphosyntax. In Proceedings of the Society for Computation in Linguistics 2021 (pp. 28–38). Association for Computational Linguistics.

Chomsky, N. (1981). Lectures on government and binding. Foris.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In Proceedings of the 2019 ACL Workshop on Blackbox NLP (pp. 276–286).

DeRose, J. F., Wang, J., & Berger, M. (2020). Attention flows: Analyzing and comparing attention mechanisms in language models. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1160–1170.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 NAACL-HLT (pp. 4171–4186).

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. CoRR, arXiv:1901.05287.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the NAACL-HLT (pp. 4129–4138).

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434–446.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 NAACL-HLT (pp. 3543–3556).

Jeon, H., Lee, D., & Park, J. (2019). Korean BERT pre-trained cased (KoBERT). SK Telecom AI Center. https://github.com/SKTBrain/KoBERT

Jeon, H. (2021). KoGPT2 ver 2.0. Hugging Face. https://huggingface.co/skt/kogpt2

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26.

Kwon, N., & Sturt, P. (2016). Attraction effects in honorific agreement in Korean. Frontiers in Psychology, 7, 1302.

Kwon, N., & Sturt, P. (2019). Proximity and same case marking do not increase attraction effect in comprehension: Evidence from eye-tracking experiments in Korean. Frontiers in Psychology, 10, 1320.

Kwon, N., & Sturt, P. (2024). When social hierarchy matters grammatically: Investigation of the processing of honorifics in Korean. Cognition, 251, 105912.

Kwon, N., & Polinsky, M. (2006). Object control in Korean: Structure and processing. Japanese/Korean Linguistics, 15, 249–262.

Lakretz, Y., Desbordes, T., Hupkes, D., & Dehaene, S. (2022). Can transformers process recursive nested constructions, like humans? In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3226–3232).

Lasri, C., Lenci, A., & Poibeau, T. (2022). Syntactic generalization and lexical heuristics in transformer-based language models. In Proceedings of the 60th Annual Meeting of the ACL (pp. 3703–3717).

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means (Version 1.10.0) [R package]. https://CRAN.R-project.org/package=emmeans

Lin, Y. C., Tan, Y. C., & Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In Proceedings of the 2019 ACL Workshop on Blackbox NLP (pp. 241–253).

McCormick, C. (2019). BERT fine-tuning tutorial with PyTorch. https://mccormickml.com/2019/07/22/BERT-fine-tuning/

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th ACL (pp. 3428–3448).

Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., & Ravindran, B. (2020). Towards transparent and explainable attention models. In Proceedings of the 58th ACL (pp. 4206–4216).

Mun, S., & Shin, G.-H. (2025). Polysemy interpretation and transformer language models: A case of Korean adverbial postposition -(u)lo. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 1555–1561).

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. Journal of Memory and Language, 34(6), 739–773.

Park, C., Na, I., Jo, Y., Shin, S., Yoo, J., Kwon, B. C., Zhao, J., Noh, H., Lee, Y., & Choo, J. (2019). SanVis: Visual analytics for understanding self-attention networks. In IEEE VIS (pp. 146–150).

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (pp. 61–71).

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? In Proceedings of the 57th ACL (pp. 2931–2951).

Sohn, H.-M. (2001). The Korean language. Cambridge University Press.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2019). Generating token-level explanations for natural language inference. In Proceedings of the NAACL-HLT (pp. 963–969).

Vig, J. (2019). Visualizing attention in transformer-based language representation models. arXiv preprint, arXiv:1904.02679.

Vázquez, R., et al. (2020). A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. Computational Linguistics, 46(2), 387–424.

Wilcox, E. G., Futrell, R., & Levy, R. (2024). Using computational models to test syntactic learnability. Linguistic Inquiry, 55(4), 805–848.

Wolf, T., et al. (2019). Transformers: State-of-the-art natural language processing. arXiv preprint, arXiv:1910.03771.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? In Proceedings of the 5th Workshop on Representation Learning for NLP (pp. 120–130).

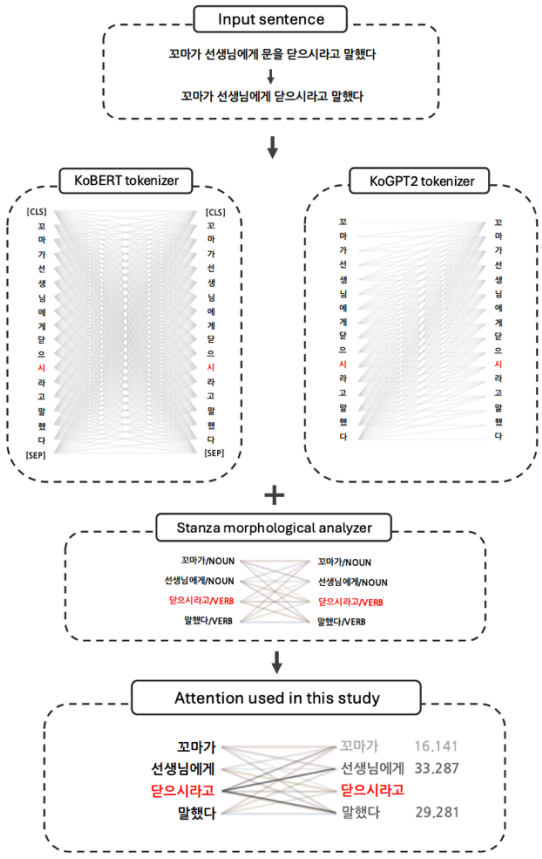
Wu, Y., et al. (2019). A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Computational Linguistics, 45(1), 163–197.

Yoon, J. (2009). The distribution of subject properties in multiple subject constructions. In Japanese/Korean Linguistics (Vol. 19, pp. 64–83).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), Article 20.

Appendices

Appendix A. N/V unit treatment and attention transformation (Example (5))



Appendix B. Classification accuracy results for pre-trained KoBERT and KoGPT-2 on target sentences involving subject–verb honorific agreement. Results are grouped by the honorific features of the two potential subject NPs (NP1 and NP2) and by sentence control type (NP1-control vs. NP2-control).

KoBERT				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	89.1 (0.01)	86.5 (0.01)	81.2 (0.01)
	NH	81.3 (0.02)	76.8 (0.01)	
NH	H	45.3 (0.02)	82.8 (0.01)	
	NH	84.2 (0.02)	84.6 (0.01)	
KoGPT2				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	93 (0.01)	94.3 (0.01)	89.8 (0.01)
	NH	69.5 (0.02)	92.5 (0.01)	
NH	H	52.4 (0.02)	86.6 (0.01)	
	NH	98.4 (0.01)	97 (0.01)	

Appendix C. Linear Mixed Effects model results for classification accuracy by pre-trained KoBERT and KoGPT-2. Coefficients, standard errors, z-values, and p-values are reported for main effects and their interactions. The model included random intercepts for sentence. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***z***	***p***
(Intercept)	2.281	0.05	45.43	.001
NP1	0.153	0.04	3.95	.001
NP2	-0.176	0.05	-3.51	.001
Model	-0.481	0.04	-12.86	.001
Control type	-0.377	0.04	-9.86	.001
NP1:NP2	0.821	0.04	21.16	.001
NP1:Model	0.176	0.04	4.75	.001
NP2:Model	0.159	0.04	4.27	.001
NP1:Type	0.173	0.04	4.53	.001
NP2:Type	-0.249	0.04	-6.53	.001
Model:Type	0.131	0.04	3.52	.001
NP1:NP2: Model	-0.285	0.04	-7.67	.001
NP1:NP2: Type	0.446	0.04	11.71	.001
NP1:Model: Type	0.223	0.04	5.98	.001
NP2:Model: Type	-0.067	0.04	-1.8	0.08
NP1:NP2: Model:Type	-0.145	0.04	-3.91	.001

Appendix D. Classification accuracy results for fine-tuned KoBERT and KoGPT-2 on target sentences involving subject–verb honorific agreement. Results are grouped by the honorific features of the two potential subject NPs (NP1 and NP2) and by sentence control type

KoBERT				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	86.5 (0.01)	86.1 (0.01)	80.8 (0.01)
	NH	74.1 (0.02)	79.2 (0.01)	
NH	H	61.8 (0.02)	84.4 (0.01)	
	NH	79.6 (0.02)	78.5 (0.01)	
KoGPT2				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	98.4 (0.01)	98.8 (0.01)	95.1 (0.01)
	NH	79.6 (0.02)	97.1 (0.01)	
NH	H	72.8 (0.02)	94.4 (0.01)	
	NH	99.4 (0.01)	97.8 (0.01)	

Appendix E. Linear mixed-effects model results for classification accuracy of KoBERT and KoGPT-2 before and after fine-tuning (Epoch 1 vs. Epoch 30). Coefficients, standard errors, *z* and *p*-values are reported for the main effects, as well as for their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***z***	***p*** <
(Intercept)	3.807	0.1	37.56	.001
Epoch	0.725	0.08	8.76	.001
Model	-2.182	0.1	-21.8	.001
Control type	-0.582	0.05	-11.46	.001
Epoch:Model	-0.905	0.08	-10.9	.001
Epoch:Type	-0.056	0.05	-1.08	0.29
Model:Type	0.321	0.05	6.33	.001
Epoch:Type	0.127	0.05	2.46	0.02

Appendix F. Linear mixed-effects model results for attention difference scores in the pre-trained state. Coefficients, standard errors, t-values, and p-values are reported for the main effects and their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***t***	***p*** <
(Intercept)	35.625	1.66	21.41	.001
Model	-41.555	1.12	-37.2	.001
Control type	-1.07	1.66	-0.64	0.52
Classification accuracy	-4.005	1.66	-2.41	0.02
Model:Type	2.881	1.12	2.58	0.02
Model: Accuracy	-4.922	1.12	-4.41	.001
Type: Accuracy	-2.368	1.66	-1.42	0.16
Model:Type: Accuracy	0.285	1.12	0.26	0.8

Appendix G. Linear mixed-effects model results for attention difference scores in the fine-tuned state. Coefficients, standard errors, t-values, and p-values are reported for the main effects and their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***t***	***p*** <
(Intercept)	-15.72	3.54	-4.45	.001
Model	15.91	2.75	5.79	.001
Control type	11.13	3.54	3.15	0.01
Classification accuracy	-9.81	3.54	-2.78	0.01
Model: Control type	-12.64	2.75	-4.6	.001
Model: Accuracy	7.24	2.75	2.63	0.01
Type: Accuracy	-8.99	3.54	-2.54	0.02
Model:Type: Accuracy	6.08	2.75	2.21	0.03

Towards Improving Low-Resource Machine Translation with Lightweight Training and Synthetic Data: Case Study of Vietnamese - Khmer

Trong Huy Nguyen[1], Thanh Huong Le[1] *, Que Nhu Tran[2],
[1]Hanoi University of Science and Technology, Vietnam
[2]Foreign Trade University, Vietnam
Huy.NT210451@sis.hust.edu.vn, huonglt@soict.hust.edu.vn,
trannhuqueht@gmail.com

Abstract

Neural machine translation for low-resource languages remains challenging due to the scarcity of high-quality parallel corpora and community support. Under the view of Vietnamese-Khmer (Vi-Km) translation task, we focus on developing a comprehensive, lightweight training pipeline that minimizes reliance on extensive parallel data and large-scale model architecture, while achieving substantial performance improvements. Then, we introduce a three-stage training strategy built upon a small pre-trained language model: (1) Small-scale Continual Pre-training on monolingual data, (2) Supervised Fine-Tuning on parallel datasets and synthetic data generated through a novel data augmentation framework, and (3) Direct Preference Optimization leveraging a newly constructed preference-ranking dataset guided by LLMs. Experimental results on the VOV dataset demonstrate that our model significantly outperforms by up to 8% in both directions on BLEU & METEOR scores over strong LLMs, and other commercial systems. These findings confirm the effectiveness of our approach for enhancing low-resource machine translation systems under low-cost computation resources and the reproducibility for other underrepresented languages.

1 Introduction

In the modern era, with thousands of languages worldwide, machine translation (MT) has become a core focus of Natural Language Processing (NLP) research. With the recent advent of pre-trained language models (PLMs) with the self-attention mechanism (Vaswani et al., 2017), machine translation quality has seen significant improvement, even in low-resource languages. At the same time, large language models (LLMs) have simplified the modeling of complex grammatical and semantic relationships between languages, resulting in more fluent and coherent translations (Zhu et al., 2024). However, both PLMs and LLMs rely on massive, high-quality corpora and computational budgets that are rarely available for extremely low-resource languages. Current research aims to address the challenges by developing a training methodology that minimizes dependence on extensive data and human resources while retaining the strengths of precedent approaches. Meanwhile, some of the latest research is shifting towards data augmentation by leveraging available monolingual data (Sennrich et al., 2016; Liu et al., 2021) or generating synthetic bilingual data. Another common method is translating through a pivot language (Kim et al., 2019). However, these approaches still face limitations in data quality, scalability, and computational requirements (Edunov et al., 2020).

Given the limited resources available for the case of Vi-Km translation and the impracticality of deploying large models in resource-constrained environments, there is a compelling need for an efficient, lightweight, and scalable training pipeline tailored to this specific task. Furthermore, existing few-shot prompting methods on LLMs struggle to deliver competitive results for this pair, emphasizing the importance of task-specific model adaptation and domain-specific fine-tuning.

The problem we address is how to leverage small language models with available datasets using an effective training approach while ensuring compatibility with limited computation infrastructure and involvement of human annotators. We propose a novel comprehensive approach for low-resource MT, with the following key contributions:

- A lightweight training pipeline, consisting of Small-scale Continual Pre-training (Sm-CPT), Supervised Fine-Tuning (SFT), and Direct Preference Optimization (DPO) (Rafailov et al., 2023), optimized for small language models on limited computational resources.

* Corresponding author: huonglt@soict.hust.edu.vn

- A scalable synthetic parallel data generation framework leveraging powerful LLMs with filtering and ranking mechanisms to construct high-quality training data, including preference-ranking datasets for DPO, without manual annotation.

- An extensive empirical study on the VOV dataset (Nguyen et al., 2022), achieving the highest results on BLEU and METEOR scores over previous works, commercial translation systems, and LLMs.

This study provides practical insights and scalable solutions for improving low-resource MT and paves the way for extending these techniques to other under-supported language pairs in the region.

2 Related Works

In the scope of addressing low-resource MT, one notable approach is pre-training a multilingual model using Transformer Encoder-Decoder (Xue et al., 2021). The model is then fine-tuned for many specific languages, showing prominent performances as in the cases of Kazakh-Russian and Russian-Tatar (Kozhirbayev, 2024). Data augmentation is also considered an effective approach. Sennrich et al. (2016) proposed back-translation, involving translating target monolingual data into the source language using a reverse translation model. The synthetic parallel data generated through back-translation is subsequently employed to train the forward translation model. This method has shown improvements in low-resource NMT, notably in the Vi-Km (Pham Van and Le Thanh, 2022; Quoc et al., 2023) research. Besides, employing paraphrase embedding and POS-Tagging is an efficient approach to augment data for machine translation by paraphrasing sentences at the word level (Maimaiti et al., 2021). However, these approaches have revealed several limitations, especially the loss of contextual information (Edunov et al., 2020). Using pivot languages is another solution to the challenges of machine translation proposed by Kim et al. (2019). The selected pivot language data is used to train a cross-lingual encoder and auto-decoder, which is then fine-tuned to produce the final translation model. Kim et al. (2019) has demonstrated the effectiveness of this method in the French-German and German-Czech translation tasks in the WMT 2019. However, this approach for low-resource machine translation has several disadvantages; notably, it often leads to the propagation of circular translation errors (Kementchedjhieva and Søgaard, 2023).

KC4MT (Nguyen et al., 2022) provides a high-quality Vi-Km bilingual dataset curated from Voice of Vietnam (VOV) news content. It was developed under a national project and validated by language experts proficient in both Vietnamese and Khmer. Besides the Vi-Km bilingual dataset, this project released an additional monolingual Khmer subset. In addition to KC4MT, the TED (Reimers and Gurevych, 2020) dataset offers a monolingual source for Vietnamese and Khmer. TED includes transcripts from over 4,000 TED and TEDx talks translated by a global volunteer community into more than 100 languages. QED (Abdelali et al., 2014), developed by the Qatar Computing Research Institute, provides multilingual subtitles from educational content across various STEM topics. While these corpora offer valuable resources, previous studies have noted challenges related to the presence of non-linguistic noise, such as special characters or inconsistent formatting.

3 Methodology

3.1 Backbone Model Selection

In this study, we leverage a pre-trained language model that offers strong multilingual support for both Vietnamese and Khmer. Among the available models, SeaLLMs (Nguyen et al., 2024) stands out as a family of large language models specifically designed for Southeast Asian (ASEAN) languages. These models are fine-tuned and optimized to enhance accessibility and performance for low-resource regional languages across ASEAN. The third version, `SeaLLMs-v3`, offers two new variants of size: 7B and 1.5B. Compared to its predecessors, `SeaLLMs-v3` is pre-trained on a massive corpus comprising general-domain and region-specific data, including sources such as Wikipedia, CC-News, CulturaX, and MADLAD-400 (Nguyen et al., 2024; Kudugunta et al., 2023). It also incorporates synthetic and translated data in training to improve support for under-supported languages in the region.

In this work, we select `SeaLLMs-v3-1.5B-Chat` (Nguyen et al., 2024) as our baseline model, a further instruction-tuned version of the 1.5B base model. This choice is motivated by its balance between model architecture and computational efficiency due to its small size, offering an effective

foundation for our experiments.

3.2 Training Methodology

To address the challenges of MT for the low-resource Vi–Km language pair, we propose a multi-stage training framework tailored for a lightweight model architecture. This approach integrates several optimization techniques to balance performance and computational efficiency. The overall framework is illustrated in Figure 1, consisting of three sequential stages:

1. **Small-scale Continual Pre-training phase**: The baseline model is continually pre-training on monolingual data using an adapter-based architecture, enabling better language modeling capacity specific to the target languages.
2. **Supervised Fine-Tuning phase**: The continual pre-trained model is then fine-tuned on a mixture of real and synthetic parallel data to enhance translation quality under limited supervision.
3. **Direct Preference Optimization Training phase**: The model generates multiple candidate translations from the source side of a gold-standard parallel dataset. These candidates are then ranked by a large language model (LLM), and the resulting preference pairs are used to train the model with a direct preference optimization objective.

For adapter type selection, we revisit two low-rank adapter variants—LoRA (Hu et al., 2022) and AdaLoRA (Zhang et al., 2023) by fine-tuning on the original VOV Vi-Km dataset. The adapter configuration that yields the highest validation performance is carried forward to all subsequent stages. After completing the full training pipeline for the Vi-Km translation direction, we reapply the same experimental stages to the reverse direction (Km-Vi). The adapter configuration, hyperparameters, and datasets are maintained to ensure a fair comparison. This bidirectional evaluation enables a more comprehensive assessment of our pipeline's scalability and its effectiveness applied to another low-resource pair.

For automatic evaluation, we use the BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) metrics, which are widely used in machine translation tasks. As both metrics rely on n-gram overlap, we employ the `khmer-nltk`[1] and `pyvi`[2] for Km and Vi tokenizers, respectively.

3.2.1 Small-scale Continual Pre-training

This stage adapts the baseline PLM to low-resource languages by conducting continual pre-training on monolingual non-labeled data. To ensure computational efficiency, it updates only a small subset of parameters — specifically, the adapter modules, embedding layers, and task-specific heading layer — while keeping the rest of the base model frozen.

This approach addresses the issue of unstable training typically caused by random initialization of adapter parameters, leading to a negative effect on model performance (Nguyen and Nguyen, 2025). The objective, as shown in Equation 1, only updates the adapter parameters ϕ and freezes model parameters θ, while optimizing the negative log-likelihood of the next token given the preceding context.

$$\mathcal{L}_{\text{NLL}}(\mathbf{x}, \phi) = -\sum_{t=1}^{T} \log P(\mathbf{x}_t | \mathbf{x}_{<t}; \theta, \phi) \quad (1)$$

The data for SmCPT are extracted from both bilingual and monolingual sources. Let $\mathcal{D}_{\text{bi}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a bilingual dataset of N pairs in train set, where $x_i \in \mathcal{L}_s$ (source language) and $y_i \in \mathcal{L}_t$ (target language). From this, two monolingual subsets are extracted: $\mathcal{D}_s = \{x_i\}$ and $\mathcal{D}_t = \{y_i\}$. Additionally, external monolingual corpora are collected independently for each language, denoted as $\mathcal{D}'_s \subset \mathcal{L}_s$ and $\mathcal{D}'_t \subset \mathcal{L}_t$. The final dataset used for SmCPT is constructed by merging all these sources: $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_t \cup \mathcal{D}'_s \cup \mathcal{D}'_t$. This composition enables the model to benefit from both translation-aligned data and naturally occurring monolingual text during continual pre-training.

3.2.2 Supervised Fine-Tuning

After SmCPT, the model is fine-tuned using the parallel dataset to further adapt to the translation task. Besides, we also propose a synthetic data generation pipeline utilizing power from a strong LLM: `GPT-4o` to enrich the parallel dataset. The synthetic dataset is mixed with the training set of the VOV dataset to supervise fine-tuning the model. Here, we use a fixed prompt during the training translation task, and the representations of the prompt were not updated in the loss function (Zhang et al., 2024). The loss function is represented in Equation

[1] https://github.com/VietHoang1512/khmer-nltk

[2] https://github.com/trungtv/pyvi

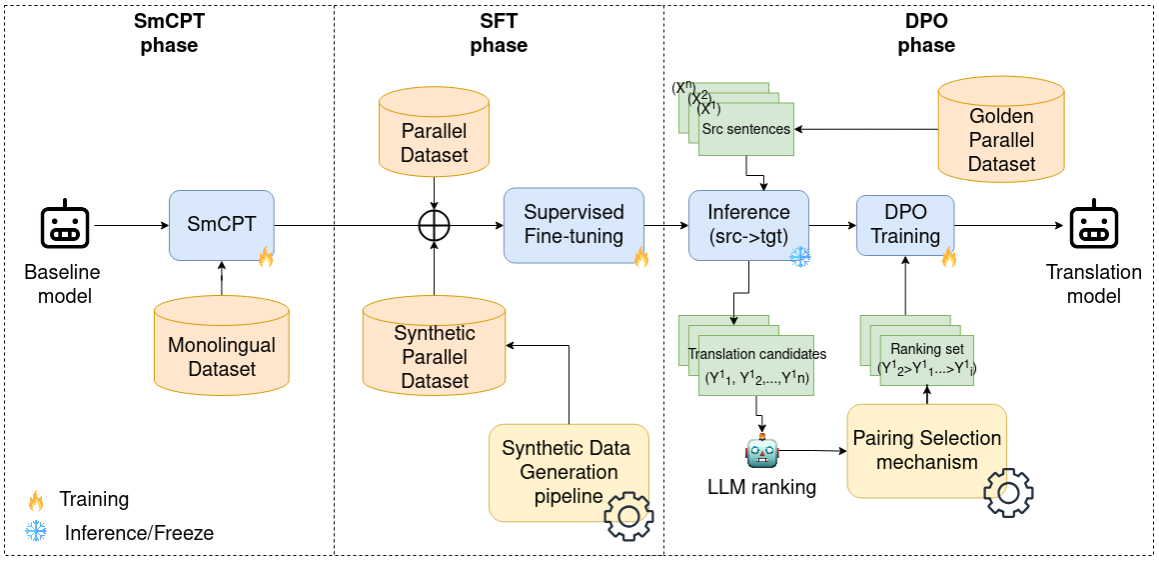


Figure 1: Our proposed training pipeline

2, in which the model parameters ϕ are updated based on the negative log-likelihood loss of the next token prediction and frozen the fixed prompt $\mathcal{I}$ and the base θ.

$$\begin{aligned}\mathcal{L}_{\text{NLL}}(\mathbf{x}, \mathbf{y}, \phi) &= -\log P(\mathbf{y}|\mathbf{x}; \mathcal{I}, \theta, \phi) \\ &= -\sum_{t=1}^{T} \log P(y_t|y_{<t}, \mathbf{x}; \mathcal{I}, \theta, \phi)\end{aligned} \tag{2}$$

3.2.3 Direct Preference Optimization

The fine-tuned model generates candidate Khmer translations from a golden subset of 2,000 Vietnamese sentences in the bilingual datasets of the VOV. These candidates are then ranked by `GPT-4o` with the provided target Khmer sentence serving as 'golden translation' to create a preference dataset in the form of pairing samples, which is inspired by the idea of *self-knowledge* (Yang et al., 2024), the student learns from its outputs and preference ranking, evaluated by a teacher model. Experiments by Yuan et al. (2024) show that models are capable of self-alignment via LLM-based judging and training on their generations through iterative DPO training. However, when forming sentence pairs for DPO training, some pairs of responses with nearly the same representations could end up being separated far apart during DPO training, leading to potential bias in the model (Yan et al., 2025). Thus, we also propose a selection mechanism to form a high-quality DPO pair dataset and control the strictness of the loss function based on β hyperparameter, as illustrated in Equation 3 (Rafailov et al., 2023), which is later shown in Appendix B.3. In which, β is a temperature parameter that controls the sensitivity to differences in rewards, $\pi_\theta(y|x)$, which is the probability of the model generating response y given input x and $\pi_{ref}(y|x)$ is the probability of a reference policy (often the referenced model) generating the same response. For each input data sample, there will be 2 representations, y_w and y_l, corresponding to the sample rated as accepted and rejected.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{r_\theta(y_w|x)}{r_\theta(y_l|x)}\right)\right] \tag{3}$$

In which,

$$r_\theta(y|x) = \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

Pairing Selection Mechanism: After ranking the candidate translations, we apply a pairing selection mechanism to create the reference-ranking dataset. For each source sentence in the seed data, there are 5 corresponding translations, and after ranking, we obtain up to 10 $\{y_{accepted} - y_{rejected}\}$ sentence pairs. In total, from the initial set of 2,000 sentences, we generate up to 20,000 samples for DPO training. To address the problem of close representation of preference ranking samples, we utilize the hybrid score of BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) for ranking the pairs as in Equation 4, wherein the $y_{accepted}$ sentence is considered as the reference

sentence and the $y_{rejected}$ sentence serves as the candidate sentence. Finally, the top $-\ p\%$ sentence pairs with the highest hybrid scores are removed from the DPO training dataset.

$$\mathrm{h}(s) = 2 \cdot \frac{\tilde{\mathrm{bleu}}_s \cdot \tilde{\mathrm{meteor}}_s}{\tilde{\mathrm{bleu}}_s + \tilde{\mathrm{meteor}}_s} \tag{4}$$

In the dataset $\mathcal{D}$ consist of $\mathcal{N}$ pairs, for each pair of sentences $s = \{s_{ref}, s_{pred}\}$, in which s_{ref} is the reference sentence, s_{pred} is the prediction sentence.

$$\tilde{\mathrm{bleu}}_s = \frac{\mathrm{bleu}_s - \min_{i \in \mathcal{N}} \mathrm{bleu}_i}{\max_{i \in \mathcal{N}} \mathrm{bleu}_i - \min_{i \in \mathcal{N}} \mathrm{bleu}_i}$$

$$\tilde{\mathrm{meteor}}_s = \frac{\mathrm{meteor}_s - \min_{i \in \mathcal{N}} \mathrm{meteor}_i}{\max_{i \in \mathcal{N}} \mathrm{meteor}_i - \min_{i \in \mathcal{N}} \mathrm{meteor}_i}$$

4 Datasets

4.1 Available Datasets

We utilize three datasets for our experiments, as summarized in Table 1. The main dataset for supervised fine-tuning is the VOV corpus (Nguyen et al., 2022), which contains 135,164 training, 2,000 validation, and 2,000 test sentence pairs. Data are collected from bilingual news articles published by the Voice of Vietnam (VOV). In addition, we use 131,055 Khmer monolingual sentences from various sources to support language modeling tasks.

Dataset	Split	# samples	
		Vi	Km
Bilingual Dataset			
VOV	train	135,164	135,164
	valid	2,000	2,000
	test	2,000	2,000
Monolingual Dataset			
VOV	-	-	131,055
TED	-	353,251	1,066
QED	-	-	344

Table 1: Dataset statistics for our experiments

To ensure the quality of parallel dataset, we apply a two-stage filtering process: (1) Language Detection using Facebook AI's language identification tool (Bojanowski et al., 2017) to remove sentence pairs where the source is not Vietnamese or the target is not Khmer; and (2) Length Filtering, which eliminates pairs that are too short, too long, or have an abnormal length ratio between source and target sentences. After filtering, 135,164 high-quality sentence pairs are retained for fine-tuning.

For monolingual data, we assemble monolingual Vietnamese and Khmer sentences from multiple sources: the VOV datasets (Nguyen et al., 2022), which provides both a parallel Vi-Km subset and a standalone Khmer monolingual subset; the TED monolingual corpora for both languages (Reimers and Gurevych, 2020); and the QED Khmer monolingual corpus (Abdelali et al., 2014). To reduce noise, we apply a three-step pre-processing pipeline: (1) Remove Duplication using the MinHash algorithm; (2) Clean Special Characters such as HTML tags, non-printable ASCII characters, and emoji; and (3) Language Filtering to discard texts not in the target language.

After processing monolingual datasets, we accumulate a total of 760,066 monolingual samples, consisting of 490,426 Vietnamese sentences and 269,640 Khmer sentences.

4.2 Synthetic Data Generation

We propose a synthetic data generation framework, as in Figure 2, leveraging monolingual Vietnamese data, which is more abundant compared to Khmer, to construct high-quality synthetic parallel corpora. Our framework begins with selecting high-quality monolingual Vietnamese sentences using the TF-IDF score. These selected samples are then labeled through black-box knowledge distillation (Yang et al., 2024), in which `GPT-4o` serves as a teacher model to generate corresponding Khmer translations for training a student model. To validate and filter the synthetic data, we employ a back-translation strategy: instead of using the same teacher model, which could introduce circular reasoning due to shared architecture or biases (Wang and Sennrich, 2020), we use `Google Translate` as an independent translation system to back-translate the generated Khmer sentences into Vietnamese. To evaluate the fidelity of the back-translated sentences against the original Vietnamese inputs, we compute a hybrid quality score combining BLEU and METEOR (Equation 4).

4.2.1 Monolingual Data Selection

Among the available Vietnamese monolingual sources, the TED dataset is selected for its simple, domain-aligned sentences, which resemble those in our target training set. As shown in Figure 2, we sampled 30K TED sentences to balance quality and domain coverage, given that our SFT parallel data comprises around 130K examples. Although several selection methods exist—e.g., Cross-Entropy

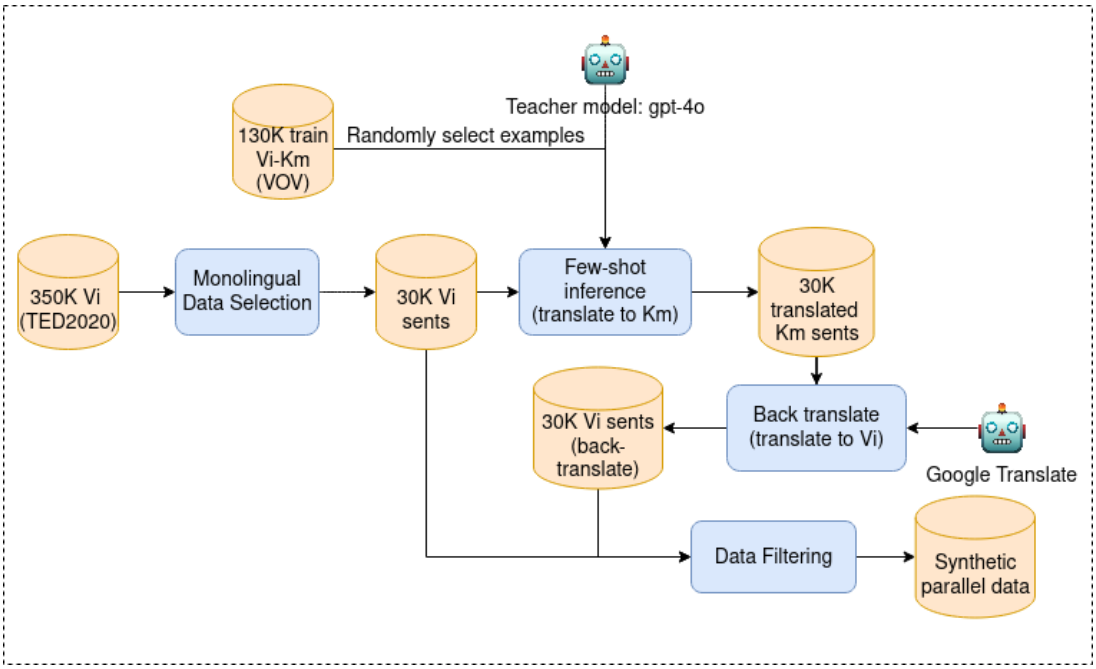


Figure 2: Synthetic data generation pipeline for **Vi-Km** translation task using VOV and TED datasets

Difference (Moore and Lewis, 2010), Feature Decay Algorithm (Poncelas et al., 2018), and TF-IDF; prior work (Silva et al., 2018) shows TF-IDF gains the best performance, and thus we adapt it in this stage. This proportion ensures a balance between input quality and domain alignment, helping to maintain strong performance on the in-domain VOV test set while preserving robustness under out-of-domain distribution shifts.

4.2.2 Synthetic Data Augmentation

We use the VOV bilingual dataset (Nguyen et al., 2022) as seed examples for few-shot prompting. For each of the 30K selected Vietnamese sentences, two randomly sampled examples are injected in the prompt to `GPT-4o` to guide diverse and high-quality Khmer translations.

After generating 30K synthetic Khmer sentences, we apply the back-translation method (Sennrich et al., 2016) to filter low-quality outputs. Specifically, each generated sentence is translated back into Vietnamese using `Google Translate` via the deep-translator library[3]. We compute a hybrid similarity score—defined as a weighted combination of BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) (see Equation 4)—between the original and back-translated Vietnamese sentences. Based on this score, we rank all sentence pairs in descending order and select the top $-\ p\%$ pairs. These original Vietnamese sentences are then matched with their corresponding Khmer outputs from `GPT-4o` to form the final synthetic parallel dataset. We choose not to use the back-translated Vietnamese sentences, as the translation quality tends to degrade after circular translation (Kementchedjhieva and Søgaard, 2023).

[3] https://github.com/nidhaloff/deep-translator

5 Experimental Results

5.1 Evaluating LoRA and AdaLoRA Efficiency on SFT

We evaluate two low-rank adapter methods, LoRA (Hu et al., 2022) and AdaLoRA (Zhang et al., 2023), with different configurations for SFT. In LoRA, the rank hyperparameter r controls the number of trainable parameters, balancing model capacity and efficiency. Meanwhile, in AdaLora, the $init\_r$ hyperparameter defines the initial low-rank dimensionality for each incremental adaptation matrix; conversely, $target\_r$ specifies the desired average rank across all adaptation matrices, directing the iterative pruning mechanism toward a predefined parameter budget and thereby balancing representational richness with computational and memory efficiency throughout training. Results comparing their effectiveness for our task are shown in Table 2.

LoRA consistently outperforms AdaLoRA on both BLEU and METEOR metrics, achieving higher scores of 58.67% and 54.18%, respectively.

Low-rank method	Configuration	%Trainable	Vi-Km BLEU/METEOR
AdaLoRA	$init\_r$: 12, $target\_r$: 8	0.8893	52.91/47.83
	$init\_r$: 20, $target\_r$: 16	1.4734	54.02/48.87
LoRA	r: 16	1.1820	**58.67/54.18**

Table 2: Low-rank adapters supervised fine-tuning results comparison

These results suggest that LoRA may be more effective in modeling the linguistic patterns, while AdaLoRA allows for dynamic rank adjustment and offers greater flexibility in parameter tuning, where its performance remained lower, even with an increased percentage of trainable parameters (1.4734%). This finding emphasizes LoRA's advantage in balancing model adaptation and translation quality. Based on this comparison, we adapt the LoRA configuration ($r = 16$, $\alpha = 32$) for the remainder of our experiments.

5.2 Main Results

Table 3 summarizes performance on both translation directions across our four training stages. The `SeaLLMs-v3-1.5B-Chat` baseline achieves quite low scores, indicating its limited zero-shot capability. In Stage 1, applying supervised fine-tuning (SFT) alone results in a substantial performance boost across both directions, improving BLEU by over 26% for Vi-Km and nearly as much for the reverse direction. Introducing SmCPT before SFT in Stage 2 further enhances translation quality, indicating that language-adaptive pretraining helps the model better align with the translation task. Stage 4 integrates DPO on top of SmCPT and SFT with synthetic preference ranking data. This approach achieves the highest performance, reaching BLEU/METEOR scores of 62.00/58.43(%) for Vi-Km and 57.12/54.61(%) for Km-Vi. These improvements confirm the effectiveness of preference-based alignment in refining model outputs.

5.3 Discussion

We conduct experiments on a diverse set of LLMs with few-shot prompts (0, 1, 2, 4, 8 shots), including GPT models, Gemini-Flash variants, and two commercial MT systems (`Google Translate`[4], and `Microsoft Translator`[5]). In parallel, we evaluate open-source PLM with Encoder-Decoder architecture baselines, including `mBART-50` (Liu et al., 2020), `NLLB-200-Distilled-600M` (Koishekenov et al., 2023).

Few-shot prompting systematically improves translation quality across GPT and Gemini models (see Table 4). Notably, the "mini" variant of these models remains substantially behind their "large" models by a clear margin. At the same time, commercial translation systems (`Google Translate`, `Microsoft Translator`) with direct translation rival many few-shot LLMs, indicating that traditional commercial MT systems remain strong baselines, especially for widely used translation tasks.

Besides, we also experiment with the efficiency of small language models over large-scale general language models on a specific task and language. Table 5 compares the fine-tuned performance of open-source NMT systems alongside previous Km-Vi translation studies. The most noticeable gains arise from fine-tuning open-source models on the domain-specific VOV dataset. Both `mBART-Large-50` and `NLLB-200-Distilled-600M` jump to over-50% BLEU after fine-tuning. One previous work, Pham Van and Le Thanh (2022) fine-tunes mBART50 on VOV and synthetic parallel data generated via back-translation and English-pivot augmentation, reporting a 54.50% BLEU on the Km-Vi VOV test set. Building on that, Quoc et al. (2023) introduce cosine-similarity-based data selection, synthetic candidate generation, and two-step filtering before fine-tuning, achieving 55.37% BLEU in the same translation direction. Meanwhile, Duc et al. (2025) focus on the Vi-Km task with a novel approach using joint-task training with the Question-Answering task, achieving 56.99% BLEU. Our fine-tuned model on the SeaLLMs-v3-1.5B-Chat baseline outperforms all these baselines—both closed-source and open-source—settings with zero-shot translation ability. These results emphasize the necessity of fine-tuning and domain-specific adaptation when leveraging LLMs for machine translation, especially in low-resource language scenarios.

6 Conclusion

This study tackles the challenge of machine translation for low-resource languages, focusing on the Vi-Km pair. We review existing approaches in low-resource machine translation, highlighting their limitations and identifying opportunities for improvement. To address these challenges, we

[4] https://cloud.google.com/translate/docs
[5] https://www.microsoft.com/en-us/translator/

Model/ Training Stage	Training Task	Augment Data	BLEU/METEOR	
			Vi-Km	Km-Vi
SeaLLMs-v3-1.5B-Chat	×	×	32.51/28.02	28.42/25.05
Stage 1	SFT	×	58.67/54.18	54.01/51.41
Stage 2	SmCPT, SFT	×	60.82/56.94	55.64/53.11
Stage 3	SmCPT, SFT	✓	61.13/57.44	56.89/54.12
Stage 4	SmCPT, SFT, DPO	✓	**62.00/58.43**	**57.12/55.61**

Table 3: Comparison of **Vi-Km and Km-Vi** translation performance at each stage

Model	# Few-shot samples	BLEU/METEOR	
		Vi-Km	Km-Vi
GPT-4o	0	49.60/51.54	52.80/51.82
	1	50.20/51.85	53.12/51.99
	2	50.85/52.02	53.55/52.21
	4	53.65/52.98	54.31/52.63
	8	54.45/54.03	55.07/54.34
GPT-4o-mini	0	50.12/44.90	49.34/45.64
	1	50.35/46.55	49.88/47.23
	2	51.06/47.34	51.32/49.22
	4	52.46/49.92	53.21/50.12
	8	53.02/49.33	54.37/52.21
GPT-3.5-turbo	0	43.74/16.30	44.32/20.21
	1	45.54/20.12	46.24/24.34
	2	47.66/23.45	47.85/25.21
	4	48.50/23.05	49.02/26.01
	8	49.00/23.98	49.32/28.88
Gemini-2.0-Flash	0	48.12/42.68	50.12/47.32
	1	49.23/43.54	51.11/48.23
	2	50.21/45.00	51.89/49.37
	4	51.68/48.21	52.32/50.75
	8	53.24/50.12	52.68/51.82
Gemini-1.5-Flash	0	39.72/16.24	40.44/18.21
	1	40.05/17.01	41.42/19.88
	2	40.24/17.44	41.78/20.21
	4	41.24/20.12	42.94/21.42
	8	42.73/22.01	43.94/23.67
Google Translate	×	49.74/50.37	51.21/52.23
Microsoft Translator	×	53.45/52.44	54.32/52.89
Ours (SeaLLMs-v3-1.5B)	×	**62.00/58.43**	**57.12/55.61**

Table 4: Results of different LLMs and software on the **Vi-Km and Km-Vi** translation task

Model/ Previous works	Fine-tune (VOV)	BLEU/METEOR	
		Vi-Km	Km-Vi
mBART-Large-50	×	37.67/8.94	31.41/5.32
	✓	50.56/48.90	52.84/49.98
NLLB-200-Distilled-600M	×	38.08/16.54	32.48/15.52
	✓	50.33/49.20	51.97/49.51
Pham Van and Le Thanh (2022) (mBART-50)	✓	-/-	54.50/-
Quoc et al. (2023) (mBART-50)	✓	-/-	55.37/-
Duc et al. (2025) (Sealion-3B (Singapore, 2024))	✓	56.99/-	-/-
SeaLLMs-v3-1.5B	×	32.51/28.02	28.42/25.05
Ours (SeaLLMs-v3-1.5B)	✓	**62.00/58.43**	**57.12/55.61**

Table 5: Results of different open-source models and previous works on the **Vi-Km and Km-Vi** translation task

propose a lightweight, modular training pipeline that leverages small-scale continual pretraining on monolingual data, supervised fine-tuning, and Direct preference optimization, with synthetic data augmentation generated by high-performance LLMs.

The empirical experiment is built upon the `SeaLLMs-v3-1.5B-Chat` model, a small multilingual ASEAN PLM. By leveraging the LoRA adapter, we efficiently multi-stage train the model and achieved the highest performance, with 62.00/58.43 and 57.12/55.61 (%) (BLEU/METEOR) on the evaluation set. These results outperform commercial translation systems (Google Translate, Microsoft Translator), LLMs such as GPT-4o and Gemini models, and also previous works (Pham Van and Le Thanh, 2022; Quoc et al., 2023; Duc et al., 2025).

These results highlight the limitations of general-purpose LLMs in handling low-resource language pairs and demonstrate the effectiveness of task-specific adaptation. The proposed training framework, which focuses on a cost-effective, lightweight, and comprehensive method, provides a promising and scalable direction for future research on machine translation in low-resource languages.

For future direction, there still exists room for expanding the quality of research, not only for Vietnamese-Khmer pair, but also for other regional low-resource machine translation. In particular, future work may investigate adaptive tokenization strategies that better capture the linguistic properties of languages, as well as the use of cross-lingual transfer learning from typologically or geographically related languages. Furthermore, integrating evaluation methods that combine automatic metrics with human-in-the-loop assessment would provide a more holistic measure of translation quality and better guide model development.

Acknowledgments

This research is funded by Hanoi University of Science and Technology (HUST) under Grant no T2024-PC-041.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *WMT@ACL*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tho Tran Duc, Huy Nguyen Trong, and Huong Le Thanh. 2025. Improving quality of vietnamese to khmer neural machine translation using multi-stage fine-tuning strategy. In *Information and Communication Technology*, pages 69–79, Singapore. Springer Nature Singapore.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yova Kementchedjhieva and Anders Søgaard. 2023. Grammatical error correction through round-trip machine translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2208–2215, Dubrovnik, Croatia. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Zhanibek Kozhirbayev. 2024. Enhancing neural machine translation with fine-tuned mbart50 pre-trained model: An examination with low-resource translation pairs. *Ingénierie des systèmes d information*, 29:831–838.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT@ACL*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource nmt guided by pos-tagging and paraphrase embedding. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(6).

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Linh The Nguyen and Dat Quoc Nguyen. 2025. Pre-training of foundation adapters for LLM fine-tuning. In *The Fourth Blogpost Track at ICLR 2025*.

Vinh Van Nguyen, Ha Nguyen, Huong Thanh Le, Thai Phuong Nguyen, Tan Van Bui, Luan Nghia Pham, Anh Tuan Phan, Cong Hoang-Minh Nguyen, Viet Hong Tran, and Anh Huu Tran. 2022. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5494–5502, Marseille, France. European Language Resources Association.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hanh Pham Van and Huong Le Thanh. 2022. Improving khmer-vietnamese machine translation with data augmentation methods. In *Proceedings of the 11th International Symposium on Information and Communication Technology*, SoICT '22, page 276–282, New York, NY, USA. Association for Computing Machinery.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018. Data selection with feature decay algorithms using an approximated target side. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 173–180, Brussels. International Conference on Spoken Language Translation.

Thai Nguyen Quoc, Huong Le Thanh, and Hanh Pham Van. 2023. Khmer-vietnamese neural machine translation improvement using data augmentation strategies. *Informatica*, 47(3):349–359. Bibliografija: str. 345-347.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium. Association for Computational Linguistics.

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yuzi Yan, Yibo Miao, Jialian Li, YipinZhang, Jian Xie, Zhijie Deng, and Dong Yan. 2025. 3d-properties: Identifying challenges in DPO and charting a path forward. In *The Thirteenth International Conference on Learning Representations*.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2024. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and

Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Training Configuration

We conduct the model training and evaluation experiments on a single GPU RTX 4090. The total time for the whole training process took over 50 hours. For loading and training the models, we utilize the modules: Transformers and TRL[6] from HuggingFace. The training process consists of three main stages: Small-scale Continual Pre-training (SmCPT), Supervised Fine-tuning (SFT), and Direct Preference Optimization (DPO). The hyperparameters used in each task are shown in Table 6.

Hyperparameter	SmCPT	SFT	DPO
number of epochs	1	2	2
batch size	8	4	4
learning rate	1e-4	1e-5	5e-6
warmup ratio	0.1	0.1	0.1
compute data type	bfloat16	bfloat16	bfloat16
optimizer	adamw_8bit	adamw_8bit	adamw_8bit
lr scheduler type	cosine	cosine	cosine
beta (β)	-	-	0.3

Table 6: Training Configuration

Small-scale Continual Pre-training (SmCPT) is conducted with a relatively high learning rate (1e-4) and a larger batch size to quickly adapt the pre-trained language model to domain-specific data while preserving general knowledge. Since the objective is to adapt to a specific task and domain rather than full-scale pre-training, only one epoch is used to prevent overfitting. In the Supervised Fine-tuning (SFT) phase, a lower learning rate (1e-5) ensures stable refinement without erasing prior knowledge, while the batch size is reduced to 4 to ensure stable updates. Since DPO training requires fine-grained updates, an even smaller learning rate (5e-6) is used to prevent drastic deviations from prior alignment. Across all training stages, computations are conducted in bfloat16 precision, which accelerates training compared to fp32 while preventing overflow when using fp16.

B Ablation Studies

B.1 Synthetic Data Filtering Effect

The overall pipeline of data augmentation is described in Section 4.2. Firstly, 30,000 Vietnamese sentences were selected from the seed dataset of TED (Reimers and Gurevych, 2020) by TF-IDF score. Then, we used `GPT-4o` model through the OpenAI API[7] with two-shot learning, taking a random combination from the training VOV dataset, to generate the translated Khmer settings. Two main hyperparameters need to be controlled during using LLM generation: *temperature* and *top_p*. In which *temperature* is a hyperparameter that controls the randomness of language model output, meaning a high temperature produces more unpredictable and creative results, and vice versa. Meanwhile, *top_p* is also used for controlling the randomness of the language model; it sets a threshold probability and selects the top tokens whose cumulative probability exceeds the threshold, which helps deliver more diverse and interesting output. For our task, we set *temperature*=0.1 and *top_p*=0.99, which is the best configuration for the translation task as in the experiment by Moslem et al. (2023).

The generated Khmer sentences are then back-translated to Vietnamese by using `Google Translate`. These sentences are compared to the original Vietnamese sentences using the hybrid score of BLEU and METEOR (see Equation 4). We conduct the experiment that combines the $top-p\%$ score pairs and VOV dataset to self-supervised fine-tune the model from the SmCPT checkpoint, as outlined in Section 3.2. As indicated from Table 7, just selecting the 90% highest score from augmented data shows the best performance in both BLEU and METEOR scores, with 61.13 and 57.44 respectively.

B.2 Preference-ranking Pairing Data Selection

We experiment with two strategies to create a DPO training dataset from the maximum of $20,000$ pref-

[6] https://huggingface.co/docs/trl/index

[7] https://platform.openai.com/docs

Threshold	Augmented	Vi-Km
$(top-p\%)$	Data Size	BLEU/METEOR
100%	30,000	60.88/57.07
95%	28,500	60.51/56.68
90%	**27,000**	**61.13/57.44**
85%	25,500	60.77/56.88

Table 7: Results of top-$p\%$ augmented data selection on the performance of fine-tuning

erence ranking pairs generated as described in Section 3.2.3. The strategy (1) is pairing the highest-ranking denoted as y_1 with the lower-ranking sentences denoted as $\{y_2, y_3, y_4, y_5\}$. The strategy (2) uses all of $20,000$ combinations, then computes the score of each pair (see Equation 4) to select $top-p\%$ highest score pairs as the DPO dataset.

No	Ranking pair		Threshold	Vi-Km
	Accepted	Rejected	$(top-p\%)$	BLEU/METEOR
(1)	y_1	$\{y_5\}$	-	61.85/58.13
	y_1	$\{y_4, y_5\}$		61.91/58.40
	y_1	$\{y_3, y_4, y_5\}$		61.74/58.06
	y_1	$\{y_2, y_3, y_4, y_5\}$		61.63/58.11
(2)	All combinations		100%	61.86/58.24
			95%	61.88/58.41
			90%	62.00/**58.43**
			85%	61.85/58.28
			80%	**62.04**/58.41
			75%	61.70/58.05

Table 8: Results of different DPO dataset selection strategies on the performance

Compared to previous stage results (see Table 7), DPO training significantly improves the performances in all cases (see Table 8). These demonstrate the importance of filtering low-quality ranking pairs to achieve optimal model parameters. At this stage, we select the model with a $top-p\%$ of 90%, achieving a BLEU score of 62.00 (slightly lower than the 80% $top-p\%$ model by 0.04). However, it peaks at a METEOR score of 58.43. Based on the conclusion from Agarwal and Lavie (2008) and our study, we conclude that METEOR offers a more rigorous and accurate evaluation compared to BLEU. This supports our decision to prioritize METEOR as a primary metric when determining the best-performing model.

B.3 Impact of Hyperparameter-beta on DPO Performance

We set up an experiment to study the effect of hyperparameter-β in DPO loss (see Equation 3). We use 2,000 preference ranking pairs that map the highest and the lowest ranking translated sentence ranked by `GPT-4o`. We adjusted the value of β from 0.05 to 0.5, and the results on BLEU and METEOR metrics corresponding to each value of β are (see Figure 3).

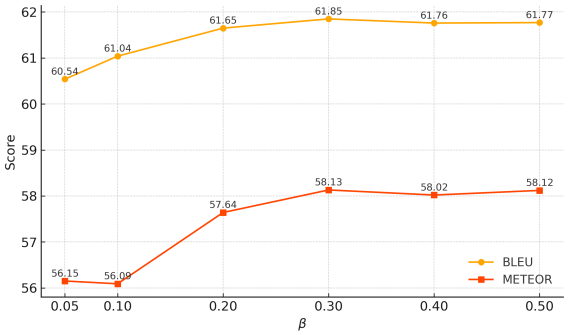


Figure 3: Impact of hyperparameter β on DPO performance

It can be observed that at β=0.3, the model achieves its best performance, with BLEU and METEOR scores of 61.85 and 58.13, respectively. Adjusting β significantly impacts the model's performance, making the selection of an appropriate β crucial for the specific downstream task. When β is smaller, the penalty term in the loss function is reduced, leading to less distinct separation between win and loss sequences. This setting is more suitable for tasks requiring creative and diverse outputs. On the other hand, for tasks where accurate outputs aligned with the input sequence are critical, choosing a larger β ensures stricter penalization and more precise outputs.

B.4 Impact of Dataset Size on SFT Performance

We provide an analysis of the impact of training data size on the SFT performance. For the experiment, we fine-tune the Adalora adapter (Zhang et al., 2023) on the training set. As shown in the results (Figure 4), there is a consistent improvement in model performance as the size of the training data increases, indicating a strong correlation between data quantity and translation quality. The BLEU score is 36.32% for a training size of 25K samples and progressively increases to 52.91% when fine-tuning on the full dataset. Similarly, the METEOR scores show a steady rise, from 31.68% for 25K samples to 47.83% for the full dataset. These findings emphasize the critical role of data size in supervised fine-tuning models, indicating that increased data may lead to better model perfor-

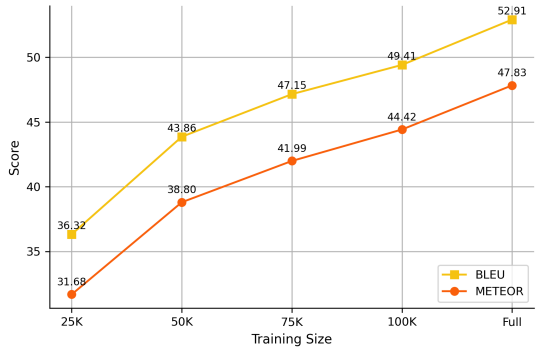


Figure 4: Impact of data training size on the SFT performance

mance, particularly for low-resource languages.

C Prompt Design

##Instruction: Translate from {source lang} to {target lang} #Input ({source lang}): {Input sentence} #Output ({target lang}):

Table 9: Prompt for fine-tuning translation task

Table 9 presents the template used for fine-tuning the translation model. The instruction specifies the translation task, with inputs labeled as "Input (Vietnamese)" and outputs labeled as "Output (Khmer)".

##Instruction: Here are some examples of {source lang}-to-{target lang} translations: #Example 1: {source lang}: {Input sample sentence 1} {target lang}: {Output sample sentence 1} #Example 2: {source lang}: {Input sample sentence 2} {target lang}: {Output sample sentence 2} ##Your task: Now, translate the following {source lang} sentence into {target lang}: {source lang}: {Input sentence} {target lang}:

Table 10: Prompt for generating Khmer translation on random two-shot examples

A two-shot learning prompt is used to generate synthetic Khmer translations with the `GPT-4o` model (see Table 10), including an instruction that provides examples of Vietnamese-to-Khmer translations and demonstrates the expected format and style of translation. After the examples, the model is assigned to translate a new Vietnamese sentence into Khmer. This prompt format incorporates in-context learning by showing examples and immediately asking the model to perform the task. In the experiment, based on the strength of the teacher model, only two examples are provided to balance the token length and the cost.

##Instruction: Your task is to rank the following {target lang} translation candidates based on the **accuracy** and **naturalness** relative to the source {source lang} sentence and the golden reference {target lang} translation. #Original {source lang}: {Input sentence} #Reference {target lang}: {Golden translation sentence} #Candidates: 1. {Candidate translation 1} 2. {Candidate translation 2} 3. {Candidate translation 3} 4. {Candidate translation 4} 5. {Candidate translation 5} ##Your task: Ranking the candidates from the best to the worst in the format: {candidate_number_1, candidate_number_2, ..., candidate_number_5}.

Table 11: Prompt for LLM translation ranking

Table 11 outlines the prompt used for ranking Khmer translation candidates, the idea of which is based on LLM-as-a-judge (Yuan et al., 2024). The instruction asks the model to rank the given candidate translations based on **accuracy** and **naturalness** relative to the original Vietnamese sentence and a 'golden' reference Khmer translation.

Researching and Rethinking the Culture and Practice of Eldercare in Hong Kong:A Corpus-assisted Discourse Analysis

Jesse W. C. Yip[1], Michelle Man-Long Pang[2]
Centre for Research on Linguistics and Language Studies (CRLLS)
The Education University of Hong Kong
[1]jwcyip@eduhk.hk, [2]michellemlpang@gmail.com

Abstract

Following the double whammy of the anti-extradition bill movement and COVID-19 in 2019, Hong Kong experienced a significant 'BNO exodus' that has reshaped the city's demographic landscape. The phenomenon accelerates Hong Kong's transition toward a super-aged society and a pool of 'stay-behind' elderly, creating shortages in eldercare personnel and further straining the city's support systems for its aging population. This corpus-assisted discourse analysis examines how the culture and practice of eldercare are represented in Chinese news media in post-2019 Hong Kong. The findings show that eldercare is generally depicted as a socio-economic issue in a neoliberal frame. News articles primarily report on the social welfare implemented by the government and the needs and community services provided for the ageing population. It is argued that the traditional virtue of filial piety to eldercare has been transformed into a neoliberal practice of the government; and that the voices of the elderly are undermined, though societal assistance for them has been explicated in the news media. Implications for journalistic practices in the portrayal of eldercare are also offered.

1 Introduction

In April 2019, the Hong Kong government introduced plans to pass an extradition bill that would allow criminal suspects in Hong Kong to be extradited to mainland China. In protest, over a million Hong Kong people marched to the government headquarters, voicing their fears that the bill would compromise the city's legal autonomy. Over the following year, as protesters' voices were met with inertia and suppression, the protests escalated in scale and violence, dividing Hong Kong society politically along generational lines (BBC, 2019). This political opposition fractured many families and parent-child relationships (Yu et al., 2023, cited in Chan and Chiu, 2025), while violent clashes paralyzed the city until COVID-19's second wave brought temporary stasis through social distancing measures (Woo et al., 2021). The National Security Law (NSL) was subsequently implemented in June 2020, triggering a surge in emigration inquiries and marking the beginning of the 'BNO exodus' (Benson, 2025; Mak, 2020, cited in Ho, 2025). From the introduction of the BN(O) visa route in 2021 to December 2023, 163,850 applications were submitted, with 157,576 approved — approximately 2.1% of Hong Kong's 2024 population — and 140,300 individuals already settling in the UK (Home Office, 2024, as cited in Lam and Fong, 2025, p.149).Most critically, these emigrants typically took only their children abroad, leaving elderly parents behind (Rolfe and Benson, 2023, cited in Chan, 2025a), creating a population of 移民遺老 'elderly abandoned in migration'that gained media attention (Chan and Chiu, 2025).

This demographic shift has intensified Hong Kong's transition to a 'super-aged society' (Choy, 2022), facing economic austerity and healthcare shortages (LegCo Research Office, 2023, cited in Lam and Fong, 2025). The shift has also created shortages in eldercare personnel, further straining the city's support systems for its aging population (Lam and Fong, 2025). While existing literature addresses ageism, voice, and agency in news coverage of the elderly during COVID outbreaks, and separately examines the BNO exodus through qualitative studies with key populations, including the elderly, their caretakers, and family members, there remains a gap in research that is both current and specifically focused on media coverage of the elderly and eldercare, along with their sociocultural implications. Eldercare has been one of the crucial issues that the Hong Kong government has to address to improve the living quality of this increasing population in the city. News media can reflect the ideologies, attitudes, and problem- solving approaches to be shaped to deal with the tasks

involved in eldercare provision. Researching news articles focusing on eldercare can enhance our understanding of the established culture and practices of eldercare in Hong Kong and offer implications for the care culture cultivation and care practices. Conducting a corpus-assisted discourse analysis, this study examines how Hong Kong's elderly and local eldercare culture are represented in news media in post-2019 Hong Kong, serving as a pilot investigation for a larger study, aiming to identify salient discourse(s) and establish future research directions.

2 Shifting Narratives of Eldercare in Post-2019 Hong Kong

At the crux of lasting health consequences post-COVID, social support shortages, and austerity measures compromising elderly mobility, the 'stay-behind' elderly face with an additional challenge of having to navigate living alone in contemporary Hong Kong, exacerbating their existing intersectional plights. The sociopolitical upheaval brought about by the 2019 circumstances not only raised pressing pragmatic concerns for our elderly population, but also a cultural change that destabilized and restructured the way this population navigates its self-orientation and identity (Chan, 2025b; Chan and Chiu, 2025). It is, therefore, essential to examine eldercare in Hong Kong.

Neoliberal Reimagination of Filial Piety. For millennia, traditional Chinese eldercare culture has revolved heavily around the notion of filial piety and the Confucian wulun (五倫), where adult children—or the "sandwich generation"—were expected to live with their elderly parents, forming multigenerational households and serving as the primary caregivers as a reciprocal gesture for the care and support they received during their younger years. Research found that traditional Chinese eldercare structures appear to have been actively evolving since 2000s, both in terms of execution and ideology. With growing society-wide pressure to prioritize financial growth, increasingly exorbitant housing costs, and an informal social security regime, most Hong Kong families' capacity for eldercare has significantly declined (Wong, 2022; Phillips, 2018). For families with better financial circumstances, their eldercare modes are characterized by commodification: care responsibilities are "subcontracted" to migrant workers and private caretakers (Leung et al., 2019), and when home accommodation is no longer feasible, families resort to institutional care as the ultimate solution (Lam, 2022). Families with limited financial resources have significantly fewer options: they often juggle full-time work and caretaking responsibilities with limited community support (Wong, 2022). Recent literature suggests that filial beliefs of the elderly have been redefined under Hong Kong's current socioeconomic climate (Bai et al., 2020). The provision of eldercare appears to be highly neoliberalized, and the traditional Chinese virtue of filial piety is hardly applicable to elucidate the culture and practice of eldercare in Hong Kong.

The 'Stay-Behind' Elderly: Eldercare in Post-2019 Hong Kong. This neoliberal adaptation of filial expectations appears to have carried forward to post-2019 Hong Kong in the context of the BNO exodus. Chan and Chiu's (2025) study on parent-child dynamics between BNO emigrants and their 'stay-behind' elderly parents argued, through transnational care theory, that these elderly are fully autonomous agents with control over how they navigate the new relational configurations, with some even having agency over their transnational mobility. However, as empowering as this narrative, the mass media generally tell tales of desolation and subsistence (see Appendix A). One could contest this "neoliberal" desire for self-reliance with the notion of "acquiescence": Bai et al. (2020) highlighted a nuanced yet paramount conflict shared by most participants—while most elderly individuals deeply yearn to age in place with family members, they hesitate to expect care from their children due to practical constraints.

While mass media may use sensationalizing language, there is a harsh truth to their depiction: the conditions for aging alone in Hong Kong remain incredibly unfavourable, particularly for individuals of lower socioeconomic status. Access to accommodation, government-funded financial support, post-retirement benefits, community services, and support networks remains disproportionately limited for elderly individuals who receive little guidance on maintaining their own interests (Hung, 2022; Wong, 2022; He and Chou, 2019).

The heated debates on the BNO exodus and the 'stay-behind' elderly are rooted in a myriad of complex and intertwined issues, including existing eldercare policies, post-retirement security, and the aging population amid economic austerity. These pieces of the puzzle, when combined, form a broader picture of the sociopolitical realities of the

elderly in Hong Kong—pieces that can be uncovered by delving into their surrounding discourse and exploring the underlying perspectives. This study aims to identify salient discourse(s) in local Chinese newspapers surrounding the elderly, and eldercare culture and practices in post-2019 Hong Kong. It conducts corpus-assisted discourse analysis to address the following research questions:

1. What are the discourses regarding eldercare in Hong Kong, and how have they been shaped in Chinese-language news media since 2019?

2. What are the implications of the media-shaped discourses for the culture and practice of eldercare in Hong Kong post-2019?

3 Research Methods

This study focuses on Chinese news articles because there are significantly more Chinese than English news publishers in Hong Kong. Focusing on Chinese news provides wider representation in voice, stance, and audience orientation.

Emergent search phrases were developed based on key themes identified in the literature and a preliminary scoping search using Google News (see Appendix B). These phrases reflect core topics related to Hong Kong eldercare and were derived from inputting 香港長者('Hong Kong' and 'elderly') into Google News. Six broad and semantically neutral search phrases were used, with word boundaries parsed by space, including 移民長者 ('migration' and 'elderly'), 人口老化香港 ('aging population' and 'Hong Kong'), 長者政策 ('elderly' and 'policy'), 香港安老 ('Hong Kong' and 'eldercare'), 香港長者晚年 ('Hong Kong' and 'elderly' and 'later years'), and 香港長者情況/現況 ('Hong Kong' and 'elderly' and 'situation'). Searches were performed using an incognito browser session and a blank account to reduce algorithmic bias. The search settings were set to 'by relevance' and accounted for the search IP address geolocation (i.e., Hong Kong). The first 50 articles per keyword were chosen based on the following criteria:

- **Range**: April 2019 - July 22, 2025
- **Language**: Chinese (distinction between varieties not drawn)
- **Format**: Online articles, excluding those that were predominantly reposts of third-party content (e.g.infographics taken from other sources)
- **Publishers**: No restrictions, unless the content was cross-posted
- **Word count**: Articles under 200 words (excluding punctuation, numbers, and symbols) were eliminated
- **Theme**: Articles must focus on Hong Kong elderly issues or the social welfare system. Excluded articles were those that: 1) mentioned the elderly only in passing; 2) focused on unrelated social issues with no direct relevance to the elderly (see Appendix C).

The initial search yielded 300 articles, with nine search results directly skipped as it was apparent from the preview that they were irrelevant to our scope. After applying the selection criteria to the 300 articles, 18 were removed and replaced due to insufficient word count, cross-posting, or thematic irrelevance. Textual data was then extracted from the articles and collated manually for analysis. The compiled corpus consists of 397,222 Chinese words. Unlike other mega-size corpora with millions of words, this corpus is undeniably small. However, "smaller corpora are more suited for studying specialist genres" (Handford, 2010, p. 256). Several previous studies have shown that small corpora can be as powerful as their larger counterparts (Sinclair, 2001; Lam, 2018). This corpus enables us not only to capture the predominant themes/discourses of eldercare in Hong Kong newspapers, but also to conduct qualitative analysis, such as cross-case analysis that examines and compares similarities and differences of multiple cases based on the reviewed literature and the news reports in the corpus.

The analysis began with generating a word frequency list using *AntConc 4.3.1*. While the list contained Chinese lexical items and function words, we filtered out the function words, focusing on the lexical items, as "lexical words are the main carriers of information and contribute more to the semantic construction and communication" (Lam, 2018: 200). The most frequently occurring lexical terms in the list represented 'the likely source of lexical cohesion both within and across the texts in the corpus, and may also be predicted to be wholly, or part of, the core of the lexical item' (Cheng, 2012: 330).

Table 2: The predominant discourse and subthemes of the top ten frequently occurring lexical

items The top 10 most frequently occurring lexical words were thematically classified into discourses/themes via an iterative analysis of their concordances, which offer analysts a general sense of the word meanings in contexts. The corpus was further investigated by examining collocates of the most frequent lexical words. Specifically, the collocation lists of each of the frequently occurring lexical words in the discourse(s)/theme(s) were generated. The specified window span was set to 8L 8R (i.e., 8 words to the left of the word of interest and 8 to the right) owing to relatively long and complex sentences in Chinese news articles. The collocates were combined and compared according to their log-likelihood scores to identify the top ten collocates of each discourses/theme. This approach aims to avoid redundant analyses of collocates shared by multiple lexical words and lengthy data analysis presentations (Yip and Kong, 2025). Each of the discourses/themes related to eldercare in Hong Kong was examined qualitatively and quantitatively based on the most frequently occurring lexical words, their collocates, and concordances. Apart from the corpus-based analysis, we also conducted a critical cross-case analysis through in-depth and iterative reading of the collected news articles and literature review on eldercare, and then compared specific cases that critically delineate the research questions to offer insights into the topic.

4 Findings

The investigation begins with the top ten most frequently occurring lexical words, which are also evenly dispersed throughout the corpus.

Word	Rank	Freq	Norm Freq	Dispersion
elderly 長者	1	4094	20378	0.923
Hong Kong 香港	2	1974	9825	0.899
government 政府	3	1199	5968	0.837
services 服務	4	1021	5082	0.857
society 社會	5	800	3982	0.880
living 生活	6	704	3504	0.878
population 人口	7	696	3464	0.878
plan 計劃	8	624	3106	0.924
needs 需要	9	525	2613	0.910
provide 提供	10	519	25836	0.888

Table 1: *Top ten most frequently occurring lexical words and their dispersion.*

Table 1 shows the generated word frequency list which indicate the word frequencies and their dispersion scores. These frequently occurring words provide a foundation for examining the predominant discourses associated with eldercare represented in the Chinese news articles. The dispersion scores of the most frequently occurring lexical words are higher than 0.8, indicating that their high frequencies are not the result of dominance by individual texts in the corpus but they are evenly distributed throughout. Thus, the lexical items in the list enable us to reveal and examine the predominant discourses/ themes of the corpus.

As word meaning is "established by the consistent co-occurrence of a form with a certain semantic environment" (Sinclair, 1991, p. 112), the concordances of these high-frequency lexical items are examined to classify them into discourse(s)/theme(s) that offer a systematic picture for understanding what the corpus is about (Baker, 2023). The words are categorized into one predominant discourse, which consists of two sub-themes, as shown in Table 2. Specifically, with the words 香港'Hong Kong', 長者'elder', and 生活'living', the predominant discourse largely contextualizes the corpus, indicating that the news focuses on the elderly and their lives in Hong Kong. More importantly, the discourse frames eldercare as a socio-economic issue in Hong Kong. The discussion regarding the elderly's living covers a range of topics, including their mental health, living environment, and financial consideration. As signaled by the lexical words 人口'population' and 社會'society', the sub-theme of population and social welfare reports the statistics and growing trends of the elderly population in Hong Kong society, along with discussion of derived social issues such as ageing, poverty, and low birth rate. The words 政府'government' and 計劃'plan' mark the HKSAR government's implementation of welfare policies for the elderly population as responses to the issues. In the sub-theme of the elderly's needs and community services, the word 需要'need' literally highlights the needs of the elderly and the words 提供'provide' and 服務'services' denote the services offered by the community, such as rehabilitation services and care home services. Residential care home services have been among the predominant social welfare policies for the elderly and disabled in Hong Kong (Yip, 2024a). The following section illustrates the specific meanings of the predominant discourse and sub-themes based on the analysis of the collocates and concordances of the concerned high-frequency lexical words.

The Predominant Discourse: The Elderly as a Socio-Economic Issue in Hong Kong. The top-

Discourse	Highly frequent words (rank)	Sub-themes	Highly frequent words (rank)	Concordance example
The elderly as a socio-economic issue in Hong Kong	Elderly 長者(1), Hong Kong 香港(2), Living 生活(6)	The population and social welfare	government 政府(3), society 社會(5), population 人口(7), scheme 計劃(8)	當中，在擁有超過100 萬人口的地區裏，香港預計在2050 年會成為全球人口老化程度最高的城市。(Hong Kong is expected to have the highest proportion of elderly residents globally by 2050 among regions with populations over one million.)
		The elderly's needs and care services	services 服務(4), need 需要(9), provide 提供(10)	不少身體、家庭情況許可的長者也會選擇居家安老，並使用由政府提供的社區照顧服務，包括日間護理中心、家居照顧支援等。(A significant proportion of elderly individuals with adequate physical capacity and family support opt for aging in place, utilizing government-subsidized community care services such as daycare centres and home support services.)

Table 2: *The predominant discourse and sub-themes of the top ten frequently occurring lexical items*

ics and foci that newspapers select in their reports to represent the elderly construct and reflect the culture and practice of eldercare in Hong Kong. The predominant discourse offers a general understanding of the newspapers' tendency to depict the elderly as a socio-economic issue in Hong Kong. The collocates and concordances of the words 長者'elderly', 香港'Hong Kong', and 生活'living' provide details of how the elderly population is portrayed as an issue in the contexts.

Collocate	Rank	Freq(Scaled)	Likelihood
allowance 津貼	1	1328	246.968
quality 質素	2	1296	185.885
stay-behind 留守	3	1120	122.960
resident 居民	4	1920	120.013
Christian 基督教	5	912	102.594
increase 提升	6	2848	98.600
population 人口	7	11136	93.484
society 社會	8	12800	92.653
the elderly 老年人	9	3008	92.572
Incentive scheme 優惠計劃	10	1136	92.310

Table 3: *Top-ten frequently occurring collocates of elder, Hong Kong and living*

As shown in Table 3, the collocate 人口'population' often precedes 香港'Hong Kong' to form the phrase 香港人口'Hong Kong population'. The news reports tend to adopt the perspective of social administration and economics to discuss eldercare in Hong Kong. Thus, demographic information, including 香港人口'general population', 勞動人口'working population' and 長者人口'elder population', as well as the collocate 社會'society', are frequently mentioned to provide background information and rationale for highlighting eldercare as an issue in Hong Kong society. It is not uncommon to see the collocation 人口問題'demographic crisis' in the corpus. For example,

> 香港人口問題早已引起社會有識之士的關注，人口老化又遇上生育率持續保持低位，令香港人口結構發生較大的變化，由原來的鑽石型轉變為倒金字塔型，而且情況仍在持續。
>
> Hong Kong's demographic crisis has drawn sustained attention from insightful observers. The dual pressures of population aging and chronically low fertility rates have drawn sustained attention from insightful observers. The dual pressures of population aging and chronically low fertility rates have led to a significant change in Hong Kong's population structure, causing a dramatic shift from a diamond-shaped distribution to an inverted pyramid configuration, with this trend continuing unabated.

The elderly is frequently associated with financial support, such as 津貼'the government subsidies', and 優惠計劃'incentive schemes'. The

collocate 津貼'allowance' spotlights 長者生活津貼'the Old Age Living Allowance', which is frequently mentioned to indicate the government's efforts to enhance the elderly's living conditions. In addition, the collocate 基督教'Christian' denotes the NGO Hong Kong Christian Service, which provide services and financial assistance for the elderly, which provide services and financial assistance for elderly in need, such as care home services and mental health support. The collocates of 質素/質量'quality', and 提高'increase' literally signal the goals of improving the elderly's quality of life after retirement by providing a better living environment.

> ...促請政府制訂政策及投放資源推廣軟餐，以照顧長者的需要；並加强對照顧者的支援，以提升長者的生活質素和福祉。
> ...urge the government to formulate policies and allocate resources to promote soft meals to address the needs of elderly citizens, as well as to enhance support for caregivers in order to increase the quality of life and well-being of the elderly.

Noteworthily, the most frequently occurring word with 長者or 老年人'the elderly terms' in the corpus is 留守'stay-behind'. The collocation stay-behind elders is specifically used to describe elderly individuals whose family members have emigrated overseas, indicating that a group of elders have been separated from their families since 2019 (Chan and Chiu, 2025). The following excerpts provide further details:

> 調查顯示留守長者最需要的服務依次為醫療保障、生活照顧及社區支援，當中子女移民未滿兩年的長者　更覺港府提供的支援不足，顯示仍處適應階段的留守長者更可能有多方面需求，建議當局及社福界推出　有關政策。
> The survey shows that the services most needed by the stay-behind elderly are medical security, daily care, and community support services. Among them, those whose children have emigrated within the past two years feel that the support provided by the Hong Kong government is insufficient, indicating that stay-behind elderly in the adaptation stage may have a wider range of needs. It is recommended that the authorities and the social welfare sector introduce relevant policies.

The stay-behind elderly have drawn the government's attention and have been surveyed to analyse their needs in society. It is concluded that their needs, which are similar to those of non-stay-behind elderly, include medical security, personal care support, and community-based assistance. The government has been urged to implement relevant welfare policies to address the needs of the elderly. As part of this discourse, the following sub-themes outline the topics discussed in news articles about eldercare.

Collocate	Rank	Freq(Scaled)	Likelihood
population 人口	1	11136	662.157
ten thousand 萬	2	2880	449.617
year 年	3	18480	398.141
special region 特區	4	2080	397.594
estimate 推算	5	1520	358.613
working 勞動	6	1392	332.743
occupy 佔	7	2896	274.175
years 歲	8	10336	274.148
scheme 計劃	9	9984	225.08
Health care voucher 醫療券	10	1968	211.137

Table 4: *Top-ten frequently occurring collocates of government, society, population and scheme*

Sub-Theme: The Population and Social Welfare. Table 4 shows the top-ten most frequently occurring collocates of the words 政府'government', 社會'society', 人口'population' and 計劃'scheme'. As mentioned, eldercare in Hong Kong is often emphasized and discussed from the perspective of socieal administration and policy. Thus, the elderly population is mentioned with collocates such as 人口'population', 萬'ten thousand, 年'year', 佔'occupy', 歲'years old' and 推算'estimate'. The news reports tend to indicate the estimated millions of people who are 60 years old and occupy a certain percentage of the overall population, suggesting that there are more and more elderly in society. The growth of the elderly population is often associated with the working population, as the increase in the elderly population means a decrease in the working population. The collocate 特區'special region' denotes that the Special Admiration Region (SAR) government plays the main role in tackling the issue. For instance,

> 人口持續老化對香港社會安老支援系統造成巨大壓力，特區政府多年來提倡「居家安老為本，院舍照顧　為後

援」的政策方針，意在通過推動「居家安老」，減輕社會和醫療體系的壓力。

The continuous aging of the population has placed enormous pressure on Hong Kong's social elderly care The continuous aging of the population has placed enormous pressure on Hong Kong's social elderly care support system. For many years, the HKSAR Government has advocated the policy principle of 'home-based elderly care as the foundation, with institutional care as a backup,' aiming to alleviate the burden on society and the healthcare system by promoting home-based elderly care.

The news articles report the government's efforts to allocate budgets to implement measures that enhance the quality of life for the elderly, including several schemes indicated by the collocates 計劃'scheme' and 醫療券'health care voucher', such as 預設照顧計劃'Advance Care Planning Scheme', 長者醫療券'Elderly Health Care Voucher', 長者醫療券大灣區試點計劃'Greater Bay Area Voucher Pilot Scheme', and 福建計劃'Fujian Scheme'. These schemes demonstrate the government's emphasis on elderly care and advocate for having their retirement lives in mainland China. For example,

所有參與基層醫療健康計劃的醫生必須加入《基層醫療健康指南》，計劃包括疫苗資助計劃、慢性疾病　共同治理先導計劃、長者醫療券計劃、大腸癌篩查計劃和普通科門診公私營協作計劃等。

All physicians participating in the Primary Healthcare Scheme must be enrolled in the Primary Healthcare Directory. The program encompasses multiple initiatives, including the Vaccination Subsidy Scheme, Chronic Disease Co-Care Pilot Scheme, Elderly Health Care Voucher Scheme, Colorectal Cancer Screening Programme, and the General Outpatient Clinic Public-Private Partnership Programme.

Sub-Theme: The Elderly's Needs and Community Services. This sub-theme highlights the needs of the elderly and the community services provided for them. To obtain more details of the discourse, the top-ten collocates of 服務'services', 需要'needs' and 提供'provide' (see Table 5) are examined.

Collocate	Rank	Freq(Scaled)	Likelihood
provide 提供	1	8304	271.003
services 服務	1	16336	271.003
centre 中心	3	6400	187.446
take care of 照顧	4	7696	170.374
assistance 支援	5	7792	144.577
community 社區	6	7568	140.178
home 家居	7	2032	135.238
daytime 日間	8	800	131.848
medical 醫療	9	8064	121.252
elderly care 安老	10	5264	100.368

Table 5: *Top-ten frequently occurring collocates of services, needs, and provide*

The collocate 提供'provide' is frequently used to show the care, assistance and services offer to the elderly at the community level. Thus, this collocate is associated with other collocates that denote the various types of support for the elderly, including 服務'services', 支援'assistance', 醫療'medical', and 安老'elderly care'. The collocates 中心'centre', 社區'community', 家居'home', 日間'daytime' and 照顧'take care of' indicate the support providers or specific support schemes for the elderly. For instance, the support providers are often NGOs subsidised by the government, such as基督教家庭服務中心'Christian Family Service Centre', 社區服務中心'community service centre' and 日間護理中心'day care centre'. For instance,

不少身體、家庭情況許可的長者也會選擇居家安老，並使用由政府提供的社區照顧服務，包括日間護理中心、家居照顧支援等。

A significant proportion of elderly individuals with adequate physical capacity and family support opt for aging in place, utilizing government-subsidized community care services such as daycare centres and home support services.

5 Discussion

The findings indicate that eldercare is generally framed as a socio-economic discourse, illustrated with survey-based statistics and supported by academic authorities from academia. As a result, one of the predominant themes of eldercare in the news media emphasizes the welfare policies implemented by the SAR government to address the

"socio-economic issue" of the increasing elderly population. In other words, eldercare has been interpreted as one of the government's main responsibilities. This corresponds to Bai et al. (2020), who suggest that eldercare is redefined under the current socioeconomic climate of Hong Kong. The virtue of filial piety in traditional Chinese eldercare culture has faded out in the culture of eldercare in Hong Kong. Although the term 居家安老'home-based eldercare' has been mentioned, it is not literally related to the Confucian concept of filial piety, which emphasizes children's care, obedience, and remembrance of parents. Home-based eldercare acts as a pragmatic advocate, encouraging reduced demand and consumption of public eldercare resources, which are already insufficient, as care for the elderly has largely been commodified through subcontracting to migrant workers and private care services institutions (Wong, 2022; Leung et al., 2019). Without a doubt, the transition to a "super-aged society" in Hong Kong (Choy, 2022) will probably lead to economic turmoil and shortages of healthcare and eldercare resources. This is not contradictory to the cultivation and advocacy of traditional core values in eldercare, such as respect and company. Nevertheless, the virtue of filial piety appears not to be illuminated in the news articles about eldercare.

Moreover, to rationalize the demand for eldercare services and resources, newspapers tend to construct a discourse of vulnerability, as indicated by the sub-theme of needs and community services. This portrayal seemingly positions the elderly as inferior in society and can result in social exclusion and reduced opportunities for older adults (Levy and Banaji, 2002; Loos and Ivan, 2018). Previous studies suggest that elderly individuals who adopt the devaluing views in social discourse are likely to suffer from lower self-esteem, mental illness, and reduced longevity (Swift et al., 2017). Moreover, these representations play a critical role in shaping both public discourse and policy decisions regarding the elderly. When older adults are predominantly portrayed as passive recipients of care or as burdensome to society, the resulting policy responses often take on a paternalistic character. Such policies tend to prioritize the provision of basic care and the containment of perceived risks, rather than fostering opportunities for empowerment, active participation, and social inclusion among older individuals. This framing can ultimately marginalize the elderly further, reinforcing stereotypes and limiting their agency within society (van Dyk, 2016).

While plenty of research discloses the impacts of the surge in emigration since 2019 on the population composition in Hong Kong (Chan, 2025a; Lam and Fong, 2025; Tran, 2025), this study shows that major newspapers appear to overlook the correlation between emigration and the increase in the elderly population, who are referred to as stay-behind elders. The stay-behind elders have become a new group within Hong Kong's elderly population, characterized as the newest additions. However, their difference from the non-stay-behind elderly appears to be overlooked. Prior research has highlighted the phenomenon that the elderly tends to opt to live alone or stay behind in Hong Kong despite their desire to age in a place with family members (Bai et al., 2020), as they are afraid of being a burden on their children's families. Thus, solely using a neoliberal approach to eldercare is unlikely to address the root of the issue and what assistance the elderly really needs from the government. For instance, the elderly might desire for travel allowance to visit their children overseas. It is important to ensure that what the support givers provide is what the support receivers need (Yip, 2024b).

6 Conclusion

This study reveals that newspapers in Hong Kong generally represent eldercare as a socio-economic issue. The discourse frames the elderly as a population in needs of support and as a financial burden on the government. The vulnerability of this population is highlighted as a means to rationalize the provision of government assistance, including subsidies and care services. A significant portion of the news coverage focuses on reporting the needs of the elderly and the welfare policies that have been implemented. However, the voices of the elderly have been inaudible in the Hong Kong news media. Without listening to the elderly's thoughts, it is difficult to provide support that caters to their real needs. Moreover, the news media play a key role in shaping public perception of eldercare and its core virtues. The virtues of filial piety deserve greater emphasis in news articles to raise citizens' awareness of taking care of their parents' physical and mental needs.

References

Xue Bai, Daniel W. L. Lai, and Chang Liu. 2020. Personal care expectations: Photovoices of Chinese ageing adults in Hong Kong. *Health & Social Care in the Community,* 28(3): 1071–1081.

Paul Baker. 2023. *Using corpora in discourse analysis.* Bloomsbury.

BBC News. 2019. Hong Kong: Timeline of extradition protests.

Michaela Benson. 2025. "Global Britain", the coloniality of migration, and the Hong Kong BN(O) visa. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 19–38, Routledge.

Yuk Wah Chan. 2025a. From colonial subjects to British? In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 39–54, Routledge.

Yuk Wah Chan. 2025b. Introduction. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 1–16. Routledge.

Yuk Wah Chan and Eunice Yin-yung Chiu. 2025. Migrants and their parents. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 183–198. Routledge.

Winnie Cheng. 2012. Exploring corpus linguistics: Language in action. Routledge.

Ray Choy. 2022. Implementation of community care policy for older adults in Hong Kong. In Vincent T. S. Law and Bernard Y. F. Fong, editors, *Ageing with Dignity in Hong Kong and Asia*, volume 16, pages 25–39. Springer.

Alex Jingwei He and Kee-Lee Chou. 2019. Long-term care service needs and planning for the future: A study of middle-aged and older adults in Hong Kong. *Ageing and Society*, 39(2): 221–253.

Wing Chung Ho. 2025. From reluctant to emotional. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 111–122. Routledge.

Michael Handford. 2010. What can a corpus tell us about specialist genres? In Anne O'Keeffe and Michael McCarthy, editors, The *Routledge handbook of corpus linguistics*, pages 255–269. Routledge.

Andrew T. W. Hung. 2022. Family caring for the elderly during the pandemic in Hong Kong: Perspective from Confucian familism. *Public Administration and Policy*, 25(1): 13–24.

Gigi Lam. 2022. Problems encountered by elders in residential care services in Hong Kong. *Asian Education and Development Studies*, 11(1): 106–117.

Phoenix W. Y. Lam. 2018. The discursive construction and realization of the Hong Kong brand: A corpus-informed study. *Text & Talk,* 38(2): 191–216.

Ka Wang Kelvin Lam and Eric Fong. 2025. The impact of post-2019 migration on Hong Kong population dynamics. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 145–160. Routledge.

Vivian W. Y. Leung, Ching Man Lam, and Yan Liang. 2020. Parents' expectations of familial elder care under the neoliberal Hong Kong society. *Journal of Family Issues*, 41(4): 437–459.

Becca R. Levy and Mahzarin R. Banaji. 2002. Implicit ageism. In Todd D. Nelson, editor, *Ageism: Stereotyping and prejudice against older persons*, pages 49–75. MIT Press.

Kevin Chi-pan Liu. 2025. 兩元乘車優惠：香港長者尊嚴與社會温度的守護者[The HK$2 Transport Subsidy: Upholding dignity and social inclusion for Hong Kong's elderly]. Hong Kong China News Agency. https://www.hkcna.hk/h5/docDetail.jsp?id=100902619&channel=4662.

Eugène Loos and Loredana Ivan. 2018. Visual ageism in the media. In Liat Ayalon and Tesch-Römer Clemens, editors, *Contemporary Perspectives on Ageism*, pages 163–176. Springer.

Hoi Yan Mak. 2020. 港版國安法:移民查詢單日急升十倍[Hong Kong national security law:

Inquiries about immigration surge tenfold in a single day].https://www.hk01.com/%E7%A4%BE%E6%9C%83%E6%96%B0%E8%81%9E/476612/%E6%B8%AF%E7%89%88%E5%9C%8B%E5%AE%89%E6%B3%95-%E7%A7%BB%E6%B0%91%E6%9F%A5%E8%A9%A2%E5%96%AE%E6%97%A5%E6%80%A5%E5%8D%87%E5%8D%81%E5%80%8D-%E5%B9%B4%E8%BC%95%E4%BA%BA%E5%B0%87%E9%A6%96%E6%9C%9F%E4%BD%9C%E8%B3%87%E6%9C%AC?utm_source=01webshare&utm_medium=referral&utm_campaign=non_native

David R. Phillips, Jean Woo, Francis Cheung, Moses Wong, and Pui Hing Chau. 2018. Exploring the age-friendliness of Hong Kong: Opportunities, initiatives and challenges in an ageing Asian city. In Tine Buffel, Sophie Handler, and Chris Phillipson, editors, *Age-Friendly Cities and Communities*, pages 119–142. Policy Press.

Heather Rolfe and Thomas Benson. 2023. From HK to UK: Hong Kongers and their new lives in Britain. Welcoming Committee for Hong Kongers. https://www.britishfuture.org/wp-content/uploads/2023/11/HK-to-UK-report.Final_.pdf.

John Sinclair. 1991. *Corpus, concordance, collocation.* Oxford University Press.

John Sinclair. 2001. Preface. In Ghadessy Mohsen, Robert L. Roseberry, and Alex Henry, editors, Small corpus studies and ELT: *Theory and practice*, pages vii–xv. John Benjamins.

Hannah J. Swift, Dominic Abrams, Ruth A. Lamont, and Lisbeth Drury. 2017. The risks of ageism model: How ageism and negative attitudes toward age can be a barrier to active aging. *Social Issues and Policy Review*, 11(1): 195–231.

Emilie Tran. 2025. The exodus of expatriates during political change and the COVID-19 pandemic. In Yuk Wah Chan and Yvonne To, editors, *Global Hong Kong*, pages 161–182. Routledge.

Silke van Dyk. 2016. The precarity of critique in times of crisis: On the limits of counter-discourse in the media representation of old age. Media, Culture & Society, 38(6): 918–933.

Jean Woo, Dion Sik Yee Leung, Ruby H.Y. Yu, Richard Wai Tong Lee, and Hung K. Wong. 2021. Factors affecting trends in societal indicators of ageing well in Hong Kong: Policies, politics and pandemics. *The Journal of Nutrition, Health and Aging*, 25(3): 325–329.

Kayla Wong, Terry Lum, and Gloria Wong. 2020. The COVID-19 long-term care situation in Hong Kong: Impact and measures. International Long-Term Care Policy Network, CPEC-LSE. https://ltccovid.org.

William W. L. Wong. 2022. Subsistence life and aging agony: A case of cultural affluence in Hong Kong. *Social Development Issues*, 43, (2).

Jesse W. C. Yip. 2024a. A discourse study on handover communication among care providers in a residential care home for persons with intellectual disabilities. *Health Communication*, 39(2): 216-228.

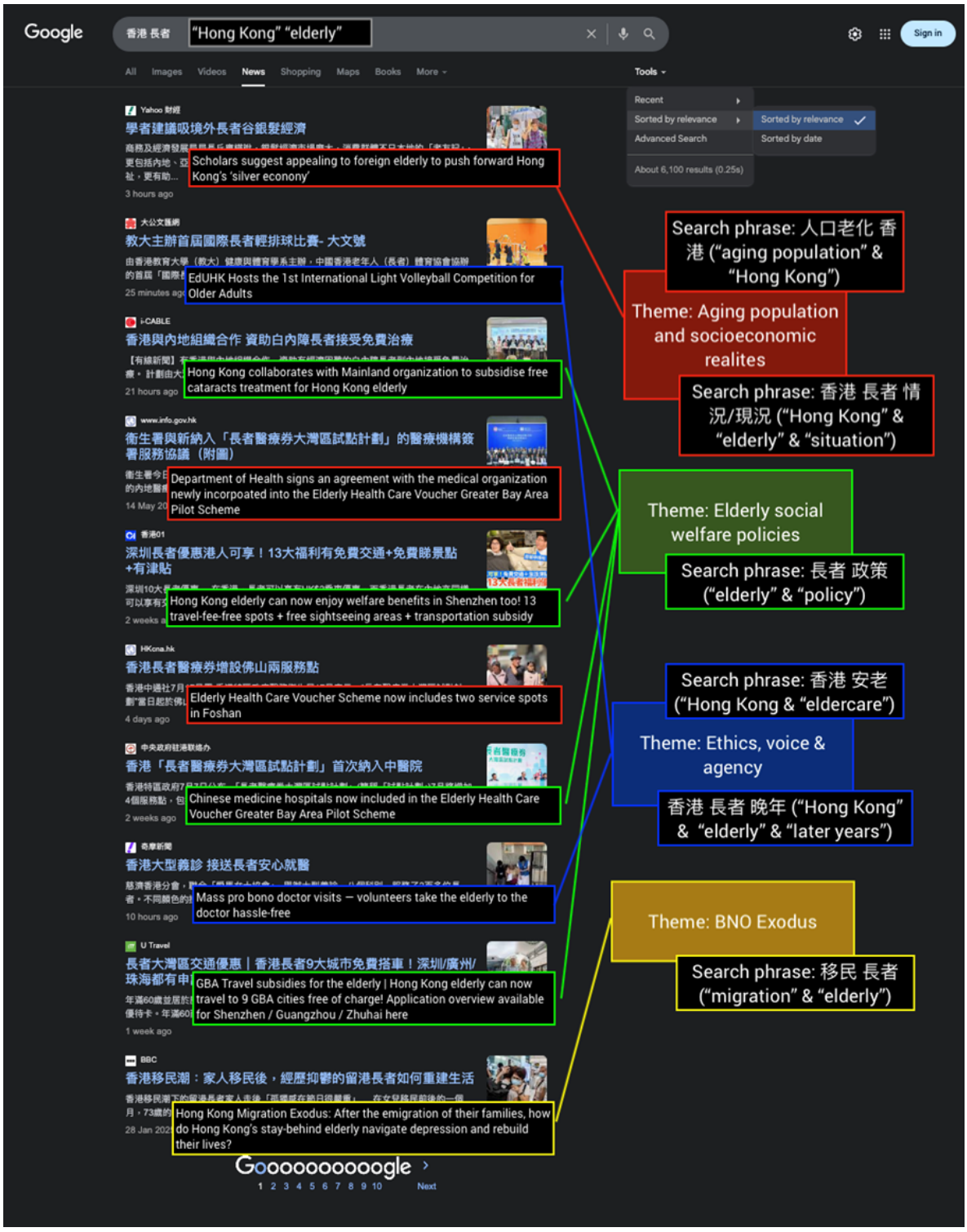
Jesse W. C. Yip. 2024b. *Discourse of online social support: A study of online self-help groups for anxiety and depression.* Springer.

Jesse W. C. Yip and Kenneth C. C. Kong. 2025. Identity construction of Hong Kong's Chief Executive in blogs: A corpus-informed study. In Sze Ming Leung and Sin Wai Chan, editors, *Applying technology to language and translation*, pages 15–40. Routledge.

Government of the Hong Kong Special Administrative Region. 2025. Govt to optimise fiscal resources.https://www.news.gov.hk/eng/2025/02/20250226/20250226_095322_529.html?type=ticker

A News tales of desolation and subsistence

B A preliminary scoping search using Google News



C Sample of excluded news articles

輸入外勞屢揭假招聘中介剋扣工資　議員促適時檢討政策

屢揭外勞遭中介剋扣工資事件。

人口高齡化和出生率下降，加劇勞動力不足的問題。立法會今日(3日)否決「優化輸入人力措施」的原議案及修正案同遭否決。提出原議案的批發及零售界議員鄧家輝指10年後香港逾1/3人口逾65歲，屬全世界最長壽的，加上低出生率，勢必惡化人手不足的問題，儘管多次提出輸入勞工，但勞工界的領導則主張釋放本地婦女或長者的勞動力，可是仍沒法改變人口結構的問題，雖然目前有輸入勞工，但只佔本港300多萬勞動人口的2%而已，反觀新加坡的外勞則佔勞動人口的25%，澳門則佔1/3。

他指收到很多業界反映，目前須以行業工資中位數僱用外勞，再要加9,600港元的兩年培訓費給政府，兼且要包住，人力成本較聘用本地更高，在現時的經營環境下不少僱主都望而卻步，或把成本轉嫁本地的消費者。他指在今時今刻北上消費的潮流之下，必需檢視行業的競爭力。

對於修正案，他指刪掉「檢討輸入勞工的工資不得少於香港相關職位的每月工資中位數的規定」這個最重要的環節，又指修正案提出鼓勵婦女投入職場，可是多年來的現實是仍然不能解決人手短缺的問題。他又指自己只是提議案出來研究而已，然而卻被刪除，質疑做法「連研究都唔得」。他又指新加坡即使引入好多外勞，但人均GDP是133,000多美元；澳門更厲害，達到134,000多美元；而香港只有7萬多美元而已，可見輸入外勞不會拖低工資。

議員顏汶羽指提出的修正案，無論「行業性輸入勞工計劃」還是「優化補充勞工計劃」，目的為了應對香港現時面對人力不足的問題。按政府的預測，去到2028年，香港將會缺少18萬人手，特別是建造業、安老及餐飲等，情況非常嚴重。然而輸入外勞只是一個短期的權宜之策，並非一勞永逸的方法，中長期是要靠本地勞動力的規劃和培訓，因此他希望政府採取更積極的措施，進一步釋放本地勞動力，特別是婦女和長者，以及提升本地勞工的工作技能。

另一議員管浩鳴指不論藍領或白領、前線或後勤，都面「人才荒」問題，運輸、零售餐飲等行業更是重災區，而隨着人口老化，就業人口持續萎縮，再加上近年出現的移民潮，進一步凸顯人力資源捉襟見肘的困境。

管指人手不足的問題會帶來連鎖反應，例如員工人數減少，薪酬水漲船高，可是生產力並沒有相應增加，企業要支撐營運就惟有加價，最終導致百物騰貴，市民生活開支百上加斤。有企業會礙於人手不足而縮短營業時間，或打消擴充意欲。故此，人力資源出現斷層會拖垮經濟復甦，最終受害都是普羅市民。如果堅持強調聘用本地人才，最後只會得不償失。

他又認為輸入外勞不會拖低本地工人的收入，反而經濟發展停滯才是最影響打工仔收入的致命傷，目前香港需要一套整全的輸入外勞政策，以解決勞工嚴重短缺的問題。他指現行的「補充勞工優化計劃」無疑在審批上比以往更具彈性和針對性，但計無兩頭利，輸入外勞自然都有缺點，例如容易衍生濫用取巧的「假招聘」、中介肆意剋扣外勞薪金等問題，期望政府能夠適時根據經濟和社會環境的轉變，精準掌握輸入勞工的配額和推出保障本地就業的措施，確保落實輸入外勞的初心及原意。

Note: "Aging population" and "the elderly" have been mentioned several times as the factor motivating migrant labour imports, but the main idea of the article is discussing how under-supervised the foreign import labour force is and why it warrants reforms.

Syllabic Distribution and Developmental Patterns of Mandarin Glides in Preschool Children

I-Ping Wan*, Xiang Li, Yu Ching Tsai
Graduate Institute of Linguistics, National Chengchi University, Taipei, Taiwan
ipwan@mail2.nccu.tw, 113555010@g.nccu.edu.tw,
113555004@g.nccu.edu.tw

Abstract

This study investigates the developmental patterns of three Mandarin glides, [j], [w], and [ɥ], in preschool children by examining their distribution across different syllabic positions. Data were collected from 45 typically developing children between the ages of 7 months and 6 years using a cross-sectional picture-naming task. The speech produced by the children was transcribed using IPA and compared with adult usage patterns extracted from a 202-hour spoken corpus of Taiwan Mandarin. Statistical analyses, including chi-square tests and linear mixed-effects models, showed that [j] and [w] were acquired earlier and produced more accurately than [ɥ]. Glide accuracy was significantly higher in onset positions and lower in coda positions and consonant-glide clusters. These findings suggest that articulatory complexity and positional effects both influence the course of glide acquisition in Mandarin. By including a comprehensive range of syllabic contexts, this study highlights the importance of structural distribution in early phonological development and provides empirical support for phonotactic modeling in child language acquisition. This study combines corpus-based and experimental evidence to address glide acquisition in a tonal language, focusing on phonological modeling, speech development, and child language research.

1 Introduction

Previous studies have shown that the acquisition of sounds in early childhood is accomplished through stages of development. The trajectory of phonological development in early ages is said to be influenced by language universal factors (e.g., Jakobson, 1968), markedness (e.g., Edwards, 1974), and the articulation ease (e.g., Locke, 1972). Certain patterns in phonological development appear to be cross-linguistically universal, while others are shaped by the specific linguistic input children receive, and still others are influenced by articulatory effort.

Numerous studies have examined the acquisition order of English glide sounds. Dodd et al. (2003) found that [w] tends to be acquired earlier than [j], consistent with earlier findings by Wellman (1931), Poole (1934), Templin (1957), and Sander (1972). To account for the variation in age of acquisition across these studies, Dodd et al. (2003) suggested that both age and gender significantly influence speech development, contributing to differences in observed acquisition timelines. Despite variability in specific ages reported, these studies consistently agree on the developmental trajectory of English glides, with [w] acquired before [j].

Based on the relationship between surface glides and their underlying vowel counterparts in Mandarin phonology, the developmental sequence of these glides can be inferred from the acquisition patterns of the high vowels. Prior research has consistently shown that the high front vowel /i/ emerges first in child speech, followed by the high back vowel /u/, and finally the high front rounded vowel /y/ (Jeng, 1979; Wu & Xu, 1979; Su, 1985; J. Hsu, 1987; Cao, 2003; Shi & Wen, 2007; H. Hsu, 2016). This developmental progression, from simpler to more complex articulatory gestures, aligns with Locke's (1972) theory of articulatory ease. Accordingly, if glides reflect the acquisition order of their vowel counterparts, it can be hypothesized that palatal rounded [j], derived from high front vowel /i/, is acquired earlier than labial-velar rounded [w] and palatal rounded [ɥ], which are derived from /u/ and /y/, respectively.

Previous studies on Mandarin phonological development have consistently shown that lip rounding presents greater articulatory difficulty for young children compared to lip spreading or neutral postures. For example, Peng & Chen (2020) and Lou (2020) reported that rounded vowels and glides ([w], [ɥ]) are typically acquired later than [j], re-

flecting higher motor control demands for rounding. Similarly, Zhang (2016) observed frequent substitution and deletion of [w] and [ɥ] in Mandarin-Taiwanese Min children, indicating articulatory instability. From a phonological perspective, Wiedenhof (2015) and Fu (2023) noted that [ɥ] has a restricted distribution and is often merged with [j], suggesting that the Mandarin phonological system itself limits the robustness of [ɥ]. Even non-native speakers of Mandarin, Thai preschoolers, learn Mandarin was found the similar patterns (Wan et al, 2024a).

There are twelve different syllable structure combinations in Mandarin (V, CV, GV, VG, VN, CVG, CVN, GVG, GVN, CGV, CGVG, CGVN). G represents the three glides [w, ɥ, j] in Mandarin. Except for the nasal sound /n/, all other consonants are strictly patterned to be in syllable initial position; glides, on the other hand, are more flexible (Chao, 1968; C.C. Cheng, 1973; Duanmu, 2007). They can appear in syllable initial (e.g., jaw), medial (e.g., ɕjaw), and final positions (e.g., xaj); however, it is not legitimate for [ɥ] to occur in syllable-final position (Duanmu, 2007; Norman, 1988). Table 1 below shows the twelve syllable structure types in Mandarin:

Syllable structure	Phonetic Transcription	Gloss
V	[i 55]	clothes
CV	[pi 21]	pencil
GV	[ja 35]	teeth
VG	[wɔ 21]	I, me
VN	[i 51]	hard
CVG	[naj 21]	milk
CVN	[san 21]	umbrella
GVG	[jow 21]	to have
GVN	[wan 21]	bowl
CGV	[xwa 55]	flower
CGVG	[njow 35]	cow
CGVN	[twan 21]	short

Table 1: The twelve syllable structure types in Mandarin

Although Mandarin allows some flexibility, it imposes strict phonotactic constraints on consonant-glide onset sequences. As noted by Wan (1999), glides show position-specific behavior in terms of both manner and place of articulation. The glide [w] co-occurs with bilabials, dentals, velars, and retroflexes; [j] occurs with bilabials, dentals, and palatals; and [ɥ] appears primarily with palatals and the nasal [n]. Table 2 shows the possible distribution of CG sequences by Wan (2003) shows the legal CG segment structures in Mandarin (see Appendix I). For the co-occurrence of vowels and glides, it is undoubtedly that the three high vowels /i/, /y/, and /u/ cannot combine with the three glides, due to the homorganicity tautosyllabic constraints (Steriade, 1988). For instance, [ij], [yɥ], and [uw] are not allowed. According to Wan and Jaeger (2003), non-high vowels can be paired with the three glides freely in the pattern of VG sequence combinations. For example, [aj], [ɥɛ], [ow], [fej], and [xaw].

Mandarin includes three glides: the labiovelar [w], the palatal [j], and the labiopalatal [ɥ]. These are generally considered surface realizations of the high vowels /u/, /i/, and /y/, respectively, when they occur adjacent to non-high vowels. None of these glides are part of the underlying phonemic inventory of Mandarin. Instead, they emerge through phonological processes that affect high vowels in specific syllabic environments. This interpretation is supported by several studies (Lin, 1989; Wu, 1994; Duanmu, 2002; Wan, 1999; Wan 2002; Fu 2023) and aligns with the view that glides in Mandarin function as allophonic variants of high vowels rather than as independent semivowels or consonants.

Wan (2002, 2003, 2006) presents a series of studies on glide behavior in Mandarin speech errors, offering compelling evidence that glides function more like nuclear elements than true consonants. First, glide errors revealed that [j], [w], and [ɥ] show positional variation and often align with the nucleus, particularly when preceded by labial, dental, or retroflex consonants. In contrast, glides form onset clusters when following palatal or velar consonants. Wan and Jaeger (2003) further observed that glides frequently substitute for their high-vowel counterparts and may trigger vowel-vowel substitutions across syllables. These interactions suggest that glides share more representational features with vowels than with consonants or other glides. Importantly, glide errors such as substitution, deletion, and insertion often occur independently, indicating their segmental status. In a follow-up study, Wan (2006) analyzed postnuclear glides and coda nasals in native speaker speech errors. Glides showed significantly fewer interactional errors than nasals (79 vs. 158), and interacted more with vowels than with consonants. Although glides occasionally co-occurred with nasals, the pattern suggested that glides are less stable as codas and more tightly integrated with the nucleus. Collectively, these findings support the analysis of glides as prosodically and phonologically affiliated

with the nucleus, rather than as independent onset or coda consonants.

In English, only vowel monophthongs can form diphthongs or triphthongs, while glides such as [j] and [w] are typically analyzed as consonants (Ladefoged and Maddieson, 1996). In contrast, Mandarin diphthongs and triphthongs are formed through the combination of a glide, which is considered the surface form of the high vowels [i], [u], or [y], and one or more non-high vowels. This contrast highlights a fundamental difference in the phonological treatment of glides in the two languages. In Mandarin, the behavior and distribution of glides correspond more closely to the properties of vowels than to those of consonants. Examples of diphthongs in Mandarin and token frequencies are shown as below.

IPA	Freq.	IPA	Freq.	IPA	Freq.
ɕja	18031	swan	1727	ʂwən	337
ɕjɛn	14461	tʂwan	1498	lwan	324
tɕja	11453	tʂʰwan	1496	tɕʰɥən	307
mjɛn	10427	tɕʰja	1337	tɕɥən	287
pjɛn	9907	kwa	1226	kʰwan	281
tɕʰjɛn	8286	tʂwa	1171	ʐwan	228
tɕjɛn	8224	kʰwa	1045	twən	228
tjɛn	8019	ɕɥən	972	swən	216
njɛn	7820	xwa	724	tswən	214
tʰjɛn	6324	tʂwən	688	nja	118
lja	5760	xwən	682	tɕʰjo	115
kwan	5689	ɕjo	641	kwən	55
tɕʰɥen	3705	tʂʰwa	619	nwan	54
xwan	2990	tʰwən	576	tʰwən	49
ljɛn	2766	tɕɥɛn	466	ʐwən	39
ɕɥen	2452	tʂʰwən	453	ʂwan	27
twan	2096	tʰwan	378	tswan	20
lwən	1874	ʂwa	361	tsʰwan	3
pʰjɛn	1806	kʰwən	348	tɕjo	1

Table 3: Sample of CGVN in IPA and token frequencies (Wan et al., 2024b)

Since glide development in Mandarin has typically been examined within the broader context of vowel development, the specific behavior and acquisition of the three glides [w], [j], and [ɥ] in different syllabic positions remain insufficiently investigated in the literature. This study addresses the following research questions, each accompanied by a predicted outcome:

1. What is the developmental order of the three Mandarin glides [j], [w], and [ɥ] in preschool children between 7 months and 6 years of age? Based on previous studies, we predict that [j] and [w] will emerge earlier, while [ɥ], due to its articulatory complexity, will be acquired later

2. How do syllabic positions (onset, post-consonantal, and coda) affect the accuracy and emergence of Mandarin glide production in child speech? We expect that glides will be produced more accurately in onset position than in post-consonantal or coda positions, reflecting structural facilitation in early speech.

3. To what extent do articulatory complexity and positional effects influence the acquisition patterns of Mandarin glides? It is predicted that the interaction of higher articulatory demands (as in [ɥ]) and marked syllabic positions will result in greater variability and delayed mastery compared to simpler contexts.

Glide acquisition in child speech has often been examined with limited positional scope, typically focusing on onset production. However, Mandarin glides present a unique opportunity to examine positional asymmetries, given their attested occurrences in onset, coda, and complex cluster contexts. Mandarin prohibits vowel sequences with a single syllable, high vowels are systematically realized as glides when adjacent to other vowels. Therefore, this study aims to address these gaps by analyzing the production accuracy of [j], [w], and [ɥ] across syllabic positions in child speech, and comparing these patterns with large-scale adult corpus data. Drawing on markedness theory and positional prominence, we ask how phonotactic constraints shape developmental trajectories, and how corpus-driven benchmarks can refine our understanding of normative phonological development.

2 Methodology

The data were drawn from a spoken corpus of Taiwan Mandarin children (N = 45; mean age = 3.8 years, SD = 1.4; 23 boys, 22 girls) collected in the Phonetics and Psycholinguistics Laboratory at National Chengchi University, Taipei, Taiwan [1].

[1] In this lab, the spoken corpus contains multi-tier linguistic annotations and encompasses diverse spoken data such as speech interactions among multiple speakers, conversations between speech therapists and Mandarin-speaking aphasic patients, language acquisition patterns in typically developing children aged 7 months to 6 years, and speech samples from children with language disorders aged 3 to 6 years. More recently, the corpus has been expanded to include not only

Participants, aged 7 months to 6 years, completed picture-naming tasks designed to elicit three target glides [j], [w], and [ɥ] in both monosyllabic and disyllabic words across different syllabic positions, as shown in Table 4

Word	IPA	Word	IPA
羊	[ja35]	月亮	[ɥɛ51 lja51]
碗	[wan21]	貝殼	[pej51 k^hɣ35]
雲	[ɥən35]	帽子	[maw51 tsɨ]
海	[xaj21]	樹葉	[ʂu51 jɛ51]
猴	[xow35]	青蛙	[tɕ^{h}i55 wa55]
掃把	[saw51 pa21]	醫院	[i55 ɥɛn51]
牙齒	[ja35 tʂɨ21]	牛奶	[njow35 naj21]
襪子	[wa51 tsɨ]	蛋糕	[tan51 kaw55]

Table 4: The test examples in the experiment

Participant information is shown in Table 5 (see Appendix II). All data were audio-recorded and transcribed by two well-trained master's students specializing in phonetics, who had prior experience transcribing over 24 languages before undertaking this task, with interrater 91%, and 9% of discrepancies were resolved through acoustic validation in Praat. To enhance reliability, Praat was used to analyze the recordings, ensuring consistency in transcriptions and validating any discrepancies. When confusions still remained, the data were discarded. Annotation was conducted semi-automatically using a Hybrid-DNN-HMM framework. For developmental analysis, children were grouped in six-month intervals (2;0–6;0), enabling observation of age-related changes in phonological development, with vocabulary production serving as an index of phonological maturity.

All pictures were presented three times in a randomized order. As preschool children may not always produce the exact target word, approximations were accepted as correct as long as the target glides were realized. For example, [tɕ^{h}i55 wa55] *'frog'* produced as [wa55 wa55] or [maw51 ts] *'hat'* as [maw51 maw51] were considered correct because the target [w] occurred. Similarly, [njow35 naj21] *'milk'* produced as [nej55 nej 55] and [xaj21] *'sea'* produced as [wej21] *'water'* were also counted as correct, since the target items involved a glide [j] in coda position. However, if the target [w] did not appear in the production, such as [tan51 kaw55] *'cake'* produced as [tan51 tan51], the response was coded as incorrect. A total of 84 test items involving true consonants and glides across different syllabic positions were examined. In this cross-sectional design, if the participant was unable to produce the target word correctly, the response was taken to indicate that the child at that age had not yet mastered the sound. The procedures followed those outlined in Wan et al. (2024b). Table 6, as seen in Appendix III, shows the correctness and error types by the 45 participants.

data from adult learners of Mandarin with L1 backgrounds in Indonesian, Vietnamese, and Thai, but also speech data from native speakers of these three languages.

3 Findings and Analysis

Data were collected through a picture-naming task involving typically developing Mandarin-speaking children in Taiwan. The task was designed to elicit productions of the glides [j], [w], and [ɥ] across various syllabic positions. Of the 7,290 expected tokens, 6,555 usable responses were obtained, with the shortfall due to omissions during the task. Among these, 6,292 productions were judged accurate, yielding an overall accuracy rate of 95.99%. Three main error types were identified in the remaining data: substitution (1.16%), where the intended glide was replaced; addition (0.26%), where an extra glide was inserted; and deletion (2.59%), where the glide was omitted. Productions containing multiple error types were excluded from analysis to ensure categorical clarity.

The glide [w] appeared most frequently in the confusion matrix, with 3,110 total instances, of which 3,102 were produced accurately. Substitution errors were rare and included realizations such as [j] (n = 1), [p] (n = 7), [p^h] (n = 1), [v] (n = 1), and [] (n = 8). Deletion of [w] was observed in 106 cases. These results indicate a high level of production accuracy for [w], although some substitution and deletion errors did occur. Table 7 shows the error type distribution according to the three glide sounds [j], [w], [ɥ] in Mandarin.

Glide / Error Type	Substitution	Deletion	Addition	Total Error
j (34.1%)	7	62	14	83
w (51.8%)	18	106	3	127
ɥ (14.1%)	33	2	0	35
Total	**58**	**170**	**17**	**245**

Table 7: Error type distribution according to the three glide sounds

This table reveals that among the three glides, [w] accounts for the highest proportion of errors (51.8%), with deletion being the most frequent type (106 instances). This suggests that [w] may be particularly vulnerable to deletion during speech pro-

duction. The glide [j] also shows a predominance of deletion errors, but with a relatively higher number of additions, indicating its potential instability or overgeneralization in certain phonological contexts. In contrast, [ɥ] is primarily affected by substitution errors, with almost no deletions or additions, implying that it may be more prone to misidentification or confusion with other sounds. Overall, deletion emerges as the most common error type, highlighting the articulatory challenges glides may pose in speech development. The observed and expected numbers of the three glide error types below show the extent to which the observed distribution deviates from the expected frequencies under the assumption of independence, forming the basis for the chi-square calculation.

Glide	Error Type	Observed	Expected
[j]	Substitution	7	19.65
[j]	Deletion	62	59.59
[j]	Addition	14	5.76
[w]	Substitution	18	30.07
[w]	Deletion	106	88.12
[w]	Addition	3	8.81
[ɥ]	Substitution	33	8.29
[ɥ]	Deletion	2	24.29
[ɥ]	Addition	0	2.43

Table 8: Expected frequencies and observed frequencies regarding the error type distribution

By examining the relationship between the three glides ([j], [w], and [ɥ]) and error types (substitution, addition, and deletion) through the chi-square test of independence, results reveal that a significant correlation can be observed ($\chi^2(4) = 128.87, p < .001$). In the test, the degrees of freedom (df = 4) were calculated, based on the categories in both variables, the three glides and the syllable position in this case. The test then compares the expected and observed frequencies to see if the variables are related. Overall, the distribution of errors is dependent on the glide type, suggesting that the errors in children's output is systematically related to the syllable position of the three glides.

$$df = (rows - 1) \times (columns - 1) \quad (1)$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

A closer observation on the three syllable positions regarding the three glides, [ɥ] is the only sound prohibited from occurring at the syllable-final position. For [w], its occurrence in the syllable-final-position was the highest. As for [j], it appears the most in the consonant-glide cluster. The total count of production according to the syllable position is 6361. Again, [ɥ] showed minimal distribution and variation, with only two syllable positions allowed in Mandarin phonotactics. Combining Table 7 and 8, the pattern of Mandarin glide syllable distribution (Table 9) can be observed: [ɥ] had the lowest overall frequency, but it showed the highest error rate ($35/292 \approx 11.99\%$); meanwhile, even though [w] and [j] had a higher occurrence compared to [ɥ], the error rate was 3.99% and 2.33% separately. According to each glide sound, the significant error types vary as well. [ɥ] exhibited the most instances of substitution; [w] had the highest deletion count; [j] experienced a high addition rate. The results from Table 7 and 8 suggest that both the occurrence frequency and syllable distribution influence the Mandarin glide production in normal children.

IPA	Syllable Position	Number of Production	Examples
[w] 3110 (48.89%)	CG	768	[xwa55]
	syllable-initial	258	[wan21]
	syllable-final	2084	[tɕʰjow35]
[j] 2959 (46.52%)	CG	1711	[ljɛn21]
	syllable-initial	373	[jaŋ35]
	syllable-final	875	[tswej21]
[ɥ] 292 (4.59%)	CG	83	[ɕɥoŋ35]
	syllable-initial	209	[ɥən35]
Total count		6361	

Table 9: Mandarin glides syllable distribution and examples

In Table 9, [[ɕɥoŋ35] "bear" became [tjoŋ35]. under the cases of substitution, [xwa55] "flower" became [fa55]. The glide [w] was deleted and the consonant [x] was replaced by [f]. The disappearance and replacement of the glide [w] and the consonant [x] showed a closer relationship between the onset and prenuclear glide. Substitutions like [tjow35] "ball" became [tj35], [o] and [w] were substituted with [ɔ]. The relationship between the postnuclear glide [w] and the nucleus [o] is bounded. They were substituted together by mid vowel [ɔ]. The substitution example satisfies the observation she found: the tautosyllabic prenuclear glide, vowel, and postnuclear glide are treated as a constituent, backed up by the ability to be substituted with a single vowel. Consequently, postnuclear glides should be considered part of the nucleus rather than the coda.

Building on their position, the status of glides

can be further suggested. The classification of glides cannot be uniform in all positions. Their status is syllable-sensitive, shaped by the structural interaction with the adjacent sounds within one syllable. The following table below shows the error types and syllable distribution of the three glides in Mandarin.

IPA	Syllable Position	Number of error	Examples
[w] (124, 54.39%)	CG	28	[tswej]→[tsej] [ja ʂwa]→[xa ʂja]
	syllable-initial	9	[wan]→[pan]
	syllable-final	87	[tɕjawʌi]→[tɕja] [njaw]→[njaŋ]
[j] (69, 30.26%)	CG	18	[ɕjaŋ tɕjaw]→[xaŋ tjaw]
	syllable-initial	5	[ja ʂwa]→[xa ʂwa] [ja ʂwa]→[a twa]
	syllable-final	46	[pej k^hʏ]→[pɐ] [xaj]→[xaə]
[ɥ] (35, 15.35%)	CG	7	[ɕɥoŋ]→[ɕjoŋ] [ɕɥoŋ]→[ʂoŋ]
	syllable-initial	28	[i ɥɛn]→[i jan]
Total count	228		

Table 10: Error types and syllable distribution

This table analyzes the distribution of pronunciation errors for the glides [w], [j], and [ɥ] across different syllable positions. The glide [w] shows the highest error rate, accounting for 54.39% of all errors, with the majority occurring in syllable-final position. The glide [j] ranks second with 30.26% of errors, mainly found in syllable-final and CG structures. The glide [ɥ] has the fewest errors (15.35%), appearing only in CG and syllable-initial positions, and not at all in syllable-final positions. These findings highlight the varying stability of glides in different phonological environments, with [w] being particularly vulnerable in syllable-final contexts. Table 11 (see Appendix IV) shows correctness and all the error types, along with the percentage out of all produced data.

The confusion matrix (see Appendix VII) helps further identify specific error patterns of the three glides, [w], [j], and [ɥ], revealing which glides are most frequently misarticulated and the phonemes with which these glides most often interact.

4 Discussion and Conclusion

Regarding the developmental order of the three Mandarin glides [j], [w], and [ɥ] in preschool children aged between 7 months and 6 years, the consistently high accuracy rates for both [w] and [j] indicate that these glides are more stable in child speech than [ɥ], which emerged later and was less accurate. Substitution errors were observed only from [ɥ] to [w] or [j], but not in the opposite direction, suggesting an asymmetrical developmental pattern. The frequency of correct productions followed the order [w] > [j] > [ɥ], aligning with Wan's (2003) distributional ranking of legal CG sequences and with adult spoken corpus data reported in Wan et al. (2024b). The correlation between children's and adults' production rankings further suggests that early glide acquisition is strongly shaped by frequency distributions in the ambient language.

In terms of syllabic positions (onset, postconsonantal, and coda), these factors significantly affect the accuracy and emergence of Mandarin glide production in child speech. Chi-square analysis confirmed a significant relationship between glide type and error type ($\chi^2(4) = 128.87$, $p < .001$), indicating that errors were not random but systematically linked to syllable position. Positional analysis showed that [w] frequently occurred in syllable-final position, [j] predominated in consonant–glide clusters, and [ɥ] was limited to initial and medial positions. Error patterns also varied across contexts: [ɥ] was most prone to substitution, [w] and [j] to deletion, and [j] exhibited a higher addition rate among the three sounds. These findings reflect the positional distribution of glides and demonstrate that syllable structure strongly conditions glide accuracy and emergence.

Finally, to some extent, articulatory complexity and positional effects influence the acquisition patterns of Mandarin glides. The findings confirm that articulatory demands and positional restrictions jointly shape glide acquisition. [w] and [j] were acquired earlier and used more consistently due to their positional flexibility and lower articulatory complexity. In contrast, [ɥ], which requires lip rounding in addition to palatal articulation, was more unstable, frequently substituted, or deleted. This supports prior research showing that lip rounding is generally more difficult than lip spreading for Mandarin-speaking children (Peng & Chen, 2020; Lou, 2020; Zhang, 2016; Wan et al., 2024a). The structural behavior of glides further corroborates earlier phonological models: in syllable-initial position, glides function as onsets or secondary articulations (Duanmu, 1990, 2002), while in syllable-final position, postnuclear glides pattern with the nucleus (Lin, 1989). Overall, the interaction of articulatory complexity and syllable structure accounts for both the variability and the

delayed mastery observed in [ɥ] relative to [j] and [w].

In addition, our findings are consistent with prior research showing that lip rounding is generally more difficult for Mandarin-speaking children than lip spreading. Rounded glides such as [ɥ] and [w] tend to emerge later, are often substituted by [j], or deleted altogether, reflecting both articulatory constraints on lip rounding and phonological restrictions in early child Mandarin (Peng & Chen, 2020; Lou, 2020; Zhang, 2016; Wan et al. 2024a).

These findings suggest that the acquisition pattern in all language systems is shaped by a combination of a universal developmental pattern, the language-specific sound inventory, and the articulatory constraints. To better understand the systematic trajectory of sound acquisition, it is essential to investigate the status of individual sound types within a particular language.

Acknowledgments

We sincerely appreciate the valuable and constructive comments from the two anonymous reviewers and the editor, which have significantly improved this manuscript. All remaining errors in the analysis and interpretation are solely our own. This research was supported in part by the National Science and Technology Council in Taiwan (NSC 100-2410-H-004-187-) to the first author.

References

Bao, Z. (1990). Fanqie languages and reduplication. *Linguistic Inquiry*, 21(3), 317-350.

Bao, Z. (1996). The syllable in Chinese. *Journal of Chinese Linguistics*, 24, 312-354.

Baxter, W. H. (1992). *A handbook of old Chinese phonology*. Berlin, New York: De Gruyter Mouton. DOI:10.1515/9783110857085.

Cao, J. X. (2003). One case study of early phonological development in Mandarin-speaking children. In *Proceedings of the 6th National Symposium on Modern Phonetics*. Tianjin Normal University.

Chao, Y. R. (1934). The non-uniqueness of phonemic solutions of phonetic systems. Bulletin of Institute of History and Philology. *Academia Sinica*, 4(4), 363-397. DOI:10.6355/BIHPAS.193401.0363

Chao, Y. R. (1948). The voiced velar fricative as an initial in Mandarin. *Le Maitre Phonetique*, 63, 2-3.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Cheng, C. C. (1973). *A Synchronic Phonology of Mandarin Chinese.* Berlin, New York: De Gruyter Mouton. DOI:10.1515/9783110866407

Cheng, R. L. (1966). Mandarin phonological structure. *Journal of Linguistics*, 2(2), 135–158. DOI:10.1017/S0022226700001444

Chiang, W. Y. (1992). *The prosodic morphology and phonology of affixation in Taiwanese and other Chinese languages.* [Doctoral dissertation, University of Delaware]. ProQuest Dissertations Publishing.

Chung, R. F. (1989). *Aspects of Kepa phonology.* [Doctoral dissertation, University of Illinois]. ProQuest Dissertations Publishing.

Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, 131(1), 442–454. https://doi.org/10.1121/1.3651823

Davis, S., & Hammond, M. (1995). On the status of onglides in American English. *Phonology*, 12(2), 159-182. DOI:10.1017/S0952675700002463

Davis, B. L., & MacNeilage, P. F. (1990). Acquisition of correct vowel production: a quantitative case study. *Journal of Speech and Hearing Research*, 33(1), 16–27. https://doi.org/10.1044/jshr.3301.16

Duanmu, S. (1990). *A Formal Study of Syllable, Tone, Stress, and Domain in Chinese Languages.* [Doctoral dissertation, MIT]. ProQuest Dissertations Publishing.

Duanmu, S. (2002). *The Phonology of Standard Chinese.* Oxford University Press.

Duanmu, S. (2007). *The phonology of standard Chinese (2nd ed.)*. New York: Oxford Uni-

versity Press Inc.

Dodd, B., Holm, A., Hua, Z., & Crosbie, S. (2003). Phonological development: a normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, 17(8), 617–643. https://doi.org/10.1080/0269920031000111348

Donegan, P. (2002). Phonological processes and phonetic rules. *Future Challenges for Natural Linguistics*, 57-81.

Edwards, M. L. (1974). Perception and production in child phonology: The testing of four hypotheses. *Journal of Child Language*, 1(2), 205-219.

Fu, B. (2023). Uncovering Mandarin Speaker Knowledge with Language Game Experiments (Doctoral dissertation, Massachusetts Institute of Technology).

Gussman, E. (2007). *The Phonology of Polish.* Oxford University Press.

Hartman, L. M. (1944). The segmental phonemes of the Peiping dialect. *Language*, 20, 28-42.

Hockett, C. F. (1947). Peiping Phonology. *Journal of the American Oriental Society*, 67, 211-222.

Hockett, C. F. (1950). Peiping morphophonemics. *Language*, 26, 63-85.

Howie, J. M. (1976). Acoustical studies of Mandarin vowels and tones (Vol. 18). Cambridge University Press.

Hsu, H. Y. (2016). *Phonological development and disorder in Taiwan Mandarin: The status of glides* (Master's thesis). National Chengchi University.

Hsu, J. (1987). A *study of the various stages of development and acquisition of Mandarin Chinese by children in Taiwan milieu* [Unpublished Master's dissertation]. Fu Jen Catholic University.

Ingram, D. (1989). *First language acquisition: Method, description and explanation.* Cambridge University Press.

Jakobson, R. (1968). *Child language: aphasia and phonological universals (No. 72).* Walter de Gruyter.

Jeng, H.-H. (1979). The acquisition of Chinese phonology in relation to Jakobson's law of irreversible solidarity. In *Proceedings of the 9th International Congress of Phonetic Sciences: Vol. 2* (pp. 155–161). Copenhagen, Denmark: University of Copenhagen.

Karlgren, B. (1954). *Compendium of Phonetics in Ancient and Archaic Chinese.* SMC Pub. Inc.

Kim, C. W., & Kim. H-Y. (1991). The character of Korean glides. *Studies in the Linguistic Sciences,* 21(2), 113-125.

Ladefoged, P., & Maddieson, I. (1996) *The sounds of the world's languages.* Oxford: Blackwell Publishers.

Li, F., & Munson, B. (2016). The development of voiceless sibilant fricatives in Putonghua-speaking children. *Journal of Speech, Language, and Hearing Research,* 59(4), 699-712.

Lin, Y. H. (1989). *Autosegmental Treatment of Segmental Process in Chinese Phonology.* [Doctoral dissertation, University of Texas]. ProQuest Dissertations Publishing.

Locke, J. L. (1972). Ease of articulation. *Journal of Speech and Hearing Research,* 15(1), 194-200. https://doi.org/10.1044/jshr.1501.194

Lou, S. (2020). Early Phonological development in Mandarin: An analysis of prosodic structures, segments and tones from babbling through the single-word period (Doctoral dissertation, University of York).

McLeod, S. (2009). Speech sound acquisition. In J. E. Bernthal, N. W. Bankson & P. Flipsen Jnr (Eds.), *Articulation and phonological disorders: Speech sound disorders in children* (6th ed., pp. 63-120 + 385-405). Boston, MA: Pearson Education.

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal*

of Speech-Language Pathology, 27(4), 1546-1571.

Miao, X., & Zhu, M. (1992). Language development in Chinese children. *Advances in Psychology*, 90, 237-276. https://doi.org/10.1016/S0166-4115(08)61894-4

Moser, D. (1991). Slips of the tongue and pen in Chinese. Department of Oriental Studies, University of Pennsylvania.

Norman, J. (1988). *Chinese*. Cambridge University Press.

Peng, G., & Chen, F. (2020). Speech development in mandarin-speaking children. In Speech perception, production and acquisition: Multidisciplinary approaches in Chinese languages (pp. 219-242). Singapore: Springer Singapore.

Poole, I. (1934). Genetic development of articulation of consonant sounds in speech. *The Elementary English Review*, 159-161.

Priester, G. H., Post, W. J., & Goorhuis-Brouwer, S. M. (2011). Phonetic and phonemic acquisition: Normative data in English and Dutch speech sound development. *International Journal of Pediatric Otorhinolaryngology,* 75(4), 592-596. https://doi.org/10.1016/j.ijporl.2011.01.027

Sander, E. K. (1972). When are speech sounds learned?. *Journal of Speech and Hearing Disorders,* 37(1), 55-63.

Scullen, M. E. (1993). *The Prosodic Morphology of French.* [Doctoral dissertation, Indiana University]. ProQuest Dissertations Publishing.

Shen, J. (1993) Slips of the tongue and the syllable structure of Mandarin Chinese. InS-C. Yau (ed.) Essays on the Chinese Language by Contemporary Chinese scholars. Paris: Centre de Recherches Linguistiques sur l'Asie Orientale-Ecole des Hautes Etudes en Sciences Sociales. 139-162.

Shi, F., & Wen, B. Y. (2007). Vowel development in Mandarin-speaking children. *Zhongguo Yuwen, 2007(5)*, 444–454.

Steriade, D. (1988). Review of CV Phonology: A Generative Theory of the Syllable, by G. N. Clements & S. J. Keyser. *Language*, 64(1), 118–129. DOI:10.2307/414790

Su, A-T. (1985). *The acquisition of Mandarin phonology by Taiwanese children.* [Master's dissertation]. Fu Jen Catholic University.

Templin, Mildred C. (1957). *Certain language skills in children: Their development and interrelationships.* Minneapolis: University of Minnesota Press.

Wan, I. P. (1999). *Mandarin phonology: Evidence from speech errors.* [Doctoral dissertation, State University of New York]. ProQuest Dissertations Publishing.

Wan, I. P. (2002). The Status of Prenuclear Glides in Mandarin Syllables: Evidence from Psycholinguistics and Experimental Acoustics. 聲韻論叢, (11), 141-162. https://doi.org/10.29753/CP.200210.0008

Wan, I. P. (2003). *Alignments of prenuclear glides in Mandarin. Taipei*: Crane Publishing.

Wan, I. P. (2006). A psycholinguistic study of postnuclear glides and coda nasals in Mandarin. *Journal of Language and Linguistics*, 5(2), 158-176.

Wan, I. P., Allassonnière-Tang, M., & Yu, P. (2024a). Early Segmental Production in Thai Preschool Children Learning Mandarin. *International Journal of Asian Language Processing*, 34(02), 2450005.

Wan, I. P., Chang, C. W., Lee, C., & Yu, P. (2024b). Probability Distributions of Sounds and Phonotactics in Taiwan Mandarin Syllables. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation* (pp. 1157-1165).

Wan, I. P., & Jaeger, J. J. (2003). The phonological representation of Taiwan Mandarin vowels: A psycholinguistic study. *Journal of East Asian Linguistics, 12*(3), 205-257. DOI:10.1023/A:1023666819363

Wellman, B. L., Case, I. M., Mengert, I. G., & Bradbury, D. E. (1931). *Speech sounds of*

young children. University of Iowa Studies: Child Welfare.

Wiedenhof, J. (2015). Appendix A The International Phonetic Alphabet. In A Grammar of Mandarin (pp. 409-409). John Benjamins Publishing Company.

Wu, Y. (1994). M*andarin Segmental Phonology.* [Doctoral dissertation, University of Toronto]. ProQuest Dissertations Publishing.

Wu, T. M., Xu, Z. Y. (1979). A preliminary analysis of language development in children during the first three years. *Acta Psychologia Sinica, 11*(2), 153–165.

Zhang, J. Q. (2016). Nonsibilant Fricative Acquisition by Bilingual Guoyu-Taiwanese Southern Min Children (Master's thesis, The Ohio State University).

Zhu, H. (2002). Phonological Development in Specific Contexts: Studies of Chinese-speaking. Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781853595899

Zhu, H. & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). Journal of Child Language, 27(1), 3-42.

Appendix

I. Possible distribution of CG sequences by Wan (2003)

Table 2 shows the possible distribution of CG sequences by Wan (2003) shows the legal CG segment structures in Mandarin: asterisk * indicates a CG sequence is not possible.

	Bilabial	**Labial**	**Dental**	**Retroflex**	**Palatal**	**Velar**
Unaspirated plosive	pj, pw		tj, tw			*kj, kw
Aspirated plosive	p^hj, p^hw		t^hj, t^hw			*k^hj, k^hw
Fricative		*fj, *fw	*sj, sw	*ʂj, ʂw zj, zw	ɕj, ɕɥ, *ɕw	*xj, xw
Unaspirated affricate			ts^hj, ts^hw	*tʂj, tʂw	tɕj, tɕɥ, *tɕw	
Aspirated affricate				*$tʂ^h$j, $tʂ^h$w	$tɕ^h$j, $tɕ^h$ɥ, *$tɕ^h$w	
Nasal	mj, mw		nj, nɥ, nw			*j, *w

Table 2: The possible distribution of CG sequences by Wan (2003)

II. Participant information

Table 5 shows all 45 participant information.

III. Correctness and error distribution

Table 6 shows the correctness and error distribution types by the 45 participants.

ID	Gender	Age
1	M	0；7；9
2	F	1；5；22
3	F	1；8；2
4	M	1；8；28
5	F	1；9；19
6	M	2；4；0
7	F	2；4；5
8	M	2；4；21
9	M	2；5；1
10	F	2；5；18
11	M	2；6；8
12	M	2；9；13
13	F	2；9；24
14	M	2；10；6
15	F	2；10；8
16	F	3；0；9
17	M	3；1；24
18	M	3；4；26
19	M	3；5；28
20	M	3；7；17
21	F	3；9；22
22	F	3；10；11
23	M	3；10；23
24	M	3；11；20
25	F	4；1；25
26	M	4；2；26
27	M	4；5；4
28	M	4；5；15
29	F	4；6；23
30	F	4；6；29
31	F	4；7；27
32	F	4；8；0
33	F	4；8；7
34	M	4；8；13
35	F	5；1；4
36	M	5；2；0
37	F	5；2；23
38	M	5；2；25
39	M	5；3；27
40	F	5；8；4
41	F	5；8；4
42	M	5；9；11
43	M	5；9；21
44	F	5；9；27
45	F	5；10；26
SD=1.38；Average age=3.8		

Table 5: Participant information

IV. Error Types

Table 11 shows correctness and all the error types, along with the percentage out of all produced data

V. Error examples

The table of the proportion of error examples for each of the three glides below (Table 12) helps provide a clear data showing which phoneme each of the three glides is most frequently substituted with during the replacement process.

VI. Confusion Matrix

The confusion matrix below helps further identify specific error patterns of the three glides, [w], [j], and [ɥ], revealing which glides are most frequently misarticulated and the phonemes with which these glides most often interact.

Participant	Correctness	Substitution	Deletion	Addition	Item counts
1	137	0	5	2	144
2	47	1	3	1	52
3	128	3	18	2	151
4	72	6	15	0	93
5	93	3	21	0	117
6	151	0	2	0	153
7	112	2	6	0	120
8	97	18	21	0	136
9	116	0	7	0	123
10	160	1	0	0	161
11	146	13	7	0	166
12	136	1	8	0	145
13	156	2	1	1	160
14	146	0	10	2	158
15	113	0	1	2	116
16	147	6	1	1	155
17	122	0	11	0	133
18	155	1	2	0	158
19	104	0	6	2	112
20	151	0	0	0	151
21	147	1	9	0	157
22	150	0	5	0	155
23	115	1	2	0	118
24	162	0	0	0	162
25	152	0	2	0	154
26	160	0	0	0	160
27	155	1	0	2	158
28	161	0	1	0	162
29	158	0	0	0	158
30	162	0	0	0	162
31	138	6	0	2	146
32	160	0	0	0	160
33	158	0	0	0	158
34	108	2	0	0	110
35	162	0	0	0	162
36	162	0	0	0	162
37	158	0	0	0	158
38	150	0	2	0	152
39	162	0	0	0	162
40	157	0	1	0	158
41	147	5	1	0	153
42	157	1	0	0	158
43	156	0	0	0	156
44	144	2	2	0	148
45	162	0	0	0	162
Total	**6292**	**76**	**170**	**17**	**6555**

Table 6: Correctness and error distribution by participant

	Correctness	%	Substitution	%	Deletion	%	Addition	%	Sum(counts)
j	2898	97.22%	7	0.23%	62	2.08%	14	0.47%	2981
w	3102	96.07%	18	0.56%	106	3.28%	3	0.09%	3229
q	292	89.30%	33	10.09%	2	0.61%	0	0.00%	327
Total	6292	96.25%	58	0.89%	170	2.60%	17	0.26%	6537

Table 11: Error types

IPA	Correctness	%	Error Types	Error counts	%	Sum
j	2898	98%	j→w	1	0.067%	2967
			j→ɥ	0	0.000%	
			j→x	2	0.067%	
			j→i	2	0.067%	
			j→ə	1	0.034%	
			j→y	1	0.034%	
			j→#	62	2.090%	
w	3102	96%	w→j	1	0.031%	3226
			w→ɥ	0	0.000%	
			w→p	7	0.217%	
			w→p^h	1	0.031%	
			w→v	1	0.031%	
			w→ŋ	8	0.248%	
			w→#	106	3.286%	
ɥ	292	89%	ɥ→j	21	6.422%	327
			ɥ→w	1	0.306%	
			ɥ→n	1	0.306%	
			ɥ→y	10	3.058%	
			ɥ→#	2	0.612%	

Table 12: The proportion of error examples for each of the three glides

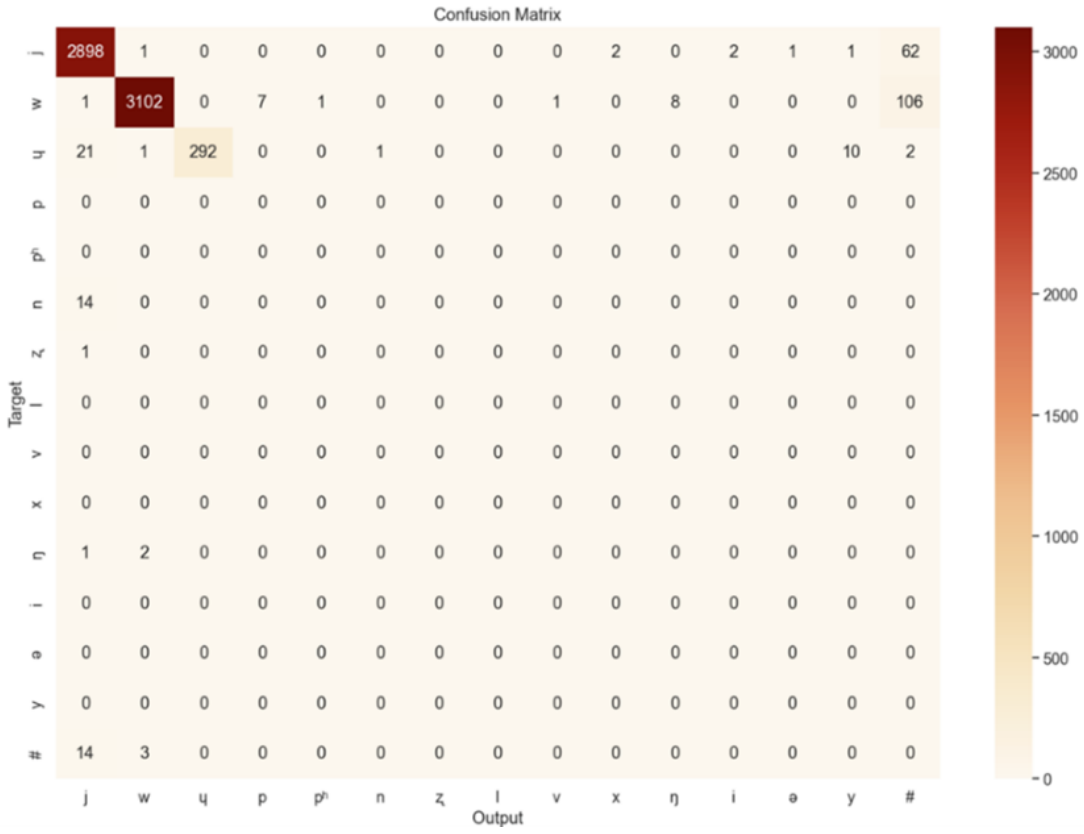


Figure 1: Confusion matrix (indicates that a phone is dropped, a deletion phenomenon)

Assessing GPT models' Sensitivity to Epistemic Meanings in Korean Periphrastic Construction

Yebin Lee[1], Arum Kang[2], Sanghoun Song[1]
yblee1018@korea.ac.kr, arkang@cnu.ac.kr, sanghoun@korea.ac.kr
[1]Korea University [2]Chungnam National University

Abstract

This study investigates whether large language models can process epistemic modality in Korean, where degrees of certainty are often expressed through periphrastic constructions with interrogative complementizers and epistemic predicates. Using GPT-4.1, two experiments tested the model's certainty judgments with and without contextual cues. Without context, the model consistently defaulted to a 50% certainty across different predicates, suggesting that its responses are categorical in nature. However, with context, responses became more varied and partly human-like, but still lacked the gradient sensitivity observed in human speakers. Further analysis revealed an exceptional pattern for the word *siph-* 'seem/believe', but this likely stems from the frequency and familiarity of the expression in the model's training corpus, leading to a holistic representation, rather than reflecting a genuine understanding of the semantic distinctions introduced by different interrogative complementizers. These results indicate that, while the model can respond to explicit contextual signals, it does not appear to encode the internal semantic distinctions that native speakers associate within epistemic modal meaning and Korean interrogative complementizers.

1 Introduction

What is unique about human language, compared to animal communication, is its ability to convey information beyond immediate reality. This ability enables humans to make counterfactual statements, such as lies or hypothetical scenarios. For example, look at example 1 below.

(1) a. John may be in his office.
b. John must be in his office.
c. John'll be in his office. Palmer (2001)

Example (1) illustrates epistemic modality in English. In (1a), the speaker expresses uncertainty about whether John is in his office by using the modal verb 'may'. In contrast, in (1b), the speaker conveys a firm judgment through the modal verb 'must'. Finally, in (1c), the speaker makes a judgment based on what generally happens with John. This contrast shows how epistemic modality enables speakers to modulate their degree of certainty, making communication more nuanced and context-sensitive. Not all languages encode epistemic modality morphologically, and the availability and granularity of such distinctions vary across languages.

In Korean, epistemic modality sometimes appears in periphrastic constructions, which are constructions in which multiple words function as a single grammatical unit. Notably, Korean has four interrogative complementizers: *-nci*, *-lci*, *-nka*, and *-lkka*. Each interrogative complementizers can be divided into more fine-grained elements, each carrying distinct semantic nuances. For instance, complementizers ending in *-kka* are more actively used in modalized questions, compared to those ending in *-ci*. In addition, complementizers with *-n-* typically mark realis or present meaning, whereas those with *-l-* mark irrealis or future meaning. However, when these complementizers combine with certain predicates, subtle semantic differences lead to variations in the level of epistemic commitment. Consider the example sentence below.

(2) Minci-ka nayil hakkyo-ey
Minci-NOM tomorrow school-LOC
o-nunci / o-lci / o-nunka / o-lkka
come-PRES/FUT/PRES/FUT.**whether**
kwungkumha-ta.
wonder-DECL
'I wonder whether Minci would come to the party.'

Each of the four complementizers combined with *o-* 'come' has its own distributional constraints on

the following predicates (Kang and Song, 2021). Specifically, *-nci* and *-lci* are ordinary interrogative complementizers compatible with factive or neutral predicates, while *-nka* and *-lkka* function as modalized (subjunctive) complementizers that yield weaker epistemic commitment and conjectural readings. These distinctions reveal that the complementizers encode different degrees of speaker commitment within epistemic modality. Because these semantic nuances are subtle distinctions that only native speakers can reliably perceive, determining whether a Korean-trained language model can discriminate among them is an important challenge in computational linguistics.

Therefore, based on the experimental design and findings of Kang and Song (2021), this study aims to determine whether large language models can distinguish between complementizers in periphrastic constructions and assign a similar degree of certainty to humans. We concluded that language models fail to distinguish between the different meanings of the interrogative complementizer, although they are able to capture the overall degree of certainty expressed by the sentence. This result shows that language models process the certainty as a whole, not a combination of the complementizer and the predicate.

The rest of this paper is organized as follows. Section 2 presents the theoretical and computational background of the study. The theoretical background focuses on epistemic modality and the subjunctive mood in Korean periphrastic constructions, while the computational background addresses the concept of certainty in language models. Section 3 details the experimental setup, including the dataset, model, and procedure. Section 4 presents the results of all experiments, along with a discussion. Finally, Section 5 concludes with a summary and a discussion of the study's limitations.

2 Background

2.1 Linguistic Background: Epistemic Modality

Epistemic modality refers to the speaker's attitude toward the reality or likelihood of a given proposition, expressing how strongly the speaker believes in its truth. In Korean, this is conveyed through various linguistic devices, including verbal endings such as *-keyss-* 'will', periphrastic constructions like *-(u)l get* 'may', and adverbs such as *ama* 'maybe' and *cheoldae* 'never' (Son, 2016). These expressions not only signal the degree of certainty or inference but also reflect the source and nature of the speaker's knowledge, whether it is derived from direct observation, memory, reasoning, or hearsay. Epistemic modality thus functions as both a semantic and pragmatic system, providing insight into how Korean speakers evaluate and encode their access to information.

Additionally, epistemic modality plays a key role in Korean interrogatives through so-called modalized questions. Recent studies such as Kang and Yoon (2019) show that sentence-final particles such as *-lkka* and *-nka* function not only as interrogative markers but also as indicators of epistemic uncertainty and emotional involvement. For instance, the question *o-ass-ulkka* 'might come'? not only seeks information but also conveys the speaker's speculation or concern. These constructions demonstrate that modality and interrogativity are tightly integrated at the clause level in Korean, forming a complex system that encodes both grammatical and affective meanings.

Another key theoretical notion underlying Korean epistemic constructions is nonveridical equilibrium (Kang and Yoon, 2020). Nonveridical equilibrium refers to a semantic state in which a proposition is evaluated as neither fully true nor fully false within the speaker's epistemic space, allowing both p and $\neg p$ worlds to remain accessible. This equilibrium characterizes linguistic environments where the speaker's commitment is suspended, such as conjecture, doubt, or emotional speculation. Specifically, complementizers like *-nka* and *-lkka* instantiate this nonveridical equilibrium by encoding a weakened truth commitment and balancing the likelihood of opposing possibilities. Thus, these complementizers not only mark interrogativity but also convey a specific epistemic stance of suspended commitment, distinguishing them from indicative forms like *-nci* or *-lci*, which presuppose veridical or factive evaluation.

To empirically verify these theoretical claims, Kang and Song (2021) conducted a collostructional analysis using the Sejong Corpus (National Institute of the Korean Language, 2009) and two follow-up experiments: an acceptability judgment task and a context-sensitive evaluation. Their results demonstrated that *-lkka* and *-nka* exhibited high collostructional strength and acceptability when paired with nonveridical predicates such as *siph-* 'seem' or 'believe', *kekcengsulep-* 'worried', and *molu-* 'do not

know', with corresponding lower certainty judgments (10–50%). In contrast, *-nci* and *-lci* appeared more frequently with veridical predicates like *al-* 'to know' and *hwaksinha-* 'be certain', and were associated with stronger commitment. These findings provide quantitative support for the hypothesis that the Korean subjunctive is marked at the complementizer level and that these markers carry systematic semantic effects in relation to certainty.

Building on this prior work, the present study extends the experimental design of Kang and Song (2021) by applying it to a large language model, especially for GPT series. Rather than relying solely on theoretical constructs, this research uses their empirical predicate-complementizer pairings and replicates their acceptability and context-based judgment tasks under comparable conditions. This allows for a direct comparison between human and model behavior, providing a more grounded assessment of how LLMs interpret degrees of certainty in structurally complex Korean constructions.

2.2 Computational Background: Language Model Uncertainty

For language models, certainty refers to the ability to express the degree of confidence in their judgments based on internal knowledge. This involves more than simply providing correct answers; it also encompasses distinguishing between known and unknown information, and using appropriate linguistic cues to convey uncertainty in ambiguous or underspecified contexts. Such capacity is essential in high-stakes domains like medical consultations and legal reasoning, where reliable communication is crucial.

In recent surveys, certainty in language models has been characterized as encompassing two interrelated aspects: the model's internal confidence in its output and the explicit linguistic expressions it employs to signal that confidence. Both dimensions are central to understanding how models handle epistemic judgments. The importance of certainty lies not only in reducing hallucinations by preventing overconfident yet inaccurate statements, but also in revealing how closely model behavior aligns with human cognitive processes.

Recent studies have pointed out structural limitations in how language models express or evaluate certainty. Suzgun et al. (2024) assessed models such as GPT-4, Claude-3, and LLaMA-3 and found that, while these models perform well on fact-based questions (Such as a question that starts with "Is it true that ..."), their accuracy drops significantly on tasks involving falsehoods or beliefs. This issue is particularly severe when models are asked to affirm beliefs that contradict factual information or are expressed in the first person (Such as a question that starts with "I believe ..., Do I believe ... ?"). Extending this line of inquiry, Li et al. (2025) investigated how models distinguish among fact, fiction, and forecast, and evaluated their use of evidence-based certainty expressions. Despite the fluency of their outputs, models often failed to select expressions that matched the strength of evidence or epistemic commitment, indicating a limited understanding of epistemic modality.

Language models thus lack a coherent internal mechanism to determine what they know and what they do not. As a result, they frequently make confident statements regardless of the information's actual reliability. Yona et al. (2024) introduced the concept of faithful response uncertainty, which quantifies the mismatch between a model's internal confidence and the decisiveness of its verbal output. Their findings show that models often express high certainty even when internally uncertain, potentially misleading users. Similarly, Krause et al. (2023), in a multilingual QA setting, found that GPT-3.5 tended to output high-confidence responses regardless of accuracy, especially in low-resource languages, revealing a persistent overconfidence bias.

To address this issue, various methods have been proposed to estimate or calibrate uncertainty in LLMs, including probability-based scoring, uncertainty quantification, and prompt-based self-assessment. However, most are limited to numerical scores or token-level probabilities, and linguistically grounded approaches that examine how certainty is expressed in natural language remain underexplored. Xia et al. (2025) provide a comprehensive survey of uncertainty estimation methods and highlight that LLMs still struggle to communicate what they do or do not know. Moreover, most prior work has focused on English, leaving a gap in understanding how models handle epistemic reasoning in languages with distinct grammatical systems for encoding certainty. Before applying advanced calibration methods, it is therefore necessary to first determine whether a model can recognize and appropriately use the linguistic markers of certainty in a given language. This study addresses that question by examining whether a language model can identify and produce the complementizer–predicate

combinations that encode degrees of certainty in Korean.

3 Method

3.1 Data

The dataset used in this study consists of full Korean sentences constructed using six epistemic predicates and four complementizers. As described in Section 2.1, epistemic modality in Korean is conveyed through periphrastic constructions that pair a complementizer with an epistemic predicate. However, as noted earlier, not all predicate–complementizer combinations are grammatically acceptable, as their compatibility is constrained by the semantic class of the predicate. For example, the complementizer *-nci*, which presupposes a strong belief that the event has occurred, cannot combine with counterfactual predicates such as *siph* 'believe'.

Acceptable combinations were selected based on Kang and Song (2021), who evaluated each construction's acceptability using a 5-point Likert scale and then converted the scores to Z-scores. To quantify the relative acceptability of each construction, we applied the cumulative distribution function (CDF) of the standard normal distribution. The resulting values range from 0 to 100, representing the percentile rank of each construction. To ensure that sentence acceptability would not confound the experimental results, we included only those combinations whose CDF-based percentile exceeded 50%. Notably, the predicate *sayangakha-* 'think' was selected to represent the 90% certainty condition, based on its high acceptability. For the *uysimsulep-* 'doubt', although it was not included in the acceptability test of Kang and Song (2021), it was included in the present study because its semantic property naturally conveys a very low degree of epistemic certainty.

Because the outputs of language models are inherently stochastic, it is difficult to generalize from a single trial. To address this, we artificially constructed a sufficient number of examples for robust evaluation. Specifically, 100 sentence pairs were handcrafted for each of the 16 valid predicate-complementizer combinations, using different verbs and nouns, yielding a total of 1,600 sentences. The complete list of selected predicates and complementizers is provided below (Table 1), and a single representative sentence pair is included in Appendix A.

	-nci	*-nka*	*-lci*	*-lkka*
kwungkumha- 'wonder'	**85.5**	**69.3**	**78.3**	**73.6**
kekcengsulep- 'worried'	**69.8**	49.4	**74.6**	**76.0**
molu- 'not know'	**63.6**	**52.9**	**68.9**	**69.9**
siph- 'seem', 'believe'	39.3	**53.0**	**59.7**	**74.1**

Table 1: Acceptability of the Periphrastic Constructions

3.2 Experiment

An experiment in this study was conducted following Kang and Song (2021). In their study, the experiment was conducted in two ways: first, to identify acceptable combinations within the periphrastic construction, and second, to determine the degree of certainty each combination conveys. The first experiment was an acceptability task in which participants judged the acceptability of given sentences. The second experiment was context-based, in which participants answered yes/no questions about whether a sentence was appropriate for a given interpretation. The main framework of the present study follows the latter approach, which is further divided into two specific subtypes.

The present study aims to achieve two goals. First, to determine whether the presence of contextual information influences the model's certainty judgments. Second, to assess whether the language model exhibits patterns similar to those reported in Kang and Song (2021). The experiment was designed analogously to the human study, consisting of two settings: a context-free task and a context-based task. In the context-free task, the model was prompted to provide an answer reflecting its level of commitment (10%, 50%, or 90%) without any supporting context. In the context-based task, the model was presented with specific discourse contexts and asked to respond with yes or no, indicating whether a given probability level of commitment (10%, 50%, or 90%) was appropriate for the situation. Both tasks require the model to evaluate the certainty of a sentence, but differ in whether they include contextual cues.

In the context-free task, the model is presented with a sentence in isolation and asked to choose a certainty level among 10%, 50%, or 90%. In contrast, the context-based task provides the model with three components: a question, a sentence, and a context. The question takes the form of a yes or no interrogative, such as "Does the following sentence accurately reflect the speaker's thoughts in response to the interlocutor's question?" The sentence is a declarative containing a complementizer–predicate combination that fits the given ques-

tion and context. The context provides information necessary to evaluate the plausibility of the sentence, indicating one of three likelihood levels (10%, 50%, or 90%). This task differs from the context-free one in that it provides an explicit epistemic frame for interpretation. Consequently, the model's response format also changes: while the context-free task requires selecting among scalar degrees of certainty, the context-based task asks for a binary judgment ("yes" or "no") on whether the sentence is appropriate given the contextual information.

3.3 Model

The language models used in this experiment are the OpenAI GPT-4 series: GPT-4.1, GPT-4o, and GPT-4.1 mini(Achiam et al., 2023). These models were selected not to compare their performance, but to evaluate whether state-of-the-art language models can distinguish different degrees of certainty based on Korean linguistic cues. All test sentences were delivered via API calls with a base temperature of 0, and no prior conversational context or inference state was provided. Each request included a shared system prompt: *"Your role is to evaluate the certainty of the following sentences as a native Korean speaker."* The maximum token length for each response was set to 50. For the sake of space and clarity, this paper reports only the results from GPT-4.1.

4 Result

4.1 Task1 & Task2 Comparison

From Table 2, we can see that the results from Task 1 and Task 2 are different in a certain way. In particular, when comparing the overall distributions between context-free and context-based tasks, a clear contrast emerges. In the absence of contextual signals, the model exhibited an overwhelming preference for certainty 50% in almost all items, suggesting either epistemic default or indecision. However, once context was introduced, responses became more varied and better aligned with human patterns. The presence of contextual information appears to activate a dormant sensitivity in the model, enabling it to shift away from the 50% default and select 10% or 90% when appropriate.

Notably, predicates such as *uysimsulep-* 'doubt' and *sayngkakha-* 'think' demonstrated striking improvements in alignment with human expectations. In the case of *uysimsulep-* 'doubt', the model began selecting 10% with much greater frequency in the context-based task, mirroring human responses reported by (Kang and Song, 2021). Likewise, for *sayngkakha-* 'think', a predicate typically associated with high certainty, the model shifted from an uncertain stance in the context-free task to confidently selecting 90% when a supportive context was provided. These shifts suggest that contextual embedding plays a crucial role in enabling the model to simulate human-like gradience in epistemic judgment.

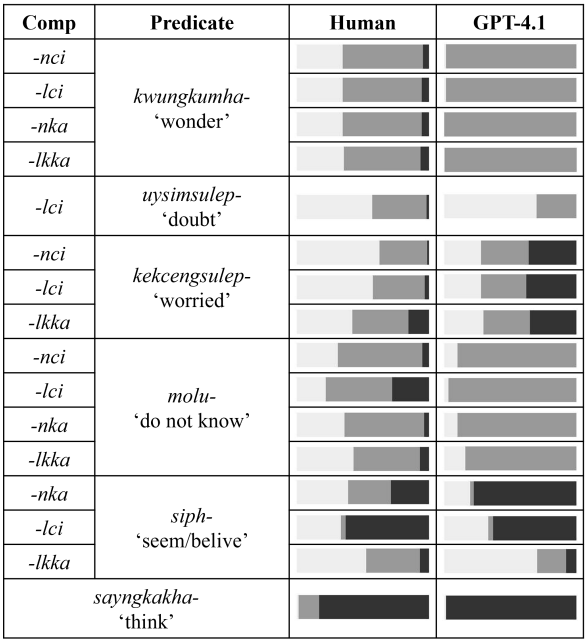


Figure 1: Overall result of Fill-in-the-Blank Task. The white, gray, and black bars represent the proportions of "yes" responses for 10%, 50%, and 90% commitment levels, respectively.

In sum, this section demonstrates that while the language model fails to exhibit gradient reasoning in isolation, the addition of contextual information enables it to move toward more human-aligned responses. The clearest evidence of this is the redistribution of responses across the 10%–50%–90% scale once context is available, with predicate-specific sensitivity emerging most notably for epistemically polarized verbs. This provides the first key finding of the study: that contextual grounding facilitates a more gradient and human-like pattern of certainty estimation in LLM.

4.2 GPT & Human Comparison

However, when compared to human responses, several critical discrepancies remain. Figure **??** presents a comparative visualization between

Comp	Predicate	Context-Free Task			Context-Based Task		
		10%	**50%**	**90%**	**10%**	**50%**	**90%**
-nci	*kwungkumha-* 'wonder'	0	100	0	1	100	0
-lci		0	100	0	1	100	0
-nka		0	100	0	0	99	0
-lkka		0	100	0	0	100	0
-lci	*uysimsulep-* 'doubt'	2	89	9	99	0	0
-nci	*kekcengsulep-* 'worry'	0	99	1	10	13	13
-lci		0	99	1	16	20	22
-lkka		0	99	1	17	20	20
-nci	*molu-* 'do not know'	0	100	0	11	100	0
-lci		0	100	0	3	100	0
-nka		0	100	0	11	100	0
-lkka		0	100	0	19	100	0
-nka	*siph-* 'seem/believe'	0	100	0	7	1	28
-lci		0	100	0	19	2	36
-lkka		0	100	0	76	24	8
-	*sayngkakha-* 'think'	0	88	12	1	0	90

Table 2: Result from Context-Free Task and Context-Based Task. In the context-free task, the language model was asked to choose which of the three certainty levels (10%, 50%, or 90%) best matched the given sentence; thus, the portions across the three levels sum to 100. In contrast, the context-based task required independent judgment of whether each sentence accurately reflected the context corresponding to 10%, 50%, or 90% certainty, resulting in separate counts for each condition.

model outputs and human acceptability-based judgments under context-provided conditions. Here, the light gray bars indicate the proportion of "yes" responses at the 10% certainty level, the medium gray bars for 50%, and black bars for 90%. At a glance, GPT-4.1's pattern appears superficially aligned with human ratings, particularly for certain predicates. However, closer inspection reveals notable differences in granularity and sensitivity.

First, the model lacks the gradient distribution across certainty levels that characterizes human responses. For example, in predicates like *kwungkumha-* 'wonder' and *molu-* 'do not know', human participants displayed a probabilistic spread, with secondary selections at 10% or 90% even when 50% was most frequent. In contrast, the model's responses clustered almost exclusively around 50%, failing to exhibit the nuanced variation observed in human reasoning. This one-sided concentration suggests a categorical bias in model predictions, in which a single certainty level dominates across instances of a given predicate. Second, although humans varied their judgments depending on the complementizer, even after controlling for the predicate (e.g., *kwungkumha-* 'wonder' vs. *kekcengsulep* 'worried'), the model did not show this sensitivity. While human participants adjusted their certainty based on fine-grained morphosyntactic cues, GPT-4.1 treated different complementizers more uniformly, leading to flatter variation. This indicates that, unlike human judgments, which are jointly shaped by the predicate and complementizer, model responses are less influenced by the compositional semantics of these constructions. Third, the predicate *kekcengsulep-* 'worried' showed the most distinct divergence. Whereas human judgments predominantly clustered around 10%, the model provided an ambiguous distribution, often split across categories without a strong preference. This discrepancy becomes clearer when examining the model's justifications: GPT-4.1 frequently classified *kekcengsulep-* 'worried' as an emotion-descriptive predicate rather than an epistemic one, asserting that it is not appropriate for evaluating certainty. This semantic misclassification suggests that the model does not recognize the epistemic implications embedded in emotion predicates like worry. Finally, one notable exception is observed with the predicate *siph-* 'seem' in conjunction with the complementizer *-lkka*, forming the construction

-lkka siph-. Here, the model's response distribution aligned closely with human data, accurately reflecting a lower degree of certainty. This suggests that, in certain cases where complementizer–predicate collocations are highly conventionalized, the model can replicate human-like interpretations. However, such alignment remains the exception rather than the rule.

5 Conclusion

This study examined whether large language models can distinguish different levels of certainty expressed through Korean periphrastic constructions. The experiment was conducted in two ways: a context-free task and a context-based task. In the context-free setting, the model consistently preferred a neutral judgment of 50%, regardless of the sentence form. When context was provided, the model began to show more variation in its responses and partially aligned with human judgments for certain combinations. However, the model did not capture the gradual variation that human speakers exhibited, suggesting a limited understanding of fine-grained certainty comprehension in Korean.

Notably, a closer comparison between human and model responses revealed several persistent divergences. First, while human responses exhibited gradient distributions across all certainty levels, the model's predictions were frequently concentrated on a single value, especially 50%. Second, human responses showed variation based on both the predicate and the complementizer, reflecting sensitivity to their interaction, whereas the model's outputs appeared largely invariant across complementizers for a given predicate. Third, for certain epistemic predicates such as *kekcengsulep-* 'worried', the model not only failed to align with human judgments but also rationalized its misclassification by labeling the predicate as non-epistemic, thereby revealing limitations in semantic categorization. These discrepancies underscore that even when context is provided, the model lacks the interpretive mechanisms necessary to represent the probabilistic and compositional aspects of Korean epistemic modality. However, the experiment also revealed only a few cases of convergence. For instance, in the expression *-lkka siph-*, the model produced outputs that closely resembled human judgments. This suggests that when complementizer–predicate pairs co-occur frequently or are structurally salient, language models may succeed in mimicking human-like certainty evaluations. Nevertheless, such instances remain isolated exceptions rather than generalizable patterns.

While these findings provide valuable insight, several limitations remain. Certainty is inherently a gradient rather than a binary concept, and further work is needed to capture this gradience more effectively. In this study, the focus was on complementizers and predicates; future analyses should also examine how epistemic constructions behave across syntactic positions, such as in embedded versus main clauses. Additionally, unlike other predicates, *siph-* 'seem' exhibited a notable tendency for the model to select *-lkka* in 10% certainty contexts, indicating a different complementizer selection pattern from predicates such as *kwungkumha-* 'wonder' or *molu-* 'do not know'. Given the unique semantic behavior of *siph-* 'seem' in Korean, further investigation is warranted. Finally, although Korean has distinctive features in expressing epistemic modality, it employs such markers less frequently than languages with richer epistemic systems, such as Italian. To determine whether language models encode the concept of certainty in a cross-linguistic sense, experiments should be extended to other languages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arum Kang and Sanghoun Song. 2021. A study on subjunctive mood in korean: Using corpus and experimental linguistic data. In *23rd Seoul International Conference on Generative Grammar*. The Korean Generative Grammar Circle. Written in Korean.

Arum Kang and Suwon Yoon. 2019. The subjunctive complementizer in korean: The interaction between inquisitiveness and nonveridicality. In *Proceedings of the 12th Generative Linguistics in the Old World & the 21st Seoul International Conference on Generative Grammar*, pages 343–358.

Arum Kang and Suwon Yoon. 2020. From inquisitive disjunction to nonveridical equilibrium: Modalized questions in korean. *Linguistics*, 58(1):207–244.

Lea Krause, Wondimagegnhue Tufa, Selene Báez Santamaría, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: exploring the calibration and expression of (un) certainty of large language models in a multilingual setting. In *Proceedings*

of the workshop on multimodal, multilingual natural language generation and multilingual WebNLG Challenge (MM-NLG 2023), pages 1–9.

Meng Li, Michael Vrazitulis, and David Schlangen. 2025. Representations of fact, fiction and forecast in large language models: Epistemics and attitudes. *arXiv preprint arXiv:2506.01512*.

National Institute of the Korean Language. 2009. The 21st century sejong project corpus. https://korean.go.kr.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge university press.

Hyeok Son. 2016. A study on use of epistemic modality markers. *Language Facts and Perspectives*, 39:249–285. Written in Korean.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the machine: Investigating epistemological blind spots of language models. *arXiv preprint arXiv:2410.21195*.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

A Example Single Sentence Set For 16 Predicate–Complementizer Combinations

Comp	Predicate	Sentences
-nci	*kwungkumha-* ‘wonder’	Minci-ka nayil hakkyo-ey o-nunci kwungkumha-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci kwungkumha-ta Minci-NOM tomorrow school-LOC come-FUT.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka kwungkumha-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka kwungkumha-ta Minci-NOM tomorrow school-LOC come-FUT.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lci	*uysimsulep-* ‘doubt’	Minci-ka nayil hakkyo-ey o-lci uysimsulep-ta Minci-NOM tomorrow school-LOC come-FUT.whether be.doubtful-DECL *‘I doubt Minci will come to school tomorrow.’*
-nci	*kekcengsulep-* ‘worry’	Minci-ka nayil hakkyo-ey o-nunci kekcengsulep-ta Minci-NOM tomorrow school-LOC come-PRES.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci kekcengsulwe-ta Minci-NOM tomorrow school-LOC come-FUT.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-nunka kekcengsulep-ta Minci-NOM tomorrow school-LOC come-PRES.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-nci	*molu-* ‘do not know’	Minci-ka nayil hakkyo-ey o-nunci molu-keyss-ta Minci-NOM tomorrow school-LOC come-PRES.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci molu-keyss-ta Minci-NOM tomorrow school-LOC come-FUT.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka molu-keyss-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka molu-keyss-ta Minci-NOM tomorrow school-LOC come-FUT.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lci	*siph-* ‘seem/belive’	Minci-ka nayil hakkyo-ey o-lci siph-ta Minci-NOM tomorrow school-LOC come-FUT.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka siph-ta Minci-NOM tomorrow school-LOC come-PRES.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka siph-ta Minci-NOM tomorrow school-LOC come-FUT.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-	*sayngakha-* ‘think’	Minci-ka nayil hakkyo-ey o-l-gerago sayngkakha-n-ta Minci-NOM tomorrow school-LOC come-FUT-will think-PRES-DECL *‘I think Minci will come to school tomorrow.’*

B Prompt Used in Each Experiment

Task	English Translation	Korean (Original)
Context-Free Task	Respond with one of the following options, indicating the degree of certainty of the speaker in the given sentence: 10% (very unlikely), 50% (uncertain), or 90% (very likely). Be sure to respond with only one of the three values, and explain the reason in a single sentence along with your answer. **Sentence:** I wonder if Minci will come to school tomorrow.	주어진 문장의 발화자가 확신하는 정도를 10% (가능성이 거의 없음), 50% (전혀 예상할 수 없음), 90% (매우 가능성이 큼) 중 하나로 응답하세요. **문장:** 민지가 내일 학교에 오는지 궁금하다.
Context-Based Task	Does the following sentence fit the given context? Start your answer with either 'yes' or 'no', and explain your reason in a single sentence. **Context(90%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Context(50%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Context(10%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Question:** Does the following sentence appropriately reflect Sun-i's thoughts in response to In-ho's question, "Will Minci come to school tomorrow?" **Sentence:** I wonder if Minci will come to school tomorrow.	다음 문장은 맥락(context)에 잘 부합합니까? '예.' 또는 '아니오.'로만 시작하고, 이유는 대답과 함께 하나의 문장으로 설명하세요. **맥락(90%):** 순이는 민지가 내일 학교에 올 가능성이 매우 크다고 생각한다. **맥락(50%):** 순이는 민지가 내일 학교에 올지 안 올지 전혀 예상할 수 없다. **맥락(10%):** 순이는 민지가 내일 학교에 올 가능성이 거의 없다고 생각한다. **질문:** 인호의 "민지가 내일 학교에 올까?"라는 질문에 다음 문장은 순이의 생각을 잘 반영하는가? **문장:** 민지가 내일 학교에 오는지 궁금하다.

For the Context-Based Task, only single context was provided at a time, and the model was asked to respond with Yes or No.

Speech-Like Cues and the Limits of Musicality: Lexical Tone Normalization in Mandarin across Speech, Rap, and Song Contexts

Yujia Tian[1], **Yanyuan Ye**[1], **Mency Lu**[1],
JIA Fanlu[2], Ran Tao[1]
[1]Department of Language Science and Technology,
Research Centre for Language, Cognition and Neuroscience,
The Hong Kong Polytechnic University
[2]Department of Psychology, Jinan University
yujaa.tian@connect.polyu.hk, yanyuan.ye@connect.polyu.hk
mency.lu@connect.polyu.hk, spe_jiafl@ujn.edu.cn
ran.tao@polyu.edu.hk

Abstract

Lexical tone normalization enables Mandarin speakers to maintain stable tone categories despite substantial pitch variability across speakers and communicative contexts. While previous research has established that speech contexts reliably facilitate tone normalization, non-speech and purely musical contexts do not, supporting the view that this mechanism is speech-specific. However, genres such as rap, which blend speech and musical elements, challenge this dichotomy. This study systematically examined whether rap music and related vocal contexts can induce lexical tone normalization in Mandarin listeners. Native Mandarin speakers categorized target syllables following six types of auditory contexts: natural speech, clear speech (elocution), typical rap, melodic rap, song, and cello. All vocal materials were produced by professional Mandarin-speaking rappers and pitch-matched to the targets. Results revealed that natural speech, elocution, and typical rap contexts robustly elicited tone normalization, as indicated by significant shifts in categorical boundaries and improved identification accuracy. In contrast, melodic rap and song produced only marginal effects, while the cello context had minimal impact. These findings indicate that speech-like cues—particularly prosody and articulation—are critical for tone normalization, whereas increasing musicality, especially melodic structure, can inhibit this process even when some speech-like features remain. Our results refine current models of speech perception by demonstrating that lexical tone normalization depends on the presence and prominence of linguistic cues, and that melody can impose clear boundary conditions on this perceptual adjustment.

1 Introduction

Speech perception is a fundamental cognitive process that enables listeners to extract meaningful linguistic units from continuous acoustic signals. For speakers of tonal languages such as Mandarin, this process is further complicated by the need to distinguish lexical meanings based on pitch patterns, or tones. Lexical tone normalization is a crucial perceptual mechanism that allows listeners to maintain stable tone categories despite substantial variability in pitch across different speakers and communicative contexts (Li, Chen, & Wong, 2021; Peng, 2006). This ability is essential for effective spoken communication in tonal languages, where even subtle pitch differences can alter word meaning.

Variability in speech arises from numerous sources, including anatomical differences, speaker identity, emotional state, and environmental context (Peng et al., 2012). Such inter- and intra-speaker variability can obscure phoneme boundaries and pose significant challenges for listeners. To overcome these challenges, listeners rely on contextual information to normalize and categorize phonemic units, a process known as talker normalization (Leather, 1983; Wong & Diehl, 2003). Despite significant acoustic variation across talkers, listeners can recognize words, highlighting the importance of contextual cues.

Previous research has shown that speech contexts rich in linguistic and prosodic cues facilitate robust tone normalization (Leather, 1983; Zhang, Peng, & Wang, 2012). In contrast, non-speech and purely musical contexts tend to elicit little or no normalization effect, supporting the view that tone normalization is governed by speech-specific

mechanisms (Peng et al., 2012; Tao et al., 2021). However, the boundary between speech and music is not always clear-cut. Rap music, for example, occupies a unique position at the intersection of speech and music, combining articulatory and prosodic features of spoken language with musical elements such as rhythm and, at times, melody.

In this study, "typical rap" is defined as a hybrid genre that combines rhythmic speech with musical accompaniment. According to the Oxford English Dictionary and Encyclopedia Britannica, rap is characterized by spoken or chanted rhyming lyrics over a musical backing. Our focus is on the continuum between speech and music, rather than a strict categorical distinction. By examining rap as an intermediate form, we aim to explore how speech-like and musical cues interact in tone normalization.

Theoretical perspectives on talker normalization have evolved over time. The "frame of reference" theory, originally developed for vowel perception, posits that listeners use contextual information to create a cognitive reference for interpreting speech sounds (Ladefoged & Broadbent, 1957; Nearey, 1978). This theory has been extended to lexical tone perception, where contextual pitch information provides a reference for categorizing tones (Wong & Diehl, 2003; Zhang et al., 2012). However, it remains unclear whether non-speech or hybrid contexts, such as rap, can provide an effective frame of reference for tone normalization.

Recent studies have begun to explore the influence of musical contexts on tone perception. While instrumental music generally fails to induce tone normalization (Tao & Peng, 2020; Zhang et al., 2013), the effects of vocal music, especially genres that blend speech and music, are less well understood. Rap, as a genre, is characterized by rhythmic speech delivered over a musical backing, often with minimal melodic content. This hybrid nature raises important questions about the limits of speech-specific processing in tone normalization: Can rap music, with its strong speech-like qualities, induce lexical tone normalization in Mandarin listeners? Or does the presence of musicality, even in a speech-like context, inhibit this perceptual adjustment?

Our previous research (Tian, Ye, Lu, Jia, & Tao, 2024) found that rap does not impede lexical tone normalization, whereas purely musical contexts fail to trigger this effect. This suggests the existence of a threshold beyond which musical variability becomes detrimental to language comprehension. However, the precise boundary between speech-like and musical cues, and their respective roles in tone normalization, remain unclear.

The present study aims to systematically investigate the role of rap music and related vocal contexts in lexical tone normalization among Mandarin speakers. Specifically, we seek to determine whether speech-like cues—such as prosody, articulation, and prosodic structure—are sufficient to trigger tone normalization, or whether increasing musicality, particularly melodic complexity, imposes boundary conditions that limit this effect. To this end, we constructed six types of auditory contexts: natural speech, elocution (clear speech), typical rap, melodic rap, song, and cello (instrumental). All vocal contexts were produced by professional Mandarin-speaking rappers to ensure consistency of voice quality and prosody. The contexts were carefully pitch-matched to the target stimuli, which consisted of Mandarin syllables varying along a tone continuum. Native Mandarin-speaking participants, with no exposure to other Chinese dialects, were presented with each context followed by a target syllable and asked to categorize the lexical tone.

We hypothesize that while the brain's language processing systems can accommodate some degree of musical variability, excessive variability may disrupt the normalization of lexical tones. Understanding the balance between musical variability and lexical tone normalization is crucial for advancing our knowledge of language processing. To further elucidate the neural mechanisms underlying this interaction, we recorded EEG data alongside behavioral experiments.

In summary, this study seeks to refine our understanding of lexical tone normalization by exploring the effects of a continuum of auditory contexts, ranging from speech to music, on Mandarin tone perception. By systematically varying the degree of musicality in the context, we aim to identify the boundary conditions that govern the effectiveness of speech-like cues in facilitating tone normalization.

2 Methodology

We utilized a similar experimental design and stimuli as in previous research (Tao et al., 2021 ; Tian, Ye, Lu, Jia, & Tao, 2024). Below is a

brief overview of the stimuli preparation and experimental procedure; for more detailed information, refer to (Zhang et al., 2013; Zhang et al., 2017).

2.1 Participants

A total of 21 native Mandarin speakers participated in our experiment, divided into two groups: a pre-experiment group (n = 5; 3 females; mean age = 21.6 years, SD = 2.79) and a formal experiment group (n = 16; 8 females; mean age = 21.8 years, SD = 3.3). All participants were university students or recent graduates from Northern China, ensuring a high degree of linguistic homogeneity and minimizing potential confounds from regional dialect exposure. Participants were screened to confirm that Mandarin was their sole language of daily communication, and none reported any exposure to other Chinese dialects or foreign languages that might influence tone perception.

All participants were right-handed, as assessed by the Edinburgh Handedness Inventory, and had normal or corrected-to-normal vision. Prior to the experiment, participants completed a health questionnaire to exclude those with a history of hearing impairment, tinnitus, neurological disorders, or language-related difficulties. All participants had at least a college diploma, and none had received formal musical training, which could potentially affect their sensitivity to musical cues. All provided written informed consent, and the study protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

2.2 Stimuli

The experimental stimuli comprised six context conditions: natural speech, elocution (clear speech), typical rap (Rap R), melodic rap (Rap M), song, and cello. The five vocal contexts were produced by four professional Mandarin-speaking rappers (two males, two females), each with over five years of professional experience and a record of Mandarin-language performances. All speakers had released albums or singles and performed at various events, ensuring high-quality and consistent vocal production. Notably, the speakers were also fluent in Cantonese, allowing for future cross-linguistic comparisons.

All recordings were made in a soundproof room using high-quality microphones. For the speech context, speakers used a normal speaking tone and speed, maintaining a natural conversational rhythm. Elocution was characterized by more focused and forceful vocalizations. For rap contexts, speakers listened to a rap accompaniment through headphones and synchronized their performance with the background music. Typical rap (Rap R) was recorded with a background music track selected for its rhythmic complexity and clear articulation, representative of Chinese Hardcore Rap. Melodic rap (Rap M), newly introduced in this experiment, featured both rhythmic and melodic elements, with melodies specified for each speaker to ensure consistency. The song context featured Chinese singing with a clear melody, while the cello context was purely instrumental, with musical notes matching the pitch height of each syllable in the rap context.

To ensure acoustic consistency, the pitch range of each speaker was measured prior to recording, and the average pitch for each context and target was calculated and aligned across conditions. The fundamental frequency (F0) of both context and target speech was analyzed using Praat software, and targets were reselected if discrepancies exceeded 10 Hz. All vocal and instrumental contexts were normalized to an intensity of 55 dB and a duration of 1800 ms. The target syllable /i/ was selected from natural recordings of the same four speakers, with Tone 1 (high-level) and Tone 2 (mid-rising) tokens chosen based on pitch trajectory and naturalness. An 11-step continuum was created between Tone 1 and Tone 2 using the STRAIGHT morphing algorithm, with steps 1, 5, 6, 7, and 11 used in the behavioral analysis to capture the full range of perceptual ambiguity.

To further distinguish typical rap from natural speech and elocution, we conducted acoustic analyses of our stimuli using Praat. The results showed that typical rap exhibited more frequent pauses (mean pause number = 2.25) and shorter average pause durations (120.67 ms) compared to speech (pause number = 1.5, pause duration = 101.71 ms) and elocution (pause number = 1.75, pause duration = 162.79 ms). These findings indicate that rap’ s rhythmic structure is more pronounced and artistically driven, whereas pauses in speech and elocution are primarily determined by semantic and syntactic boundaries. Additionally, the word pitch range in rap (17.34 Hz) was lower than in speech (21.10 Hz) and elocution (40.02 Hz), reflecting differences in prosodic variation.

Each context was further manipulated to cre-

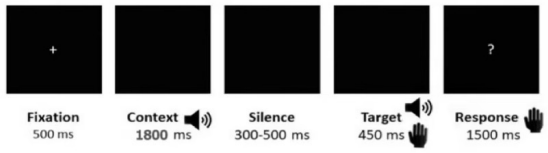


Figure 1: The trial procedure of the Mandarin word identification task.

ate two F0 conditions: F0-lowered and F0-raised, achieved by shifting the entire F0 trajectory by three semitones up or down. This manipulation allowed for the examination of contrastive context effects on tone perception. Fillers were constructed using the same procedure, with the context "接下来我会说出" ("Next I will say") and target fillers "衣" (/i55/, "clothes") and "疑" (/i35/, "misbelieve").

2.3 Experiment Procedure

Participants completed a practice block followed by six experimental blocks, each corresponding to one context condition. The order of blocks was counterbalanced across participants to control for order effects. In each trial, participants listened to a context stimulus, followed by a target syllable, and were instructed to identify the tone by pressing designated keys ("left" for Tone 1, "down" for Tone 2) with their right hand. A forward mask symbol (+) was displayed for 500 ms, followed by the context stimulus. After a brief silence (300–500 ms), the target syllable was presented. A question mark appeared after the target, prompting participants to respond within 1500 ms. Reaction times were not analyzed, as the focus was on tone identification accuracy. Each context condition included two F0 shifts (high, low), two types of content from four talkers, and five steps for the target (with the middle three steps repeated twice as often as the endpoints), plus fillers. This resulted in 72 trials per context condition. The six experimental blocks were counterbalanced to prevent order effects.

2.4 Analysis

In line with previous studies (Chen, et al., 2016; Tao, et al., 2021; Zhang et al., 2023), we analyzed the Tone 2 identification rate to assess lexical tone normalization. Our data were analyzed using Probit analysis to estimate categorical boundaries and effect sizes across contexts (Finney, 1971). The identification rate of Tone 2 was calculated for each context, F0 shift, and target step. Repeated measures ANOVAs were conducted to assess the influence of context type and F0 shift, with Greenhouse-Geisser correction applied where necessary. Post hoc comparisons were performed using Bonferroni correction.

Table 1: Derived categorical boundary positions for each type of context with high and low mean F0.

Context	High F0	Low F0	Difference
Cello	2.653	2.747	0.095
Rap M	2.942	3.050	0.107
Song	2.850	3.036	0.186
Rap R	2.618	3.209	0.591
Elocution	2.654	3.301	0.647
Speech	2.628	3.400	0.771

3 Results

The data revealed a robust hierarchy in the effectiveness of different contexts for eliciting lexical tone normalization. The speech context produced the most pronounced normalization effect, with effect sizes exceeding 20% at steps 2, 3, and 4, and peaking at nearly 30% at step 4. Elocution demonstrated a comparable, though slightly smaller, effect, with effect sizes approximately 5% lower than speech at each step and an overall mean of around 20%. Typical rap (Rap R) also induced robust normalization, particularly at step 4, where its effect size (24.11%) surpassed that of elocution (22.88%).

In contrast, melodic rap (Rap M) and song contexts produced only marginal normalization effects, with effect sizes below 10% at the middle three steps and negative values at certain steps (e.g., -0.45% for Rap M at step 4 and -2.57% for song at step 5), indicating little or no facilitation. The cello context failed to induce any significant effect, with effect sizes consistently below 5% and a mean of approximately 2.05%, suggesting minimal impact.

A 6 (context type: cello, elocution, melodic rap, rap, song, speech) Œ 2 (shift of context frequency: high, low) repeated-measures ANOVA on the categorical boundary (Greenhouse–Geisser corrected) showed no main effect of context, $F(3.26,48.86) = 1.92$, $p = .134$, $\eta_p^2 = .11$. Frequency shift, however, strongly shifted the boundary, $F(1,15) = 19.16$, $p < .001$, $\eta_p^2 = 0.56$, and this

effect was qualified by a significant interaction, $F(3.21, 48.09) = 7.83$, $p < .001$, $\eta_p^2 = .34$. Post-hoc contrasts revealed that high-frequency contexts advanced the boundary most for rap, melodic rap and song (all ps < .01), modestly for cello (p = .042), and not for speech or elocution (ps > .10).

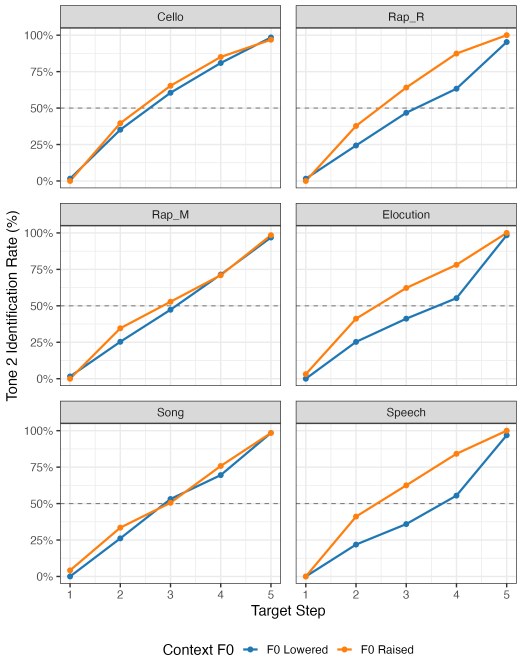


Figure 2: Average Tone 2 response by step and shift for each context (cello, melodic rap, song, rap, elocution, and speech).

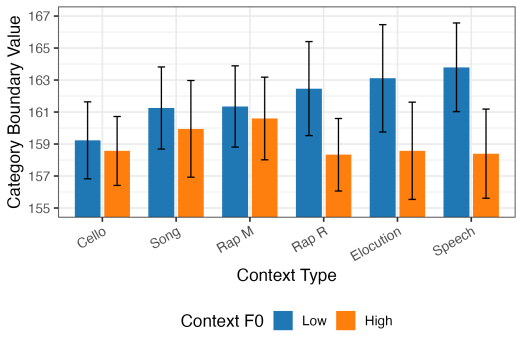


Figure 3: Category boundary values of each context type and frequency. Blue bars represent higher mean F0 contexts, whereas orange bars represent lower mean F0 contexts. Higher category boundary values indicate more Tone 2 responses.

Post hoc analyses using the LSD method were conducted to further examine differences between context types in their effects on the categorical boundary. The results indicated a significant difference between the cello context and melodic rap (Rap M) (p = .020, mean difference = -0.4054, 95% CI: [-0.7472, -0.0635]). Comparisons between the cello context and speech (p = .072), as well as the cello context and elocution (p = .088), approached significance. No other pairwise comparisons reached significance. These findings highlight the pronounced normalization effects in speech and elocution contexts, and the limited impact of melodic and instrumental contexts.

In summary, the speech context exhibited the most significant lexical tone normalization effect, with effect sizes exceeding 20% at key steps, underscoring its strong influence. Elocution, while slightly less impactful, still demonstrated substantial effects, highlighting the importance of clear linguistic cues. A typical rap context could also induce lexical tone normalization in Mandarin, particularly when pitch and rhythm align closely with speech, as seen with Rap R's effectiveness at certain steps. Conversely, Rap M and song contexts had minimal effects, likely due to melodic interference, while the cello context showed negligible impact. These results emphasize the inhibitory role of melody in tone normalization.

4 Discussion

This study provides clear evidence that speech-like cues are essential for lexical tone normalization in Mandarin. Both natural speech and elocution contexts robustly elicited normalization, as did typical rap when its rhythmic and pitch characteristics closely resembled those of speech. In contrast, contexts with greater melodic complexity, such as melodic rap and song, failed to induce significant normalization, highlighting the inhibitory role of melody.

Our acoustic analyses clarify the distinction between melody and lexical tone. By examining the mean pitch (melody) and pitch range (lexical tone variation) of each syllable across contexts, we found that rap and speech-maintained pitch contours consistent with natural Mandarin prosody. In these contexts, the pitch range within individual syllables was relatively large, indicating preserved tonal variation. However, in song and melodic rap, the pitch contours followed the imposed melody rather than natural tonal patterns, resulting in reduced pitch variation within syllables. This suggests that melody can override lexical tone cues, flattening the pitch contour and diminishing the information available for tone normalization.

To illustrate these findings, Figure 4 and Figure

5 present the mean pitch trajectories (melody) and word pitch range (lexical tone variation) for each syllable across contexts. As shown in Figure 4, the mean pitch trajectories for song and melodic rap are more stable and follow distinct melodic patterns, while speech, elocution, and typical rap display more natural pitch fluctuations that reflect Mandarin prosody. Figure 5 further demonstrates that the word pitch range is substantially reduced in song and melodic rap, indicating a flattening of tonal variation. In contrast, speech and elocution contexts maintain a much wider pitch range, and typical rap falls in between, preserving some degree of tonal variation.

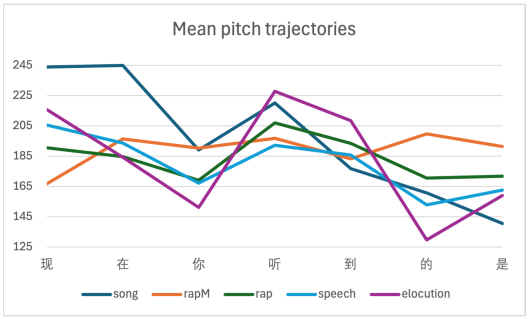


Figure 4: Mean pitch trajectories (melody) for each syllable across contexts.

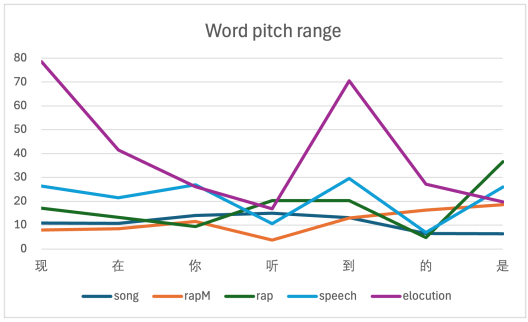


Figure 5: Word pitch range (lexical tone variation) for each syllable across contexts.

A closer look at the data reveals that, for each syllable, the pitch range in song and melodic rap remains consistently low (e.g., for “现”, song: 10.83 Hz, rapM: 7.90 Hz), while speech and elocution show much higher variability (speech: 26.35 Hz, elocution: 78.56 Hz). This pattern is consistent across all syllables, supporting the conclusion that melody in song and melodic rap overrides lexical tone cues, thereby diminishing the information available for tone normalization.

Together, these visualizations and data clearly show that as musicality increases, especially with the introduction of a strong melodic structure, both the mean pitch trajectory and the pitch range become less reflective of natural tonal variation. This supports the conclusion that melody can inhibit lexical tone normalization by reducing the availability of tonal cues in the auditory context.

Table 2: Word pitch range (Hz) for each syllable across contexts.

Syllable	Song	Rap M	Rap	Speech	Elocution
现	10.83	7.90	17.04	26.35	78.56
在	10.69	8.50	13.21	21.43	41.46
你	14.00	11.52	9.45	26.93	26.06
听	15.04	3.70	20.24	10.60	16.81
到	13.17	13.04	20.21	29.51	70.48
的	6.51	16.33	4.73	6.91	27.11
是	6.35	18.54	36.53	25.95	19.68

These results refine our understanding of the mechanisms underlying tone normalization. While previous research has emphasized the speech-specific nature of this process (Peng et al., 2012; Tao et al., 2021), our findings suggest that the boundary between speech and music is not absolute. Instead, there appears to be a threshold of musicality—particularly the presence of a strong melodic structure—beyond which the cognitive system can no longer effectively normalize lexical tones. This threshold is most evident in the transition from typical rap to melodic rap and song.

To further clarify the effects of different context types, we conducted post hoc comparisons using the LSD method on the categorical boundary values. Although the interaction between context type and context frequency was not statistically significant, this does not undermine the main objective of our study, which was to assess the relative efficacy of each context type. The lack of interaction may be attributable to the limited sample size, which could have reduced statistical power and obscured potential effects. Some conditions exhibited clear normalization effects while others did not, suggesting that with a larger sample, more pronounced differences and interactions might emerge. Future research should therefore consider increasing the number of participants to enhance statistical power and provide a more definitive assessment of interaction effects.

Our results also contribute to ongoing debates regarding domain-general versus language-specific mechanisms in pitch processing. The lack

of normalization in purely musical contexts supports the language-specific view, while the effectiveness of speech-like rap suggests that certain musical forms can engage speech processing mechanisms under specific conditions. This aligns with the frame of reference theory, which posits that listeners use contextual cues to interpret speech sounds.

Several factors may account for the inhibitory effect of melody observed in melodic rap and song. First, a dominant melodic structure may override or interfere with the listener' s ability to use pitch information for tone normalization. Second, processing complex musical cues may increase cognitive load, reducing resources available for linguistic processing. Third, the alteration of tonal pitch information in melodic contexts may disrupt the mapping between context and target, further reducing normalization effectiveness.

Our study also highlights the importance of retaining complete tonal pitch information in the context for inducing lexical tone normalization. In song and melodic rap, although vowels and consonants are preserved, tonal cues are weakened or flattened to fit the melody, resulting in diminished normalization effects. Therefore, tonal information appears to be the most critical factor for successful lexical tone normalization.

This study has several limitations. The sample size was relatively small, and the stimuli were limited to a specific set of contexts and speakers. Future research should explore a broader range of musical genres and linguistic backgrounds and employ more detailed EEG analyses to further elucidate the neural mechanisms involved. Additionally, cross-linguistic comparisons with other tonal languages would provide valuable insights into the generalizability of these findings.

In conclusion, our results support the hypothesis that while the brain's language processing systems can accommodate some degree of musical variability, excessive musicality—particularly strong melodic structure—can disrupt the normalization of lexical tones. The threshold beyond which musical speech variability hinders language comprehension appears to lie between typical rap and melodic rap. Further research is needed to define this threshold and explore the neural mechanisms underlying this interaction.

5 Conclusion

This study demonstrates that lexical tone normalization in Mandarin is robustly supported by speech and speech-like contexts, including typical rap with strong rhythmic and articulatory features. However, as musicality increases, particularly in melodic rap and song, the normalization effect diminishes, highlighting the inhibitory role of melody. Our findings suggest that there is a threshold of musicality beyond which the cognitive system can no longer effectively normalize lexical tones.

These results refine current models of speech perception by emphasizing both the necessity of speech-specific cues and the boundary conditions imposed by musical structure. Future research should further investigate the neural mechanisms underlying these effects, explore the precise threshold between speech and music, and consider broader implications for language learning and rehabilitation.

Acknowledgments

This study has been supported by an internal grant from The Hong Kong Polytechnic University (Project No. P0051041).

References

[1] Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, 17, 103–109.

[2] Chen, F., & Peng, G. (2016). Context effect in the categorical perception of Mandarin tones. *Journal of Signal Processing Systems*, 82(2), 253–261. doi:10.1007/s11265-015-1008-2.

[3] Dechovitz, D. (1977). Information conveyed by vowels: A confirmation. *Haskins Laboratory Status Report on Speech Research*, SR-53/54, 213–219.

[4] Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11(2), 149–175.

[5] Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America*, 125(6), 3983–3994.

[6] Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88, 642–654.

[7] Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), 98–104.

[8] Leather, J. (1983). Speaker normalization in the perception of lexical tone. *Journal of Phonetics*, 11, 373–382.

[9] Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Indiana University Linguistics Club, Bloomington, IN.

[10] Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85, 2088–2113.

[11] Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, 38(4), 616–624.

[12] Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & Wang, W. S.-Y. (2012). The effect of intertalker variations on acoustic–perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2), 579–595. doi:10.1044/1092-4388(2011/11-0025).

[13] Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 40–61.

[14] Tao, R., & Peng, G. (2020). Music and speech are distinct in lexical tone normalization processing. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

[15] Tao, R., Zhang, K., & Peng, G. (2021). Music does not facilitate lexical tone normalization: A speech-specific perceptual process. *Frontiers in Psychology*, 12, 717110. doi:10.3389/fpsyg.2021.717110.

[16] Tian, Y., Ye, Y., Lu, M., Jia, F., & Tao, R. (2024). Effect of rap music context on lexical tone normalization. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 1279–1286.

[17] Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2), 413–421. doi:10.1044/1092-4388(2003/033).

[18] Ye, Y., & Peng, G. (2024). Mental representation of Mandarin Tone 3: An integrated phonetic and phonological reflection. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 1295–1300.

[19] Zhang, C., Peng, G., & Wang, W. S.-Y. (2012). Unequal effects of speech and non-speech contexts on the perceptual normalization of Cantonese level tones. *Journal of the Acoustical Society of America*, 132(2), 1088–1099. doi:10.1121/1.4731470.

[20] Zhang, K., Sjerps, M. J., & Peng, G. (2021). Integral perception, but separate processing: The perceptual normalization of lexical tones and vowels. *Neuropsychologia*, 156, 107839. doi:10.1016/j.neuropsychologia.2021.107839

[21] Zhang, K., Tao, R., & Peng, G. (2023). The advantage of the music-enabled brain in accommodating lexical tone variabilities. *Brain and Language*, 247, 105348. doi:10.1016/j.bandl.2023.105348.

RMCP: Enhancing LLM-based Translation via Prompting with Retrieved Monolingual Corpora

Wenhui Shi and Ziman Han
Institute of Language Sciences
Shanghai International Studies University
Shanghai, China
shiwenhui712@163.com, hanziman@shisu.edu.cn

Abstract

Previous work has proved that incorporating external linguistic resources into translation models can effectively improve their adaptability to diverse translation scenarios. However, most existing studies rely on limited bilingual translation memories and require additional model training, significantly restricting their application. This study attempts to address these issues simultaneously with the Retrieved Monolingual Corpora Prompting (RMCP) framework. This framework leverages a pre-trained multilingual sentence embedding model to retrieve the top-*k* semantically similar sentences of the source text from a monolingual corpus and incorporates them into the prompt of large language models (LLMs) as in-context examples. Experiments demonstrate that applying the framework to various LLMs significantly improves their translation performance across different language pairs, including those involving low-resource languages. It even outperforms the powerful commercial machine translation system Google Translate. Notably, the inference model shows great potential in utilizing monolingual examples. Qualitative analysis reveals that RMCP improves the quality of LLMs' translations by providing lexical, syntactic, pragmatic, and formatting guidelines.

1 Introduction

Translation Memory (TM) is a computer-based tool that enables translators to consult a database of previous translations, retrieving similar sentence pairs to guide and assist in translating new content (Somers, 2003). The integration of TM into machine translation (MT) systems has long been pursued as a way to combine the accuracy of human-translated segments with the scalability provided by automated models (Bouthors et al., 2024; Hao et al., 2023; Zhang et al., 2018; Wang et al., 2014). These efforts have resulted in significant improvements in translation quality, consistency, and efficiency in MT applications.

However, previous TM-augmented translation approaches rely heavily on bilingual parallel data, which are often limited in both coverage and scale in practical scenarios (Wolk and Marasek, 2015). This reliance on bilingual resources significantly impedes advancements in MT for many low-resourced languages and specialized domains. Although there have been attempts to incorporate monolingual corpora into translation systems, most existing methods depend on either back-translation pipelines (Sennrich et al., 2016) or architectural adaptations (Cai et al., 2021). These methods often result in the loss of translation expertise, higher computational expenses, and difficulties in achieving real-time updates.

The rise of LLMs presents a promising solution to these challenges. A key advantage of LLM-based translation over statistical machine translation (SMT) and neural machine translation (NMT) is their prompt-based interface, which offers extended context length and flexible formatting (Khair and Sawalha, 2025). This architecture minimizes text fragmentation from chunking, thereby enhancing information continuity across longer contexts. Additionally, the simple prompt-based customization enables users to readily define the role of monolingual corpora in the translation task. These intrinsic features create numerous possibilities for incorporating monolingual corpora into the translation process.

To this end, we propose RMCP, a plug-and-play framework that enhances translation quality through the direct utilization of monolingual corpora. Specifically, our framework uses Language-Agnostic BERT Sentence Embedding (LaBSE) to retrieve the top-*k* sentences from an external monolingual corpus. These sentences are then seamlessly integrated into the prompt to steer the LLM's translation via in-context learning (ICL). This entire process eliminates the need for parallel data and model retraining, offering a lightweight yet

powerful solution for translation augmentation.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to systematically demonstrate that monolingual corpora can be directly harnessed as a source of translation knowledge through a straightforward prompting approach. This finding paves new pathways for low-resource language translation.
- We propose a practical framework that seamlessly integrates pre-trained sentence retrievers with readily available LLMs, thereby significantly reducing the barriers to adoption.
- We perform extensive empirical evaluations across a variety of language pairs, LLM architectures, and experimental settings. The results not only demonstrate the superiority of our method compared to vanilla LLM translation and Google Translate baselines, but also unveil critical insights, including the great potential of inference models to utilize retrieved monolingual examples for translation.

2 Related Work

Extensive research has highlighted the importance of high-quality retrieved sentence pairs, commonly referred to as "fuzzy matches" or "translation memories," in enhancing machine translation performance.

Related research showed remarkable diversity and innovation within the SMT paradigm. Koehn and Senellart (2010) employed XML markup to enable SMT systems to concentrate on non-matching segments, effectively merging the precise matches offered by TM with the generalization capabilities of SMT. Ma et al. (2011) implemented discriminative learning techniques to promote translation consistency, resulting in notable improvements in BLEU scores for English-Chinese technical documents. Additionally, Wang et al. (2014) introduced dynamic merging of TM phrase pairs and an enhanced integration model, addressing the discrepancies between TM databases and SMT training datasets. These groundbreaking studies have revealed the untapped potential of external linguistic resources in MT systems.

Subsequent studies have also made significant strides in integrating translation memory into traditional encoder-decoder NMT models. A range of lightweight methods (Zhao et al., 2018; Zhang et al., 2018; Bulte and Tezcan, 2019; Xu et al., 2020; Reheman et al., 2023) have been proposed successfully. Meanwhile, some researchers have broadened the resource pool from bilingual translation memories to include monolingual corpora (Reheman et al., 2024). However, the lack of predefined roles for monolingual data in the input framework of traditional NMT often requires careful design for effective incorporation. One common strategy is back-translation (Sennrich et al., 2016; Fadaee et al., 2017; Edunov et al., 2018), where monolingual target-language sentences are translated to the source language to create synthetic parallel data. Alternatively, performance improvements can also be achieved through model adaptation approaches, such as architectural modifications (Cai et al., 2021) or additional training (Cai et al., 2021; Tamura et al., 2023).

In recent years, the emergence of LLMs has significantly transformed the landscape of machine translation. Researchers have begun to explore the integration of retrieved translation segments with LLMs, a methodology referred to as the Retrieval-Augmented Translation (RAT) paradigm (Hoang et al., 2022). Some existing works have struck a balance between efficiency and quality through innovative designs (Shi et al., 2022; Mu et al., 2023; Wang et al., 2024; Zhu et al., 2024). However, despite these notable advancements, the availability of bilingual corpora necessary for these studies is still considerably more limited compared to the abundance of monolingual data (Sennrich et al., 2016). Therefore, this paper explores a more accessible approach: directly utilizing monolingual corpora through prompting to enhance LLM-based translation.

3 The RMCP pipeline

Inspired by the In-Context Retrieval-Augmented Language Models (RALM) proposed by Ram et al. (2023), this study employs a black-box RAT pipeline that requires no fine-tuning. The core feature of this pipeline is the direct integration of a pre-trained sentence retriever with an off-the-shelf LLM serving as the translation engine. The entire process consists of three main stages: Sentence Retrieval, Prompt Construction, and Translation Generation. Figure 1 presents an example to illustrate the workflow.

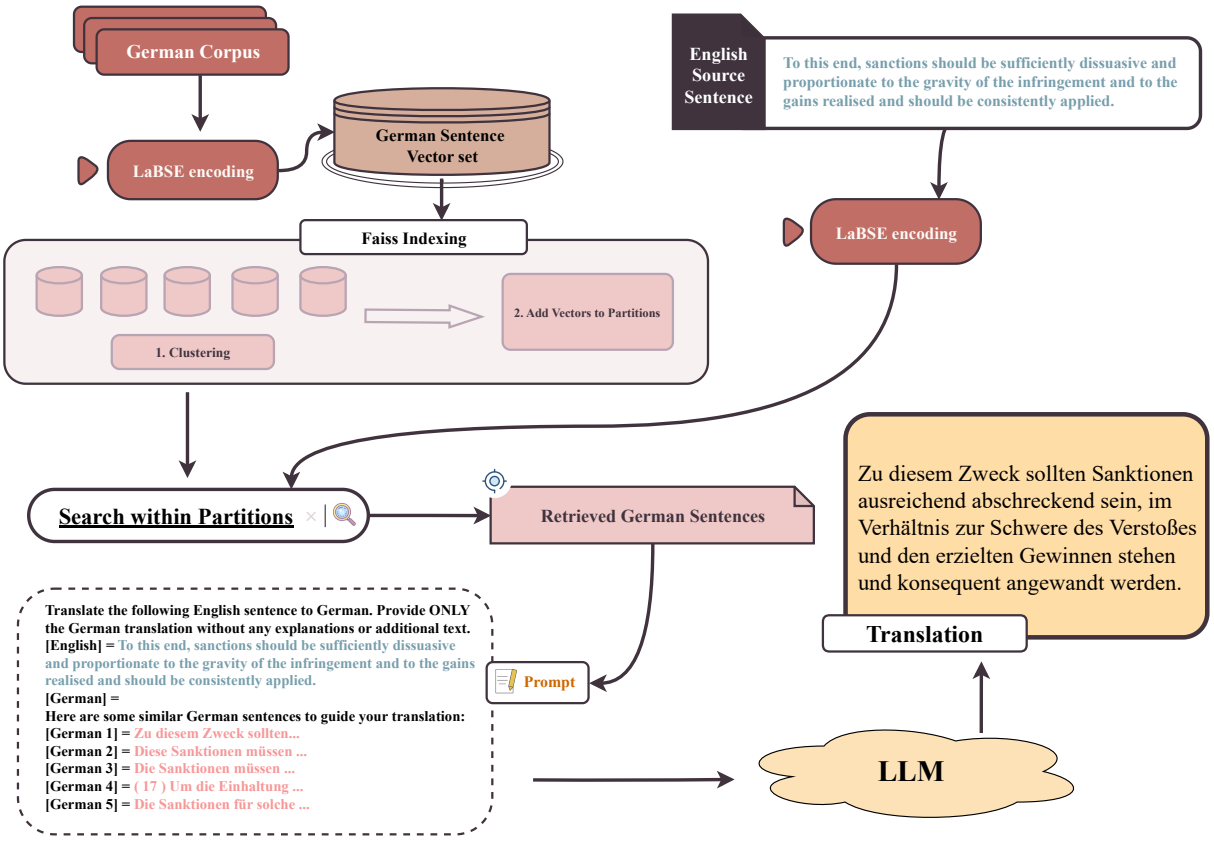


Figure 1: An example of RMCP. Note that the clustering and partitioning steps are specific to the IndexIVFFlat index.

3.1 Sentence Retrieval

In the retrieval stage, we encode both the input source sentences and retrievable monolingual corpora into a shared semantic space using LaBSE (Feng et al., 2022), which is renowned for its strong performance on cross-lingual similarity tasks. Subsequently, we leverage Faiss (Douze et al., 2024)[1] to manage and query these embeddings efficiently. To enable fast retrieval of the top-k most semantically similar sentences, we construct an appropriate search index from the sentence embeddings based on the corpus size. Similarity is measured by the maximum inner product search (MIPS) within Faiss.

3.2 Prompt Construction

Once the top-k sentences are retrieved, they are formatted as in-context examples and integrated into a prompt for the LLM. As shown in the example in Figure 1, a typical prompt includes a clear task instruction, the retrieved sentences presented as guiding examples, and the source sentence to be translated. This structure leverages the ICL capabilities of LLMs, providing them with powerful contextual cues to enhance translation quality.

3.3 Translation Generation

In the final stage, the constructed prompt is fed to a large language model. The model processes the full context and generates the final translation for the source sentence.

4 Experimental Setup

4.1 Datasets and Preprocessing

We tested our method on two datasets: 1) the JRC-Acquis (JRC-A) dataset (Steinberger et al., 2006), which constitutes a multilingual parallel corpus covering 24 European languages with aligned documents in legal and administrative domains[2] and 2) Bible Jv↔Id (Cahyawijaya et al., 2021), a Bible corpus for Javanese-Indonesian (Jv↔Id) translation.[3] For the JRC-A dataset, we focused on four language pairs, conducting experiments in both translation directions: German↔English (de↔en),

[1] https://github.com/facebookresearch/faiss

[2] https://wt-public.emm4u.eu/Acquis/JRC-Acquis.3.0/corpus/

[3] https://github.com/IndoNLP/indonlg

English↔Spanish (en↔es), English↔French (en↔fr), and German↔French (de↔fr). We adopted a data splitting strategy consistent with Reheman et al. (2023). Specifically, for each language pair, we randomly selected 3,000 sentence pairs to constitute the test set, while the remaining sentences were used as the retrievable corpus for translation augmentation.

For the Bible Jv↔Id corpus, we used the provided test set for evaluation, while the combined training and validation sets served as the retrievable corpus.

It is important to note that the retrievable corpora from both datasets initially consisted of bilingual sentence pairs. Since our study focuses on the utility of monolingual corpora, we disaggregated these bilingual pairs into source-side and target-side monolingual corpora, which were then used independently for retrieval in our experiments.

Detailed statistics for these two datasets are presented in Table 1.

Dataset	Lang.	Testset	Mono. Corpora
JRC-A	De↔En	3,000	423,315
	En↔Es	3,000	432,858
	En↔Fr	3,000	424,300
	De↔Fr	3,000	846,502
Bible Jv↔Id	Jv↔Id	1,193	6,765

Table 1: Statistics of the JRC-A and Bible Jv↔Id datasets.

4.2 Models and Baselines

We experiment with three leading LLMs: two generative models, DeepSeek-V3 and GPT-4.1, and one reasoning model, DeepSeek-R1. These models can represent the current performance frontier in their respective model categories. Their performance under our RMCP framework is compared against the zero-shot setting to quantify the improvements.

Furthermore, we benchmark our results against Google Translate[4]. As a user-friendly and powerful commercial system, it serves as an ideal baseline to demonstrate the practical viability and competitiveness of our approach.

[4] https://translate.google.com/

4.3 Evaluation Metrics

We evaluate translation quality using established automatic metrics.

Our primary metrics are BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017), both of which measure n-gram overlap with reference translations and are implemented via sacreBLEU (Post, 2018)[5] to ensure reproducibility. Specifically, for the JRC-A dataset, we report BLEU scores, employing the default "13a" tokenizer. For the Bible JvId dataset, chrF++ is used instead of BLEU due to the morphological richness of Javanese and Indonesian.

We also report COMET[6] scores (wmt22-COMET-da) (Rei et al., 2022) as a complementary metric providing deeper semantic insights.

5 Results

5.1 The Effectiveness of RMCP

To comprehensively evaluate the effectiveness of our proposed method, we compared the translation performance of LLMs with and without RMCP augmentation across eight translation directions. In this experiment, we utilized the P2.D prompt template (see Appendix A for details) and set the number of retrieved examples to 5.

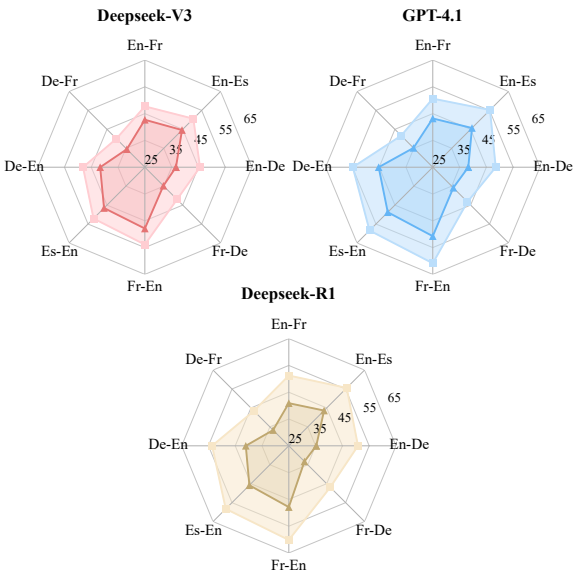


Figure 2: BLEU scores for different LLMs w/ and w/o RMCP augmentation on the JRC-A dataset.

Table 2 presents the detailed BLEU and COMET scores for all configurations. All reported improvements of our method over the zero-shot setting

[5] https://github.com/mjpost/sacrebleu
[6] https://github.com/Unbabel/COMET

BLEU									
Model	**Setting**	**En-De**	**De-En**	**En-Es**	**Es-En**	**En-Fr**	**Fr-En**	**De-Fr**	**Fr-De**
DeepSeek-V3	Zero-shot	36.61	41.68	44.62	46.55	42.71	47.89	34.51	34.79
	Few-shot (RMCP)	45.58***	48.08***	50.61***	52.09***	47.71***	53.88***	40.04***	42.00***
GPT-4.1	Zero-shot	38.26	45.20	45.70	48.76	43.20	50.74	35.18	35.85
	Few-shot (RMCP)	48.50***	**54.85*****	55.21***	**58.40*****	50.60***	**60.85*****	41.67***	43.31***
DeepSeek-R1	Zero-shot	35.25	41.06	43.62	45.82	40.91	47.88	33.30	33.26
	Few-shot (RMCP)	**50.82*****	53.64***	**55.57*****	58.29***	**51.24*****	60.15***	**43.43*****	**46.89*****
Google Translate	–	45.40	53.28	51.81	53.73	47.71	56.67	37.96	40.13
COMET									
Model	**Setting**	**En-De**	**De-En**	**En-Es**	**Es-En**	**En-Fr**	**Fr-En**	**De-Fr**	**Fr-De**
DeepSeek-V3	Zero-shot	83.49	82.77	85.46	84.31	86.00	84.87	82.78	82.62
	Few-shot (RMCP)	84.79***	83.74***	86.32***	85.09***	86.59***	85.50***	83.80***	83.60***
GPT-4.1	Zero-shot	84.04	83.12	85.73	84.55	86.16	85.15	83.02	82.81
	Few-shot (RMCP)	**85.61*****	84.90***	**87.22*****	**86.10*****	**87.32*****	**86.58*****	**84.52*****	**84.20*****
DeepSeek-R1	Zero-shot	83.23	82.41	85.16	84.09	85.61	84.64	82.48	82.30
	Few-shot (RMCP)	85.25***	84.41***	86.89***	85.67***	87.07***	86.21***	84.45***	84.15***
Google Translate	–	85.39	**85.21**	86.56	85.25	86.89	85.65	83.57	83.54

Table 2: BLEU and COMET scores for different LLMs w/ and w/o RMCP augmentation on the JRC-A dataset. Bold text denotes the highest score in each translation direction. Statistically significant improvements of Few-shot (RMCP) over its corresponding Zero-shot baseline are marked as follows: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

have been tested for statistical significance using bootstrap resampling with 1000 iterations. The results clearly demonstrate that RMCP consistently and substantially enhances the translation quality across all tested LLMs and language pairs. For instance, when augmented with retrieved monolingual examples, DeepSeek-V3's BLEU score for En-De translation improves from 36.61 to 45.58, and GPT-4.1's score for Fr-En translation increases from 50.74 to 60.85. This consistent positive impact of incorporating monolingual translation memories, which can also be intuitively observed from Figure 2, underscores the general applicability and efficacy of our approach. Crucially, this enhancement allows LLMs, particularly the advanced GPT-4.1, to achieve translation quality that is not only competitive with but often surpasses the Google Translate baseline.

A particularly insightful comparison arises when examining DeepSeek-R1 and DeepSeek-V3, which possess a similar parameter count. In the zero-shot setting, their translation performance is largely comparable. However, upon the application of RMCP, DeepSeek-R1 exhibits a markedly more substantial improvement in performance compared to DeepSeek-V3. In En-De translation, DeepSeek-R1's BLEU score increases by +15.57 points (from 35.25 to 50.82), whereas DeepSeek-V3's score improves by a smaller margin of +8.97 points (from 36.61 to 45.58). This pattern is consistent across other language pairs, showcasing the great potential of reasoning models like DeepSeek-R1 in leveraging monolingual examples for translation tasks. Nevertheless, the observation that DeepSeek-R1 did not consistently surpass the performance of the generative model GPT-4.1, coupled with its high computational costs, suggests that its practical utility warrants further evaluation.

To further validate the effectiveness of our method in low-resource scenarios, we conducted experiments on the Bible Jv↔Id dataset, with ChrF++ and COMET scores presented in Table 3. The results indicate that the retrieval and utilization of monolingual corpora remain effective for this morphologically richer low-resource language pair. However, the improvements brought by monolingual corpora in the Id-Jv direction were relatively limited. This suggests that while RMCP can enhance LLM translation quality, its effectiveness might still be constrained by the LLM's foundational capabilities in that specific direction.

5.2 Impact of the Language of the Monolingual Corpora

In this section, we compare the performance of GPT-4.1 when retrieving from target-side monolingual data against retrieving from source-side monolingual data. The results are detailed in Table 4.

The findings indicate that while utilizing source-side monolingual data occasionally provides marginal, sometimes statistically significant, gains over the zero-shot baseline, its impact is inconsistent and significantly less pronounced than that observed with target-side monolingual data. Mono-

Model	Setting	Id-Jv		Jv-Id	
		ChrF++	COMET	ChrF++	COMET
DeepSeek-V3	Zero-shot	35.29	85.27	58.26	87.30
	Few-shot (RMCP)	35.51*	85.07	60.83***	87.88***
GPT-4.1	Zero-shot	37.30	86.32	61.95	88.83
	Few-shot (RMCP)	38.66***	**86.62***	**67.26*****	**89.98*****
DeepSeek-R1	Zero-shot	35.07	85.48	56.96	86.79
	Few-shot (RMCP)	35.56*	85.65	58.83***	87.26***
Google Translate	–	**39.30**	84.29	65.89	88.36

Table 3: ChrF++ and COMET scores for different LLMs w/ and w/o RMCP augmentation on the Bible Jv↔Id dataset. Bold text denotes the highest score in each translation direction. Statistically significant improvements of Few-shot (RMCP) over its corresponding Zero-shot baseline are marked as follows: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

lingual data from the target language is a far more effective resource for retrieval-augmented translation within our prompting framework.

5.3 Impact of *k*

To determine the optimal number of retrieved in-context examples (k), we analyzed translation performance on the En-De direction by varying k from 0 to 10. Referring to Figure 3, it can be observed that the reasoning model DeepSeek-R1 demonstrates a remarkable ability to leverage a small number of examples. It reaches a decent BLEU score with only k=1 and approaches its peak performance rapidly with k=3 to k=5 examples. In contrast, while the generative models, DeepSeek-V3 and GPT-4.1, also benefit from increased k, their performance curve shows a more gradual ascent, typically requiring more examples to reach their respective optimal scores. This suggests that reasoning models such as DeepSeek-R1 could be particularly efficient in extracting and using information from limited in-context examples.

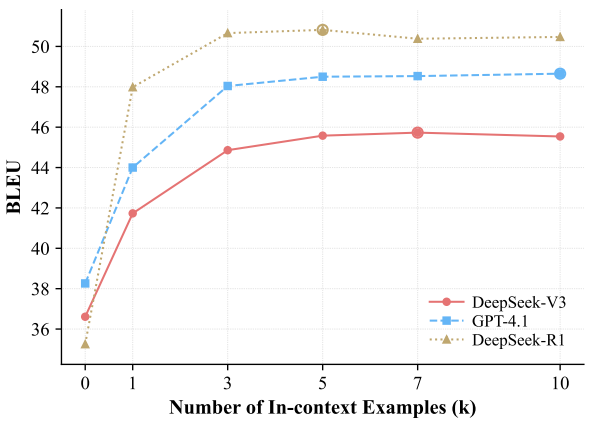


Figure 3: BLEU scores for different models with varying values of k.

5.4 Impact of Prompt Design

We evaluated two categories of prompt templates for GPT-4.1: Explicit Guidance (P1 series), which clearly defines the role of the monolingual examples, and Implicit Guidance (P2 series), which presents the examples with minimal instruction, relying on the LLM's in-context learning capabilities to infer their utility.

Results (Tables 5 and 6 in Appendix A) indicate that prompt structure indeed matters. While specific implicit prompts can achieve strong results in some cases, the Explicit Guidance prompts generally demonstrate more stable and consistently high performance.

5.5 Impact of the Size of the Monolingual Corpora

We investigated the effect of the retrievable monolingual corpus size on translation performance using GPT-4.1 for En-De translation. The size of the target-side monolingual data was varied from 0% to 100% of the available data, in 20% increments.

As shown in Figure 4, there is a clear positive correlation between the size of the monolingual corpus and translation quality. Both BLEU and COMET scores consistently improve as more monolingual data is made available for retrieval. This suggests that a larger monolingual corpus provides a richer source of information for the model, leading to better translation performance.

6 Analysis

To offer a more in-depth understanding of RMCP's effectiveness, this section delves into specific improvements across various linguistic and stylistic dimensions by analyzing translations generated

BLEU								
Setting	**En-De**	**De-En**	**En-Es**	**Es-En**	**En-Fr**	**Fr-En**	**De-Fr**	**Fr-De**
Zero-shot	38.26	45.20	45.70	48.76	43.20	50.74	35.18	35.85
RMCP.src	38.03	45.14	46.04*	49.20**	43.00	51.27***	35.05	35.92
RMCP.tgt	**48.50*****	**54.85*****	**55.21*****	**58.40*****	**50.60*****	**60.85*****	**41.67*****	**43.31*****
COMET								
Setting	**En-De**	**De-En**	**En-Es**	**Es-En**	**En-Fr**	**Fr-En**	**De-Fr**	**Fr-De**
Zero-shot	84.04	83.12	85.73	84.55	86.16	85.15	83.02	82.81
RMCP.src	84.13**	83.20*	85.86***	84.59	86.20	85.19	83.06	82.83
RMCP.tgt	**85.61*****	**84.90*****	**87.22*****	**86.10*****	**87.32*****	**86.58*****	**84.52*****	**84.20*****

Table 4: BLEU and COMET scores for GPT-4.1 on the JRC-A dataset under three settings: Zero-shot, RMCP.src (w/ source-side monolingual corpora), and RMCP.tgt (w/ target-side monolingual corpora). Bold text denotes the highest score in each translation direction. Statistically significant improvements of Few-shot (RMCP) over its corresponding Zero-shot baseline are marked as follows: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

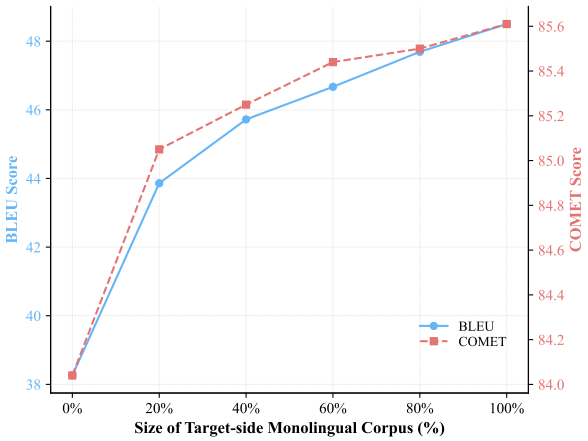


Figure 4: Translation Performance of GPT-4.1 across different monolingual corpus sizes

with and without monolingual examples. All illustrative cases, detailed in Appendix B, are from the Fr-En translation direction of the JRC-A dataset, with GPT-4.1 as the translation model.

6.1 Precise Lexical Choice and Terminology Application

Compared to the zero-shot approach, RMCP demonstrably enhances the LLM's ability to select more precise vocabulary and terminology appropriate for specific contexts, especially within legal and official texts, by leveraging retrieved examples from target-side monolingual corpora.

In Case 1, the phrase "instruments de défense commerciale" is rendered as the abbreviation "TDI" in the reference translation. The zero-shot method produces the full term "trade defense instruments." In contrast, RMCP, guided by retrieved Example 1, generates a translation that includes both the full term and its abbreviation, more closely aligning with the reference. Beyond identifying abbreviations, RMCP also aids in translating domain-specific terminology. In Case 2, RMCP successfully identifies the name of a regulation, employing the capitalized and pluralized term "Agreed Minutes" and the capitalized "Government." This is a significant improvement over the zero-shot output, which deviates from conventions.

6.2 Optimized Syntactic Structures and Expressive Fluency

By referencing monolingual examples from the target language, RMCP guides the model in generating translations with natural syntactic structures and smooth logical connections. This effect is particularly pronounced when translating sentences with specific stylistic features, such as the lengthy clauses and fixed phrasings common in legal texts.

In Case 3, the zero-shot translation appears redundant and awkward due to its repeated use of "of + gerund" structures. The RMCP translation, however, adopts the more conventional "Whereas" opening and features a more concise and fluent parallel structure. This improvement is attributable to retrieved examples demonstrating how to organize and connect multiple parallel actions in the target language. Furthermore, Case 4 illustrates that RMCP successfully learns and incorporates key legal phrasal fragments from retrieved examples, such as "shall check that ... are complied with." This indicates that the formality of LLM translations can be enhanced by leveraging examples that feature recurrent and domain-specific syntactic structures and connectives.

6.3 Appropriate Pragmatic Functioning

In certain situations, RMCP enables the LLM to adjust the translation's tone based on retrieved examples, producing outputs that better align with the requirements of specific text types.

For instance, in Case 4, the LLM successfully acquires and employs the modal verb "shall" based on the retrieved examples. In legal discourse, "shall" commonly indicates a binding obligation (Garner, 2014), thereby making the RMCP translation more formal and semantically precise than the zero-shot version's use of the simple present "consider." This choice also enhances consistency with the reference translation. Additionally, in Case 5, although the retrieved examples do not directly provide the opening "A procedure should be established," they contain multiple instances of suggestive expressions with "should be." Consequently, the LLM employs the modal verb "should," accurately conveying the implicit recommendation or obligation in the source sentence.

6.4 Strict Adherence to Textual Formatting and Typographical Norms

In addition to improvements in lexical, syntactic, and pragmatic aspects, RMCP-generated translations also demonstrate a greater adherence to standard formatting and typographical conventions, such as list numbering, capitalization of proper nouns, and the appropriate use of special symbols. This is mainly attributable to the high quality and well-formatted nature of the retrieved examples.

For instance, in Case 6, the LLM learns to capitalize "Regulation" and adopts the British spelling "authorised" based on the retrieved texts. Similarly, Case 2 highlights RMCP's effectiveness in conforming to these conventions.

In summary, the retrieved monolingual examples function as strong contextual cues that guide the LLM not only in producing more accurate and fluent translations but also in adhering to formatting and typographical norms, ultimately enhancing the overall presentation and professionalism of the output.

7 Conclusion

This paper introduced a highly practical approach that enhances LLM-based machine translation by prompting with retrieved monolingual corpora. Our findings indicate that external monolingual corpus resources can improve the translation performance of various LLMs across different language pairs, surpassing the robust and user-friendly commercial baseline Google Translate. The proposed method establishes a lightweight framework for translation improvement that requires neither parallel data nor model retraining, providing flexible solutions to meet diverse real-world translation needs.

Limitations

This study is not without its limitations. First, the lack of similarity-aware filtering mechanisms raises concerns about the potential introduction of noise from low-similarity monolingual examples. Second, the increased inference latency and computational costs stemming from retrieval and longer prompts were not evaluated in this work. Finally, while our findings emphasize the great potential of reasoning models (e.g., DeepSeek-R1) in utilizing monolingual examples, a comprehensive and multidimensional comparative analysis against generative models has yet to be conducted.

References

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7307–7318, Online. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

B.A. Garner. 2014. *Black's Law Dictionary*. BLACK'S LAW DICTIONARY. Thomson Reuters.

Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. 2023. Rethinking translation memory augmented neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2589–2605, Toronto, Canada. Association for Computational Linguistics.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2022. Improving robustness of retrieval augmented translation via shuffling of suggestions. *Preprint*, arXiv:2210.05059.

Mohammad Mohammad Khair and Majdi Sawalha. 2025. Automated translation of islamic literature using large language models: Al-shamela library application. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 53–58, Abu Dhabi, UAE. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning - a translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1239–1248, Portland, Oregon, USA. Association for Computational Linguistics.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Abudurexiti Reheman, Yingfeng Luo, Junhao Ruan, Chunliang Zhang, Anxiang Ma, Tong Xiao, and JingBo Zhu. 2024. Exploiting target language data for neural machine translation beyond back translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12216–12228, Bangkok, Thailand. Association for Computational Linguistics.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13519–13527.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models

with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022. XRICL: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5248–5259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

H. Somers. 2003. *Computers and Translation: A Translator's Guide*. Benjamins Translation Library. John Benjamins Publishing Company.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Takuya Tamura, Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. 2023. Target language monolingual translation memory based NMT by cross-lingual retrieval of similar translations and reranking. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 313–323, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Kun Wang, Chengqing Zong, and Keh-Yih Su. 2014. Dynamically integrating cross-domain translation memory into phrase-based machine translation during decoding. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 398–408, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-RAG: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978, Bangkok, Thailand. Association for Computational Linguistics.

Krzysztof Wolk and Krzysztof Marasek. 2015. Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 118–125, Da Nang, Vietnam.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Yang Zhao, Yining Wang, Jiajun Zhang, and Chengqing Zong. 2018. Phrase table as recommendation memory for neural machine translation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4609–4615. International Joint Conferences on Artificial Intelligence Organization.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

A Prompt Templates and Evaluation Results

In order to explore the impact of prompt design on the performance of GPT-4.1, we use 8 prompt templates in the De→En direction of the JRC-A dataset for experiments. The results are shown in Table 5. and Table 6.

B Qualitative Analysis Examples

This appendix provides detailed examples supporting the qualitative analysis presented in Section 6. For each case, we present the source sentence (French), the human reference translation (English), the output from the Zero-shot LLM, the output from our RMCP method, and the top 5 retrieved English monolingual sentences that guided the RMCP translation.

Case 1

Source (fr): (441) Ces actions sont des exemples de mesures faisant appel aux instruments de défense commerciale , imposées par les États-Unis sur les importations du produit concerné , et viennent s' ajouter aux mesures de sauvegarde mentionnées ci-dessus.

Reference (en): (441) These measures are examples of the TDI measures imposed by the US on imports of the product concerned and are in addition to the safeguard measures mentioned above.

No.	Prompt Template	BLEU	COMET
P1.A	Translate the following English text to German. Provide ONLY the German translation without any explanations or additional text. Please use the style, phrasing, and fluency demonstrated in the German examples below as a reference. Example 1: Zu diesem Zweck sollten die Maßnahmen ... Example 2: Diese Sanktionen müssen wirksam , verhältnismäßig ... Example 3: Die Sanktionen müssen in angemessenem Verhältnis ... Example 4: (17) Um die Einhaltung dieser ... Example 5: Die Sanktionen für solche Zuwiderhandlungen ... Text to Translate: To this end , sanctions should be sufficiently... German Translation:	51.26	**85.89**
P1.B	Translate the following English text to German. Provide ONLY the German translation without any explanations or additional text. The following are German sentences from the same domain. Pay attention to the domain-specific terminology and common phrasings when generating your translation. Example 1: Zu diesem Zweck sollten die Maßnahmen ... Example 2: Diese Sanktionen müssen wirksam , verhältnismäßig ... Example 3: Die Sanktionen müssen in angemessenem Verhältnis ... Example 4: (17) Um die Einhaltung dieser ... Example 5: Die Sanktionen für solche Zuwiderhandlungen ... Text to Translate: To this end , sanctions should be sufficiently... German Translation:	50.35	85.79
P1.C	Translate the following English text to German. Provide ONLY the German translation without any explanations or additional text. Consider the following German statements as relevant background information: Zu diesem Zweck sollten die Maßnahmen ... Diese Sanktionen müssen wirksam , verhältnismäßig ... Die Sanktionen müssen in angemessenem Verhältnis ... (17) Um die Einhaltung dieser ... Die Sanktionen für solche Zuwiderhandlungen ... Text to Translate: To this end , sanctions should be sufficiently... German Translation:	50.93	85.61
P1.D	Translate the following English text to German. Provide ONLY the German translation without any explanations or additional text. To help you, here are some high-quality sentences in German that reflect the desired output quality and style. Example 1: Zu diesem Zweck sollten die Maßnahmen ... Example 2: Diese Sanktionen müssen wirksam , verhältnismäßig ... Example 3: Die Sanktionen müssen in angemessenem Verhältnis ... Example 4: (17) Um die Einhaltung dieser ... Example 5: Die Sanktionen für solche Zuwiderhandlungen ... Text to Translate: To this end , sanctions should be sufficiently... German Translation:	50.32	85.82

Table 5: Translation Performance of GPT-4.1 using Explicit Guidance prompts (P1 series)

No.	Prompt Template	BLEU	COMET
	Zu diesem Zweck sollten die Maßnahmen ... Diese Sanktionen müssen wirksam , verhältnismäßig ... Die Sanktionen müssen in angemessenem Verhältnis ... (17) Um die Einhaltung dieser ... Die Sanktionen für solche Zuwiderhandlungen ...	49.00	85.57
P2.A	Translate the following English text to German. Provide ONLY the German translation without any explanations or additional text. Text to Translate: To this end , sanctions should be sufficiently... German Translation:		
	Translate from English to German. Provide ONLY the German translation without any explanations or additional text. German Examples: Zu diesem Zweck sollten die Maßnahmen ... Diese Sanktionen müssen wirksam , verhältnismäßig ...	**51.74**	85.87
P2.B	Die Sanktionen müssen in angemessenem Verhältnis ... (17) Um die Einhaltung dieser ... Die Sanktionen für solche Zuwiderhandlungen ... Source (English): To this end , sanctions should be sufficiently... Target (German):		
	German: Zu diesem Zweck sollten die Maßnahmen ... German: Diese Sanktionen müssen wirksam , verhältnismäßig ... German: Die Sanktionen müssen in angemessenem Verhältnis ... German: (17) Um die Einhaltung dieser ... German: Die Sanktionen für solche Zuwiderhandlungen ...	50.91	85.62
P2.C	Provide ONLY the German translation of the following English text. To this end , sanctions should be sufficiently...		
	Translate the following English sentence to German. Provide ONLY the German translation without any explanations or additional text. [English] = To this end , sanctions should be sufficiently... [German] = Here are some similar German sentences to guide your translation:	48.50	85.61
P2.D	[German 1] = Zu diesem Zweck sollten die Maßnahmen ... [German 2] = Diese Sanktionen müssen wirksam , verhältnismäßig ... [German 3] = Die Sanktionen müssen in angemessenem Verhältnis ... [German 4] = (17) Um die Einhaltung dieser ... [German 5] = Die Sanktionen für solche Zuwiderhandlungen ...		

Table 6: Translation Performance of GPT-4.1 using Implicit Guidance Prompts (P2 Series).

Zero-shot (en): These actions are examples of measures involving trade defense instruments, imposed by the United States on imports of the product concerned, and are in addition to the safeguard measures mentioned above.
RMCP (en): (441) These actions are examples of measures involving trade defence instruments (TDIs), imposed by the United States on imports of the product concerned, and are in addition to the safeguard measures mentioned above.
Retrieved Monolingual Examples:

- Example 1: (445) The decrease in the level of imports of the above products to the US can be explained , primarily , by reference to the direct effect of the measures referred to above on exporters whose goods became subject to duties as part of trade defence instruments (TDIs) ; but also by the indirect dissuasive effect of those measures and the investigations on all exporters of the product concerned to the US.
- Example 2: Imports into the Community of the products listed in the Annex and originating in the United States of America are hereby made subject to retrospective Community surveillance based on import declarations or any other equivalent import document required before the entry into force of this Regulation.
- Example 3: A 14 % ad valorem duty additional to the customs duty applicable under Regulation (EEC) No 2913 / 92 shall be imposed on the products originating in the United States of America listed in the Annex to this Regulation .. "
- Example 4: The protective measures shall be applied in compliance with any obligations arising under international agreements binding on the Community.
- Example 5: This may be particularly true to a degree for those countries which have taken protective measures to counteract the US safeguard action.

Case 2

Source (fr): relative à la signature et à la conclusion d' un procès-verbal agréé entre la Communauté européenne et le gouvernement de la République de Corée sur le marché mondial de la construction navale
Reference (en): on the signing and conclusion of Agreed Minutes between the European Community and the Government of the Republic of Korea relating to the world shipbuilding market
Zero-shot (en): concerning the signing and conclusion of an agreed minutes between the European Community and the government of the Republic of Korea on the world shipbuilding market
RMCP (en): on the signing and conclusion of Agreed Minutes between the European Community and the Government of the Republic of Korea concerning the world shipbuilding market
Retrieved Monolingual Examples:

- Example 1: The Agreed Minutes between the European Community and the Government of the Republic of Korea relating to the world shipbuilding market are hereby approved.
- Example 2: concerning the conclusion of the framework Agreement for Trade and Cooperation between the European Community and its Member States , on the one hand , and the Republic of Korea , on the other hand
- Example 3: on the conclusion of an Agreement between the European Community and the Government of the People 's Republic of China on cooperation and mutual administrative assistance in customs matters
- Example 4: on the signing , on behalf of the Community , of the Agreement between the European Community and the Kingdom of Denmark on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters
- Example 5: on the signing , on behalf of the Community , of the Agreement between the European Community and the Kingdom of Denmark on the service of judicial and extrajudicial documents in civil or commercial matters

Case 3

Source (fr): considérant que l' application de ces normes doit avoir pour effet d' éliminer du marché les produits de qualité non satisfaisante , d' orienter la production de façon à satisfaire aux exigences des consommateurs et de faciliter les relations commerciales sur la base d' une concurrence loyale ,

en contribuant ainsi à améliorer la rentabilité de la production;
Reference (en): Whereas the application of the standards should eliminate poor quality products from the market , direct production to meet consumer requirements and facilitate trade on the basis of fair competition , so helping to improve profitability;
Zero-shot (en): considering that the application of these standards should have the effect of eliminating from the market products of unsatisfactory quality, of directing production so as to meet consumer requirements and of facilitating trade relations on the basis of fair competition, thus helping to improve the profitability of production;
RMCP (en): Whereas the application of these standards should have the effect of removing products of unsatisfactory quality from the market, directing production to meet consumer requirements, and facilitating trade relations on the basis of fair competition, thereby helping to improve the profitability of production;
Retrieved Monolingual Examples:

- Example 1: Whereas the effect of these standards must be to remove products of unsatisfactory quality from the market , gear production to satisfying consumer requirements and facilitate trade relations on the basis of fair competition , thereby helping to make production more profitable;
- Example 2: Whereas application of those standards should have the effect of keeping products of unsatisfactory quality off the market , directing production to meet consumers ' requirements and facilitating trade under fair competitive conditions , thereby helping to improve the profitability of production;
- Example 3: Whereas application of these standards should have the effect of keeping products of unsatisfactory quality off the market , guiding production to meet consumers ' requirements and facilitating trade under fair competitive conditions , so helping to improve the profitability of production;
- Example 4: (3) Whereas the result of applying these standards should be to remove products of unsatisfactory quality from the market , to bring production into line with consumer requirements , to facilitate trade relationships based on fair competition , and thereby to help make production more profitable;
- Example 5: (3) Whereas the effect of these standards must be to remove products of unsatisfactory quality from the market , gear production to satisfying consumer requirements and facilitate trade relations on the basis of fair competition , thereby helping to make production more profitable;

Case 4
Source (fr): La Commission s' assure du respect de l' article 11 et du paragraphe 1 point b) du présent article par des contrôles à effectuer conformément au titre VI et , à la suite de ces derniers , demande , le cas échéant , aux États membres de retirer les reconnaissances accordées.
Reference (en): The Commission shall check that Articles 11 and paragraph (1) (b) of this Article are complied with by carrying out checks in accordance with Title VI and in the light of such checks shall , where appropriate , call on Member States to withdraw recognition.
Zero-shot (en): The Commission ensures compliance with Article 11 and paragraph 1 point (b) of this Article by carrying out checks in accordance with Title VI and, following these checks, requires, where appropriate, the Member States to withdraw the recognitions granted.
RMCP (en): The Commission shall check that Article 11 and paragraph 1(b) of this Article are complied with by carrying out checks in accordance with Title VI and, as a result of those checks, shall, where appropriate, request Member States to withdraw the recognition granted.
Retrieved Monolingual Examples:

- Example 1: In order to ensure that Article 5 and paragraph (1) (b) of this Article are complied with , the Commission shall carry out checks and in the light of such checks may , where appropriate , request that Member States withdraw recognition.
- Example 2: The Commission shall check that paragraphs 2 and 4 (b) are complied with by carrying out checks in accordance with Title VI and may , as a result of these checks , ask a Member State to withdraw recognition.
- Example 3: The Commission shall check that paragraph 1 and 3 (b) are complied with by carrying out checks and may , as a result

of those checks , request a Member State to withdraw recognition.

- Example 4: The Commission shall check that Article 39 and paragraph 1 (b) of this Article are complied with by carrying out checks and in the light of such checks shall , where appropriate , call on Member States to withdraw recognition.
- Example 5: Member States shall report to the Commission on the derogation concerning paragraphs (a) and (b) in accordance with Article 11.

Case 5

Source (fr): (7) Il y a lieu d' établir une procédure de détermination du statut épidémiologique par rapport à l' ESB d' un État membre , d' un pays tiers et d' une de leurs régions (ci-après dénommés " pays ou régions ") sur la base du risque d' introduction initiale (en anglais " incident risk ") , de propagation et d' exposition humaine en utilisant les informations disponibles.
Reference (en): (7) A procedure should be established for the determination of the epidemiological status of a Member State , a third country and of one of their regions , hereinafter referred to as " countries or regions " with respect to BSE , on the basis of the incident propagation and human exposure risk , using information available.
Zero-shot (en): It is necessary to establish a procedure for determining the epidemiological status with regard to BSE of a Member State, a third country, or one of their regions (hereinafter referred to as "countries or regions") on the basis of the risk of initial introduction ("incident risk"), spread, and human exposure, using the available information.
RMCP (en): (7) A procedure should be established for determining the epidemiological status with regard to BSE of a Member State, a third country and any of their regions (hereinafter referred to as "countries or regions") on the basis of the risk of initial introduction ("incident risk"), spread, and human exposure, using the available information.
Retrieved Monolingual Examples:

- Example 1: (7) On the basis of the draft list drawn up by the Commission in agreement with each of the Member States concerned , which also identifies sites hosting priority natural habitat types or priority species , a list of sites selected as sites of Community importance should be adopted.
- Example 2: (6) Based on a common integrated risk analysis model , the Agency should carry out risk analyses in order to provide the Community and the Member States with adequate information to allow for appropriate measures to be taken or to tackle identified threats and risks with a view to improving the integrated management of external borders.
- Example 3: (10) Certain ruminant tissues should be designated as specified risk material on the basis of the pathogenesis of TSEs and the epidemiological status of the country or region of origin or residence of the animal concerned.
- Example 4: (7) Preparatory measures need to be taken at Member State and Community level to enable assistance intervention teams in emergencies to be mobilised rapidly and coordinated with the requisite flexibility and to ensure , through a training programme , the effective response capability and complementarity of assessment and / or coordination teams , intervention teams and other resources , as appropriate.
- Example 5: (5) On the basis of the draft list drawn up by the Commission in agreement with each of the Member States concerned , which also identifies sites hosting priority natural habitat types or priority species , a list of sites selected as sites of Community importance should be adopted.

Case 6

Source (fr): considérant dès lors qu' il doit être possible à l' agence européenne pour l' évaluation des médicaments de procéder à l' évaluation , sur demande d' une firme , de tout médicament vétérinaire contenant une nouvelle substance active qui , à la date d' entrée en vigueur du présent règlement , n' a pas encore été autorisée par un État membre pour l' usage chez les animaux;
Reference (en): Whereas therefore it must be possible for the European Agency for the Evaluation of Medicinal Products to carry out an evaluation , at the request of a company , of any veterinary medicinal product containing a new active substance which , on the date of entry into force of this

Regulation , was not authorised by any Member State for use in animals;
Zero-shot (en): whereas it must therefore be possible for the European Agency for the Evaluation of Medicinal Products to carry out the evaluation, at the request of a company, of any veterinary medicinal product containing a new active substance which, on the date of entry into force of this regulation, has not yet been authorized by a Member State for use in animals;
RMCP (en): Whereas it should therefore be possible for the European Agency for the Evaluation of Medicinal Products to carry out, at the request of a company, the evaluation of any veterinary medicinal product containing a new active substance which, on the date of entry into force of this Regulation, has not yet been authorised by a Member State for use in animals;
Retrieved Monolingual Examples:

- Example 1: 'Veterinary medicinal products containing a new active substance which , on the date of entry into force of this Regulation , was not authorised by any Member State for use in animals'.
- Example 2: Furthermore , a Member State should be able to suspend the examination of an application for authorization to place a veterinary medicinal product on the market which is currently under active consideration in another Member State with a view to recognizing the decision reached by the latter Member State.
- Example 3: Whereas , pending the adoption by the Commission of the animal health certification for importation of fresh poultrymeat from the countries specified in this list , Member States may continue to apply , on importation , their animal health requirements in force on 1 January 1994;
- Example 4: Furthermore, a Member State should be able to suspend the examination of an application for authorization to place a medicinal product on the market which is currently under active consideration in another Member State with a view to recognizing the decision reached by the latter Member State.
- Example 5: Whereas a period of 60 days should be allowed before the entry into force of this Regulation in order to allow Member States to make any adjustment which may be necessary to the authorizations to place the veterinary medicinal products concerned on the market which have been granted in accordance with Council Directive 81 / 851 / EEC (4) , as last amended by Directive 93 / 40 / EEC (5) , to take account of the provisions of this Regulation ; ex II.

Preserving Ambiguity: Prompt Sensitivity in Gender-Neutral Literary Translation by GPT Models

Jin Yim
Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu,
Seoul 03760, Republic of Korea
jin.yim@ewha.ac.kr

Abstract

This study aims to evaluate how effectively various prompting strategies influence GPT model translations of gender-neutral Korean third-person singular expressions into English, using the queer literary novel *Concerning My Daughter* (Kim, 2017) as a corpus. Specifically, it investigates how the effects of prompting strategies for translating a specific expression have evolved across GPT model versions (GPT-3.5-turbo, GPT-4, and GPT-4o). Through quantitative analyses (BLEU, TER, BERTscore), multivariate statistical techniques (MANOVA, PCA, CA), and qualitative examinations, this research demonstrates significant improvements in translation quality and gender neutrality when context-aware prompts are employed. Meta prompting explicitly emphasizing gender ambiguity was most effective with advanced GPT versions (GPT-4 and GPT-4o), though overly complex prompts did not consistently improve quality metrics. The findings highlight critical limitations in current evaluation frameworks, advocating for specialized criteria and ethical frameworks in AI-generated literary translation.

1 Introduction

As AI-based machine translation (MT) has become commonplace and increasingly applicable to various types of texts, recent studies have begun to explore whether AI can adequately handle sophisticated and subtle literary translation tasks (Alsajri, 2023; Hu & Li, 2023; Li, 2024; Mukti et al., 2024).

This study focuses specifically on the underresearched issue of translating gender-marked references, which poses particular challenges in Korean-to-English literary translation. By examining multiple versions of a single generative AI (GAI) model, the research investigates how sophisticated prompting strategies contribute to producing literary translation outputs increasingly closer to human-level quality.

In particular, gender-neutral language usage is an important concern in MT research, as previous studies have reported gender stereotypes in both neural machine translation (NMT) models (Stanovsky et al., 2019; Vanmassenhove et al., 2018) and large language models (LLMs) (Piazzolla et al., 2024). Research has particularly focused on characterizing and resolving gender bias occurring between specific language pairs such as English and languages with grammatical gender systems, including French and Italian (Ghosh & Caliskan, 2023; Piazzolla et al., 2024; Sant et al., 2024). Although improvements in MT quality have been reported in terms of mitigating gender bias, several challenges still remain unresolved (Piazzolla et al., 2024). Moreover, existing research has predominantly concentrated on binary gender distinctions, while attention toward non-binary gender issues in translation is increasingly gaining prominence (Kostikova, 2023; Yim, 2025; Yu, 2025).

Against the backdrop, this study qualitatively and quantitatively evaluates how the translation outputs of Korean expressions referring to non-binary gender in queer literature vary according to different prompts (zero-shot, context, and meta prompting) and model versions (GPT-3.5 turbo, GPT-4, and GPT-4o) of an LLM, using human translation as the baseline. In doing so, this research aims to contribute to the literature on ethical considerations in AI-generated language usage and literary translation into English. The research questions are specifically formulated as follows:

- How do automated translation quality metric scores for literary sentences containing non-binary gender expressions vary according to prompts and model versions?
- How do gendered or ungendered translation patterns of non-binary gender references differ depending on variations in model versions and prompting strategies?

This study extends previous research (Yim, 2025), which examined prompt sensitivity in a single GPT version (GPT-4o). This comparison allows for an analysis of how prompt sensitivity has evolved across different GPT versions. Furthermore, translating non-binary gender is closely linked to ethical issues in LLM-generated language usage, specifically regarding gender bias, and this comparison will provide insight into how these ethical considerations have evolved across GPT models. Given the observation that non-binary gender has been particularly underexplored in translation and linguistics studies involving Korean (Yu, 2025), this research contributes to expanding existing linguistic scholarship beyond the MT literature by addressing a critical gap in the field.

2 Literature Review

This section first discusses why translating non-binary gender references from Korean to English raises significant ethical and linguistic issues, particularly from a human translation perspective (2.1). It then reviews previous research on how MT has addressed these challenges, highlighting the linguistic implications of these approaches (2.2). Finally, the chapter explores how translation outcomes have been improved across recent GPT models with prompting techniques (2.3).

2.1 Translating Non-binary Gender

The manner in which gender is expressed varies significantly across linguistic systems. English follows a natural gender system, meaning that only nouns and pronouns explicitly indicate a person's gender (Sabato & Perri, 2020). Korean similarly employs the identical gender system; however, its third-person pronoun usage notably differs from English. Korean lacks a fully developed third-person singular pronoun system, instead utilizing demonstratives such as i ("this"), *geu* ("that"), and *jeo* ("that over there") combined with nouns like *saram* ("person"), *i* ("one"), *nom* ("guy"), and *ja* ("individual") to denote third-person referents (Ko & Koo, 2020). Importantly, these combined forms generally do not explicitly indicate grammatical gender, with only a few specialized exceptions, such as *geunyeo* ("she"), which originated from translations of the Japanese equivalent of the English pronoun "she" (Ahn, 2001).

In English, third-person singular pronouns explicitly indicate gender through forms such as "he" and "she," "her" and "his," or "him." However, a significant issue arises because masculine pronouns in English have historically been employed generically to refer to individuals of all genders—a practice increasingly recognized as sexist language (Sabato & Perri, 2020, p. 334). As alternatives, more inclusive forms such as "s/he" and the singular "they" have been adopted. The use of singular "they" has gained formal recognition, appearing in major dictionaries, including the Oxford English Dictionary (2024), where it is explicitly described as suitable for contexts requiring gender neutrality or for referring to individuals who do not identify within the binary gender framework.

In summary, translating third-person singular references from Korean into English necessitates context-based inference to determine whether binary gendered pronouns ("he" or "she") or more inclusive non-binary forms ("s/he" or "they") should be employed. Additionally, it is possible to refer to an individual using a proper noun or a general noun that is either explicitly gendered (e.g., "the man" or "the woman") or non-gendered (e.g., "the person"). This inferential process significantly impacts translation accuracy, intensifying the challenge of maintaining gender-neutral language usage. Especially when gender ambiguity is intentionally encoded in the source text, the translator's choice between gendered or ungendered forms can considerably affect the discourse functions of the translated output, as demonstrated in previous studies such as Aguilar (2023), Ivan (2024), and Yim (2025).

2.2 Issues of Gender Translation in MT

When translating gender-neutral items whose grammatical gender is unknown into gender-marked languages, the process of gender inference becomes essential. If contextual information is insufficient, MT systems either infer and assign a gender or translate in a manner that preserves gender ambiguity. Issues of gender in MT have been primarily explored in relation to gender bias. Previous research has indicated that, when MT systems translate source texts without explicit gender information into target languages that obligatorily mark gender, they tend to exhibit a default bias toward masculine forms (Ghosh & Caliskan, 2023; Piazzolla et al., 2024; Stanovsky et al., 2019; Vanmassenhove et al., 2018). Large language models (LLMs) are reported to be even more susceptible to this bias (Sant et al., 2024). For instance, ChatGPT frequently translates gender-neutral pronouns

into explicitly gendered forms such as "he" or "she" (Ghosh & Caliskan, 2023). Similarly, DeepL has shown a notable tendency to overuse the pronoun "he" in backtranslations from Finnish, Estonian, and Indonesian into English. This bias is particularly influenced by sentence context and verbs used, demonstrating high reproducibility across repeated translations (Barclay & Sami, 2024). Such gender bias can significantly impact translation accuracy and discourse effectiveness by introducing gender markers absent from the original text or incorrectly inferring a different gender.

2.3 Improved Translation with Prompting

LLM prompting strategies are effective the translation outputs (Yamada, 2024). Accordingly, various studies have explored prompt design to enhance translation performance across multiple GPT model versions (He, 2024; Peng et al., 2023; Sant et al., 2024; Wang et al., 2023). Several strategies have been found beneficial in enhancing LLM-generated translations: For example, providing contextual information such as detailed translation guidelines and domain-specific knowledge (Peng et al., 2023); assigning a translator persona to the GPT model (He, 2024); and instructing the model to translate at the document level to leverage broader contextual understanding (Wang et al., 2023). It is also known that multi-shot prompts, which provide actual translation examples, improve translation quality more effectively than zero-shot prompts (Sant et al., 2024). Particularly, explicitly instructing the model about the translation's specific purpose and emphasizing the reduction of gender bias has proven effective in decreasing biased outcomes (Sant et al., 2024). Recently, meta-prompting techniques informing the LLM of task details and subsequently allowing it to generate its own prompts have also demonstrated potential for improving translation quality (Suzgun & Kalai, 2024).

However, some studies suggest that simpler prompts might be preferable to overly complex ones. Specifically, concise and effective prompts, such as zero-shot prompts (He, 2024), have been reported to yield better translation quality improvements compared to prompts containing detailed translation briefs and elaborate instructions (Peng et al., 2023). As discussed above, translation outputs significantly vary according to prompt configurations. In terms of GPT model performance, GPT-4 has shown substantial improvements over GPT-3.5 in both translation quality (Jiao et al., 2023; Yan et al., 2024) and other performance metrics (Chen et al., 2024). In particular, GPT-4 reportedly outperforms junior translators but still falls short of the translation quality produced by experienced human translators, showing a tendency toward literal translation (Yan et al., 2024).

In summary, two critical gaps currently exist in GPT models' handling of non-binary gender translations. First, while ChatGPT's translation performance has improved in recent versions, it still has not matched the quality of experienced human translators. Moreover, most prompt-related studies discussed above have primarily focused on single-model versions, leaving uncertainty about how the same prompting strategies affect translation outcomes across different GPT versions. Second, despite extensive prompting research to address gender-related translation problems (Sant et al., 2024), gender biases and translation errors persist. The fact that translation issues remain even in relatively clear-cut cases involving binary gender suggests that the challenges become considerably more complex when translating non-binary gender references. Addressing this significant issue, particularly from the perspective of inclusive language use, calls for future interdisciplinary research evaluating inclusive translation across diverse languages, textual contexts, and GPT model versions. Given these research gaps, the present study aims to examine how translations of sentences containing non-binary gender expressions—requiring heightened gender sensitivity—have evolved across GPT model versions, which currently represent the most widely adopted LLM technology. The findings will indirectly contribute to identifying approaches to accurate gender inference through contextual and grammatical cues, while also addressing methods to preserve the discourse effects of the original text and uphold ethical principles related to gender-neutral language use. Consequently, this study provides significant insights into the current state of AI-generated translation.

3 Methodology

3.1 Corpus

The source text corpus used in this study is drawn from the Korean novel *Concerning My Daughter* (Kim, 2017) by novelist Hyejin Kim. This queer novel portrays the internal struggles experienced by a mother who struggles to accept that her daughter is living with a same-sex partner, narrated from

the mother's perspective. Throughout the novel, the mother consistently refers to her daughter's non-binary gender partner using the third-person singular pronoun *geu ae* ("that person/child"), maintaining gender neutrality. From the novel, 183 sentences containing references to *geu ae* were identified and compiled into a corpus. Using this corpus, the study measures how gender neutrality is maintained or altered according to different prompting strategies across multiple GPT model versions. For baseline comparison, this paper used Jamie Chang's English translation (Kim, 2022), which was viewed by critics as "precise and pared-back" renditions of the original narrative "in a careful, balanced way" (West-Knights, 2022). The original text and the corresponding human translation corpus were also utilized in a previous analysis conducted by Yim (2025).

3.2 Process

To investigate translation choices concerning non-binary gender references, each sentence was translated independently using the API. The translation conditions and prompts used were consistent with those employed by Yim (2025) (see Appendix A). The context prompt has the gender and character name, while the simplified form of meta prompt includes the persona, genre, and gender ambiguity instructions.

3.3 Analysis

The analysis was divided into two parts: translation quality (TQ) and gender representation. First, TQ scores were quantitatively evaluated by combining multiple metrics, following the recommendation of Kocmi et al. (2021). Specifically, BLEU (Papineni et al., 2002) was used to assess basic similarity to the human baseline; TER (Snover et al., 2006) was applied to measure the practical effort required for post-editing; and BERTscore (Zhang et al., 2020) was employed to evaluate how effectively the contextual meaning of the original text was captured. To explore how translation quality scores varied according to prompts and GPT model versions, a MANOVA test was conducted. This was followed by PCA (Biber, 1988) to visually illustrate prompt sensitivity showing how prompts make LLM translations deviate from human translations across model versions. Additionally, qualitative analysis involved to identify distinctive patterns and variations.

Second, the analysis examined the representation of non-binary gender in translation outputs (gender representation). The frequency of gendered versus ungendered reference expressions was analyzed across the nine generated corpora and compared against the human translation baseline. Descriptive statistical analysis (chi-square test) and explanatory analysis (Correspondence Analysis; Glynn, 2014) were performed to determine which translations most closely resembled the human baseline in terms of gender-neutrality patterns, while also identifying distinctive translation characteristics according to prompts and GPT model versions.

Python 3.11.8 (July 14, 2025)	**R 4.5.1 (July 20, 2025)**
pandas: 1.5.3	FactoMineR: 2.12
numpy: 1.24.0	factoextra: 1.0.7
scipy: 1.9.3	CA: 0.71.1
matplotlib: 3.6.3	**GPT-3.5 turbo, 4, 4o**
seaborn: 0.11.2	API version: 1.13.3
openpyxl: 3.0.10	Temperature: 0.7
nltk: 3.8.1	Max tokens: 300
sacrebleu: 2.3.1	Top-p: 1.0
bert-score: 0.3.13	Penalty: 0
Okt, konplay 0.6.0	Date: July 3, 2025

Table 1: System and package information

Corpus compilation, data analyses, and statistical testing were conducted using Python and R within the Google Colab environment. Information on the specific models and packages utilized is presented in Table 1.

Corpus (version_prompt)	**ID**	**Token**
Source text	ST	3108
Human translation	HT	2610
GPT-3.5_simple	GPT35_TT1	2948
GPT-3.5_context	GPT35_TT2	2927
GPT-3.5_meta	GPT35_TT3	2868
GPT-4_simple	GPT4_TT1	2794
GPT-4_context	GPT4_TT2	2780
GPT-4_meta	GPT4_TT3	2737
GPT-4o_simple	GPT4o_TT1	2836
GPT-4o_context	GPT4o_TT2	2878
GPT-4o_meta	GPT4o_TT3	2765

Table 2: Corpus size

4 Results

4.1 Translation Quality Metrics

Information about the corpora (number of tokens) generated based on the analysis process presented in Section 3.3 is provided in Table 2.

Table reports corpus-level mean $\pm$ SD ($n = 183$ sentences) for each model-prompt condition. TER is reported as TER inverse so that higher values indicate better quality. Across prompting

TQ per corpus		GPT35	GPT4	GPT4o
BLEU	**TT1**	0.111 ± 0.115	0.123 ± 0.124	0.132 ± 0.127
	TT2	0.119 ± 0.129	0.148 ± 0.147	0.162 ± 0.157
	TT3	0.12 ± 0.136	0.126 ± 0.142	0.152 ± 0.157
	Pairwise	-	-	-
TER inverse	**TT1**	11.228 ± 33.889	12.411 ± 37.514	14.627 ± 34.519
	TT2	16.958 ± 31.414	21.303 ± 32.501	23.459 ± 32.174
	TT3	14.818 ± 32.157	16.318 ± 34.866	23.006 ± 30.915
	Pairwise	-	-	-
BERTScore	**TT1**	0.921 ± 0.024	0.923 ± 0.024	0.925 ± 0.023
	TT2	0.924 ± 0.024	0.929 ± 0.025	0.932 ± 0.025
	TT3	0.924 ± 0.024	0.924 ± 0.026	0.933 ± 0.025
	Pairwise	-	-	tt1 < tt2 *; tt1 < tt3 *

Table 3: TQ description and pairwise comparison

strategies, context prompting (TT2) and meta prompting (TT3) generally produced higher mean scores than simple prompting (TT1) for BLEU, TER_inverse, and BERTScore. TT2-TT3 comparisons were mixed: meta prompting exceeded context prompting only for GPT-3.5 on BLEU and for GPT-4o on BERTScore; in all other cases, TT2 showed the higher mean.

For each model, prompt effects were tested with a Kruskal-Wallis omnibus test followed by Dunn pairwise tests with Holm adjustment. Non-significant pairwise contrasts are omitted. Significant differences emerged only for GPT-4o on BERTScore ($tt1 < tt2^*$; $tt1 < tt3^*$; * $p < .05$, Holm-adjusted).

To further assess how prompt types and GPT model versions influenced changes in these metrics, a MANOVA test was performed (Appendix B). Although MANOVA generally requires multivariate normality, the large sample size ($n = 183$) ensures that the analyses conducted remain robust due to the central limit theorem.

The MANOVA results revealed statistically significant effects of prompt type (Wilks' $\lambda = 0.9888$, $F(6, 3284) = 3.08$, $p = .005$) and model version (Wilks' $\lambda = 0.9847$, $F(6, 3284) = 4.24$, $p < .001$) on the three translation quality metrics. In contrast, the interaction between prompt type and model version was not significant (Wilks' $\lambda = 0.9961$, $F(12, 4328.74) = 0.53$, $p = .899$). Although the Wilks' values were close to 1, indicating that the overall effect sizes were small, both prompt type and model version contributed to significant variation in translation quality scores.

Finally, a Principal Component Analysis (PCA) was conducted using the three translation quality metrics scores across corpora to visually represent how prompts make translations deviate from human translations across versions (see Figure 1 and Appendix C for detailed results).

PC1 accounted for comprehensive translation quality, evenly reflecting all three metrics (75.87%), while PC2 captured differences between BERTScore and the other two metrics (12.8%). Compared to BLEU and TER, BERTScore reflects semantic similarity rather than surface-level overlap. The relatively higher BERTScore therefore suggests that, even when lexical realizations diverge, the generated translations retain meaning-level consistency. Together, these two principal components explained 88.67% of the total variance across corpora. Prompt sensitivity was visualized by calculating convex hull areas based on PCA scores. Results indicated that the changes in the areas across prompts remained substantial in GPT-4 (TT1 → TT2: -1.575, TT2 → TT3: 5.011) and GPT-4o (TT1 → TT2: 2.206, TT2 → TT3: 2.859), compared to GPT-3.5 (TT1 → TT2: 0.893, TT2 → TT3: 0.174). This suggests that GPT-4 and GPT-4o exhibited greater sensitivity to prompt changes, reflected in broader variations in translation quality scores compared to the earlier version.

This trend is evident in Example 1 (Appendix D), which shows that score changes increased in conjunction with TT2 and TT3 in GPT-4 and GPT-4o.

4.2 Gender Representation

Corpus	Gendered				Ungendered			
	Female pronoun	Female noun	Male		Proper noun	Pronoun	Noun	
Human	101	21	1	**123**	110	11	2	**123**
GPT35_tt1	78	1	15	**94**	0	26	129	**155**
GPT35_tt2	198	14	1	**213**	25	16	29	**70**
GPT35_tt3	207	2	0	**209**	42	18	17	**77**
GPT4_tt1	51	2	32	**85**	0	19	146	**165**
GPT4_tt2	159	11	2	**172**	112	11	5	**128**
GPT4_tt3	167	5	0	**172**	127	10	5	**142**
GPT4o_tt1	49	10	10	**69**	0	51	134	**185**
GPT4o_tt2	201	12	2	**215**	37	15	5	**57**
GPT4o_tt3	154	4	0	**158**	100	13	3	**116**

Table 4: Frequency of gender representation

Next, to address Research Question 2 regarding gender neutrality, the analysis focused on the frequency of gender reference expressions used across

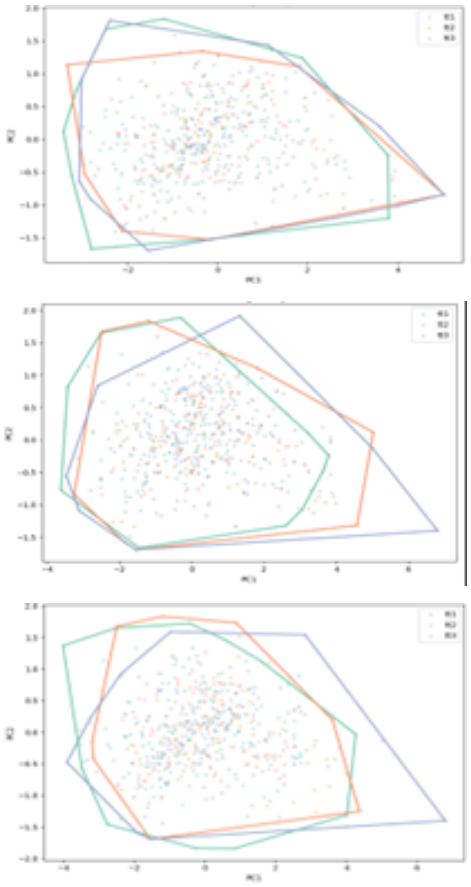
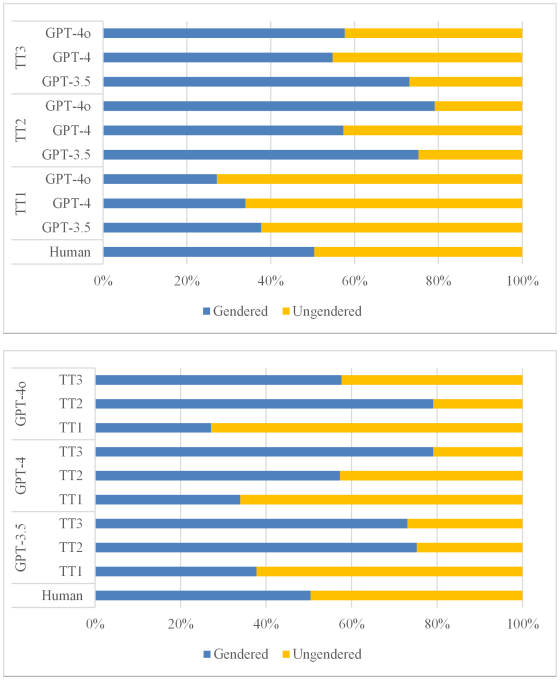


Figure 2: Modelwise prompt effects (GPT-3.5, 4, 4o, top to bottom)



Note: Panel (a) groups the results by prompt type (TT1–TT3), while Panel (b) groups them by model. Both panels display the same dataset in horizontal bar format to improve readability and highlight different comparative perspectives.

Figure 1-a (top), b (bottom). Frequency of gendered vs. ungendered expressions across human and LLM outputs.

each corpus. Following the analytical categories established in previous research (Yim, 2025), frequencies of English translations corresponding to Korean ungendered third-person references were measured. Table 4 presents the corpus-specific frequencies of expressions identified as the most commonly used English translations for the Korean term *geu ae* ("that person/child").

Since some cells contained frequencies of five or less, a chi-square test was conducted using aggregated frequencies of gendered and ungendered expressions across each corpus. For simplicity of comparison, we employed chi-square tests, which capture only overall associations. Such analysis lies beyond the present scope and will be pursued in future work. Although the chi-square test is limited in that it does not capture interaction effects between variables, more advanced approaches such as log-linear modeling would be required for this purpose. The test revealed statistically significant differences among corpora ($\chi^2 = 306.490$, $df = 9$, $p = 1.09E - 60$).

In particular, for the TT1 prompt, which provided no explicit gender-related instructions, the frequency of gendered expressions—indicating binary gender assignment through inference—decreased progressively from GPT-3.5 (94) to GPT-4 (85), and further to GPT-4o (69). This trend suggests increased gender sensitivity in higher GPT model versions (Figure -a).

For context prompting (TT2), which explicitly provided the character name and gender information, the frequency of gendered expressions was highest for GPT-4o (215), followed by GPT-3.5 (213), and GPT-4 (172). Notably, GPT-4 produced the highest frequency of gender-neutral references (128) despite explicit gender information (Figure -a).

In meta-prompting (TT3), which highlighted genre characteristics and emphasized the importance of gender ambiguity, all GPT models showed increased frequencies of ungendered expressions compared to context prompting (TT2). However, the magnitude of this increase varied considerably across models: GPT-3.5 showed an increase of 7 instances (TT2: 70 → TT3: 77), GPT-4 an increase

of 14 (TT2: 128 → TT3: 142), while GPT-4o exhibited a notably higher increase of 59 (TT2: 57 → TT3: 116) (Table 4, Figure -b).

Notably, GPT-4 and GPT-4o under the TT3 prompting condition exhibited proportions of gendered and ungendered expressions most similar to the human translation baseline. This suggests that prompts incorporating contextual information and explicit instructions emphasizing gender ambiguity were more effective than zero-shot prompting in achieving human-like gender neutrality. However, GPT-3.5 did not achieve human-level performance in representing gender ambiguity.

To further provide explanatory statistical insights, a two-dimensional correspondence analysis (CA) was conducted, offering a detailed visual representation of distances between the human translation and the nine MT-generated corpora concerning specific expressions (Figure 3). Contributions and coordinates for each corpus and expression are provided in Appendix E.

Table 5 shows that the eigenvalue for the first dimension (Dim1) was notably high at 0.554, surpassing the conventional threshold (≥ 0.4) for a robust dimension. Although the eigenvalue for the second dimension (Dim2) was relatively lower at 0.109, it was still meaningful for providing additional explanatory value. Regarding explained inertia, Dim1 accounted for 71.84% of the total variance, and the cumulative inertia up to Dim2 was 86.82%, indicating sufficient reliability and explanatory power for interpreting the analysis results.

Dimension	Eigenvalue	Explained_Inertia
Dim1	0.554	71.84%
Dim2	0.110	14.28%

Table 5: CA Inertia

Figure 3 shows the correspondence analysis map using the row principal method (FactoMineR + factoextra default). The rows (corpora) are represented in principal coordinates, while the columns (lexical items) are shown as supplementary points. CA results revealed that GPT-4 TT2, GPT-4 TT3, and GPT-4o TT3 were positioned closest to human translation regarding the use of gendered reference expressions. Conversely, as expected, the three TT1 corpora without explicit contextual information diverged significantly from human translation. Despite the provision of context and instructions em-

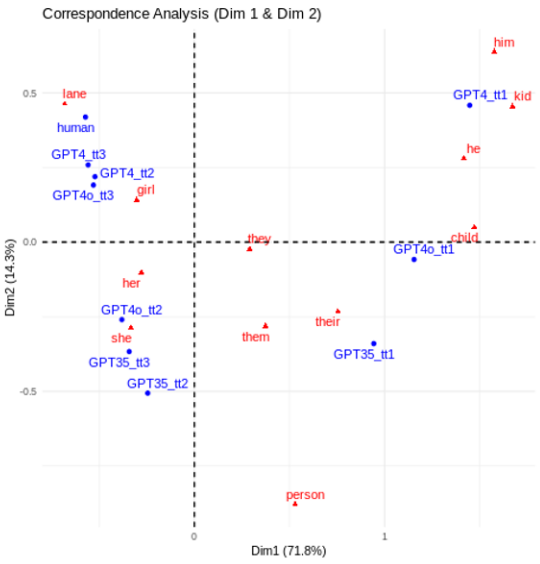


Figure 3: CA Results.

phasizing gender ambiguity, GPT-3.5 TT3 still frequently produced gendered translations. This suggests that the effectiveness of prompts emphasizing gender ambiguity increased with the advancement of the GPT model versions, leading to more human-like translations.

In Dim1, which accounted for most of the corpus differences, the expressions with the highest contributions were *child* (34.8%), *kid* (24.4%), and *Lane* (17.0%). Predictably, the three TT1 corpora—which lacked the explicit character name—were positioned on the opposite end of Dim1. Human translation, GPT-4o TT3, GPT-4 TT2, and GPT-4 TT3 were closely positioned around the proper noun *Lane*, whereas GPT-4o TT2 focused more explicitly on the gender information provided rather than employing the character's name. Additionally, in the TT1 condition, GPT-3.5 favored inclusive forms, whereas GPT-4 defaulted toward masculine forms (Example 2 in Appendix D).

In Example 2, it is also worth mentioning that the pronoun "they" in the human translation could either represent a singular pronoun or, based on context, a plural form referring collectively to Lane and the protagonist's daughter. Given that GPT-3.5 TT1 appeared close to "they" in Figure 3, an examination of the corpus revealed two instances (Examples 3 and 4 in Appendix D) where GPT-3.5 likely employed the gender-inclusive singular form of "they." In contrast, the newer models GPT-4 and GPT-4o translated these references differently,

using "the child" (GPT-4 and GPT-4o, Example 3), gendered expressions such as "the kid" and "he" (GPT-4, Example 4), or "the girl" and "she" (GPT-4o, Example 4). These translation choices led general nouns like "child" and "kid" to significantly contribute to distinguishing the corpora along Dim1. Additionally, these tendencies explain why GPT-4o TT1 was closely positioned to explicitly masculine pronouns such as "he" and "him."

Moreover, despite detailed prompting instructions, GPT-3.5 and GPT-4 occasionally failed to correctly infer omitted subjects from the Korean source text or produced misgendered translations (Example 5, Appendix D). In contrast, GPT-4o accurately translated the source meaning across all prompting conditions in the same example, successfully reflecting the intended gender ambiguity through prompting.

5 Discussion and Conclusion

The evaluation of translation quality metrics across three prompting strategies and GPT model versions revealed that quality scores were more strongly influenced by prompts than by model versions. Specifically, context and meta prompting conditions yielded higher scores compared to zero-shot prompting. However, comparisons between context and meta prompting showed mixed outcomes: average scores decreased in six corpora but increased in only three corpora, suggesting that excessively complex prompts may not consistently enhance translation quality. Based on MANOVA results and convex hull visualization from PCA analysis indicated that newer GPT models tended to exhibit greater sensitivity to prompting variations. These findings indicate that prompt-induced improvements in translation quality are negligible for earlier models but emerge with GPT-4o, and primarily on a semantic similarity metric (BERTScore), suggesting that newer model families may better leverage context/persona instructions.

Furthermore, the analysis of frequencies of key gender-marking expressions demonstrated significant differences among corpora in their use of gendered versus ungendered forms. Human translation maintained a balanced 1:1 ratio. All three GPT models under zero-shot prompting tended toward ungendered forms, whereas context prompting resulted in substantially higher proportions of gendered forms. Meta prompting, which provided genre-specific context and explicit instructions emphasizing gender ambiguity with persona, led to increased use of ungendered expressions, with GPT-4o showing the largest increase.

The implications of these findings are as follows. First, although translation quality improved with context prompting compared to zero-shot prompting, adding more detailed information such as genre and gender ambiguity instructions failed to yield further improvements in some corpora, implying two possibilities. On one hand, it partially supports prior research indicating that overly complex prompts may hinder translation quality (He, 2024). On the other hand, it aligns with studies suggesting that providing context can enhance translation performance (He, 2024; Peng et al., 2023; Sant et al., 2024).

Second, the greater effectiveness of meta prompting over context prompting in maintaining gender ambiguity suggests limitations of current automated translation quality metrics in adequately capturing literary translation characteristics. Given the complexity of literary translation evaluation, it is necessary to adopt additional specialized metrics derived from corpus-based human translation studies (Liu et al., 2024) or to develop tailored evaluative frameworks (Zhang et al., 2025). This study particularly emphasizes the importance of incorporating translation criticism approaches—addressing ethical and ideological considerations—into assessments of AI-generated literary translations.

Third, instances where GPT-3.5 translated ambiguous references using singular "they," whereas newer models produced explicitly gendered translations (e.g., "the girl–she," "the child–he"), suggest a concerning trend. Specifically, more recent GPT outputs might not necessarily reflect increased inclusivity regarding non-binary gender. This finding, while preliminary, indicates the need for systematic further analysis. Although this study demonstrates that the latest models can produce gender-neutral translations when explicitly instructed, continued advancements in AI should explicitly aim to promote more inclusive language practices.

Despite the significance of these findings, this study has several limitations. First, the scope was limited to multiple versions of a single GPT model; however, given GPT's widespread use, this limitation is somewhat justified. Future studies should expand the analysis to include diverse models such as DeepL, Gemini and Claude to further examine developments in translating gender-neutral references. Second, the prompts utilized in this study specif-

ically focused on literary translation and gender neutrality. Although narrow, this linguistic focus effectively allows for a detailed exploration of translation guidelines and prompt efficacy. Lastly, due to the creative nature of literary translation tasks, the temperature setting was intentionally increased to 0.7; however, translations were generated through a single iteration. Future research should verify the stability of these findings through repeated translation tasks.

The implications of this study are threefold. First, by highlighting the challenges of generative AI translation between languages with differing gender grammars, this research provides significant insights not only into AI literature but also for translation studies and English writing education. Second, by examining how generative AI has evolved regarding gender grammar, this study underscores the necessity for ongoing research into ethical considerations within AI-generated translations. Lastly, by emphasizing the importance of prompt engineering in contemporary AI models, this study contributes to advancing creative translation research, particularly in literary domains utilizing AI.

Acknowledgments

This study is an extended follow-up to Yim (2025). The source text, baseline human translation corpus, and prompts used in this study are identical to those in the previous study (Yim, 2025) for comparability. The author deeply appreciates the anonymous reviewers' careful reading and insightful comments, which greatly helped improve this article. Any remaining errors or limitations are the author's sole responsibility.

References

Primary Sources

Hye-Jin Kim. 2017. *Concerning My Daughter*. Minumsa, Seoul.

Hye-Jin Kim. 2022. Concerning My Daughter, translated by Jamie Chang. Picador, London.

Secondary Sources

Young-hee Ahn. 2001. Translation of third-person pronouns <He> and <She> into Japanese and Korean: New fictional discourse through translation. *Journal of Japanese Language and Literature*, 17:147–172.

Abdulazeez Alsajri. 2023. Challenges in translating Arabic literary texts using artificial intelligence techniques. *EDRAAK 2023*, (February 2023):5–10. https://doi.org/10.70470/edraak/2023/002

Daniel Herencia Aguilar. 2023. Translating gender ambiguity in literatura: The case of *Written on the Body*. Skokpos, 12:137–160.

Peter J. Barclay and Ashkan Sami. 2024. Investigating markers and drivers of gender bias in machine translations. In Proceedings of the 2024 IEEE International Conference on Software Analysis, Evolution and Reengineering *(SANER)*, March 12, 2024. IEEE, Rovaniemi, Finland, pages 455–464. https://doi.org/10.1109/saner60148.2024.00054

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2). https://doi.org/10.1162/99608f92.5317da47

Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, August 08, 2023. ACM, Montréal QC Canada, pages 901–912. https://doi.org/10.1145/3600211.3604672

Dylan Glynn. 2014. Correspondence analysis: Exploring data and identifying patterns. In *Human Cognitive Processing*, Dylan Glynn and Justyna A. Robinson (eds.). John Benjamins Publishing Company, Amsterdam, pages 443–485. https://doi.org/10.1075/hcp.43.17gly

Sui He. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Kaibao Hu and Xiaoqian Li. 2023. The creativity and limitations of AI neural machine translation: A corpus-based study of DeepL's English-to-Chinese translation of Shake-

speare's plays. *Babel*, 69(4). https://doi.org/10.1075/babel.00331.hu

Alexandra Maria Ivan. 2024. Gender identity in translation: The impossibility of transposing non-binary characters into Romanian. *Forum for Contemporary Issues in Language and Literature*, 4:27–43. https://doi.org/10.34739/fci.2023.04.03

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. Retrieved from http://arxiv.org/abs/2301.08745

Young-Kun Ko and Bon-Kwan Koo. 2018. *Urimal Munbeopron [Korean Grammar Theory]*. Jipmoondang, Paju, Korea.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 478–494.

Aida Kostikova. 2023. Gender-neutral language use in the context of gender bias in machine translation (a review literature). *JCAL-JCAL*, 1(July 2023):94–109. https://doi.org/10.33919/JCAL.23.1.5

Qi Li. 2024. Bridging languages: The potential and limitations of AI in literary translation—A case study of the English translation of *A Pair of Peacocks Southeast Fly*. *AHR*, 8(1):1–7. https://doi.org/10.54254/2753-7080/8/2024091

Cuilin Liu, Se-Eun Jhang, Homin Park, and Hyunjong Hahm. 2024. A corpus-based multilingual comparison of AI-based machine translations. *kjell*, 24(January 2024):257–276. https://doi.org/10.15738/kjell.24..202404.257

Muhammad Abdee Praja Mukti, Muhamad Trian Maulana, Kharisma Nur Rohmah, Fourus Huznatul Abqoriyyah, and Andang Saehu. 2024. Effectiveness of artificial intelligence usage as translation medium among English literature student of UIN Sunan Gunung Djati Bandung. *JEEF*, 4(2):107–112. https://doi.org/10.29303/jeef.v4i2.683

Oxford English Dictionary. 2024. *Oxford English Dictionary Online*. Oxford University Press. Retrieved July 11, 2025, from https://www.oed.com/

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4390455

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*, Pennsylvania, USA, pages 311–318. https://doi.org/10.1002/andp.19223712302

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2024. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *Hermes – Journal of Language and Communication in Business*, 63:209–225. https://doi.org/10.7146/hjlcb.vi63.137553

Bruna Di Sabato and Antonio Perri. 2020. Grammatical gender and translation: A cross-linguistic overview. In *The Routledge Handbook of Translation, Feminism and Gender,* pages 363–373.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. https://doi.org/10.48550/ARXIV.2407.18786

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers,* pages 223–231.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. https://doi.org/10.48550/arXiv.1906.00591

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. https://doi.org/10.48550/arXiv.2401.12954

Eva Vanmassenhove, Christian Hardmeier, and

Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 3003–3008. https://doi.org/10.18653/v1/D18-1334

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. https://doi.org/10.48550/arXiv.2304.02210

Imogen West-Knights. 2022. In a Korean Best Seller, Women Have Biases, but No Options. The New York Times. https://www.nytimes.com/2022/09/26/books/review/concerning-my-daughter-kim-hye-jin.html

Masaru Yamada. 2024. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. https://doi.org/10.48550/arXiv.2308.01391

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. https://doi.org/10.48550/arXiv.2407.03658

Jin Yim. 2025. Gender ambiguity in human and AI translations: "That Person" in *Concerning My Daughter*. Korean Journal of English Language and Linguistics, 25:2–34. doi:10.15738/kjell.25..202507.912

Huijae Yu. 2025. A linguistic approach to queer translation – Focusing on non-binary identity. *The Journal of Translation Studies*, 26(2):79–103. https://doi.org/10.15749/JTS.2025.26.2.003

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. https://doi.org/10.48550/arXiv.2410.18697

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. https://doi.org/10.48550/arXiv.1904.09675

A Prompts (Yim, 2025)

Corpus	Prompt
TT1(Zero-shot)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. Please translate each Korean sentence into English."
TT2 (Context)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. In this context, '그 애' refers to Lane, a woman. Please translate each Korean sentence into English."
TT3 (Meta)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. You are a professional literary translator with deep sensitivity to gender identity, emotional nuance, and queer relationships. Please translate each Korean sentence into fluent, expressive, and natural English. Preserve ambiguity where appropriate, and maintain the rhythm, tone, and intimacy of the original text. In this context, '그 애' refers to Lane, a woman. Please translate each Korean sentence into English."

B MANOVA Results

Effect	Wilks' λ	F	df1	df2	p
Prompt	0.9888	3.0814	6	3284	0.0052
Model	0.9847	4.2355	6	3284	0.0003
Prompt × Model	0.9961	0.5265	12	4328.74	0.8991

C PCA Results

Explained variance and loadings

Principal Component	Metrics	Value
PC1	Explained Variance (%)	75.87
	BLEU_z Loading	0.579
	TER inverse z Loading	0.582
	BERTScore_z Loading	0.571
PC2	Explained Variance (%)	12.8
	BLEU_z Loading	-0.495
	TER inverse z Loading	-0.305
	BERTScore_z Loading	0.813

Centroid coordinates by Corpus

Corpus	PC1	PC2
GPT3.5 TT1	-0.311	-0.036
GPT3.5 TT2	-0.107	-0.019
GPT3.5 TT3	-0.142	-0.009
GPT4 TT1	-0.189	-0.019
GPT4 TT2	0.197	-0.005
GPT4 TT3	-0.088	-0.037
GPT4o TT1	-0.064	-0.001
GPT4o TT2	0.369	0.033
GPT4o TT3	0.336	0.095

Convex hull area changes across prompting

Version	TT1	Δ	TT2	Δ	TT3
GPT-3.5	18.161	0.893	19.054	0.174	19.228
GPT-4	19.932	1.575	18.356	5.011	23.368
GPT-4o	17.718	2.206	19.924	2.859	22.784

D Examples from Corpora

Example 1: ID #731
ST: 그 애가 잠시 고개를 들어 나와 눈을 맞춘다.
HT: She raises her head for a moment, meets my gaze.

GPT-3.5

TT1: That child briefly raises their head and meets my eyes.

TT2: She raises her head for a moment and meets my eyes (BLEU: 0, TER: -20, BERTScore: +0.0062).

TT3: She raises her head for a moment, meets my gaze (BLEU: 0, TER: 0, BERTScore: +0.005).

GPT-4

TT1: The kid lifts his head for a moment and makes eye contact with me.

TT2: Lane momentarily raises her head and makes eye contact with me (BLEU: 0.0081, TER: +80, BERTScore: +0.0086).

TT3: Lane lifts her head for a moment, locking eyes with me (BLEU: 0.0081, TER: +80, BERTScore: +0.0093).

GPT-4o

TT1: The child briefly lifts their head and makes eye contact with me.

TT2: Her eyes, which had been lingering at a point past me, meet mine (BLEU: -0.02 TER: +40, BERTScore: +0.0093).

TT3: She briefly lifts her head and meets my eyes (BLEU: -0.02, TER: +40, BERTScore: +0.0093).

Example 2: ID #773
ST: 그 애가 오고 난 후 내가 가장 처음 건넨 말은 그것이다.
HT: Those were the first words I said to her after they moved in.
GPT-3.5 TT1: The first thing I said after he/she came was that.
GPT-4 TT1: The first thing I said to him after he came was that.
GPT-4o TT1: The first thing I said after that child arrived was that.

Example 3: ID #669
ST: 그 애는 차분하게 자신의 이름을 말하고 찾아온 용건을 설명한다.
HT: Lane gets up in surprise and evenly states name and purpose of visit.
GPT-3.5 TT1: They calmly state their name and explain the reason for their visit.
GPT-4 TT1: The child calmly states his name and explains the purpose of his visit.
GPT-4o TT1: The child calmly states their name and explains the reason for their visit.

Example 4: ID #3693
ST: 주방에서 과일을 씻던 그 애는 젠이 잠든 줄로만 알았다고 했다.
HT: Lane, who was washing fruit in the kitchen, thought she was asleep.
GPT-3.5 TT1: The person washing fruits in the kitchen said that they thought Zen had fallen asleep.
GPT-4 TT1: The kid, who was washing fruits in the kitchen, said he thought Zen was asleep.
GPT-4o TT1: The kid, who was washing fruits in the kitchen, said he thought Zen was asleep.
GPT-3.5 TT2: She said that she thought Jen had fallen asleep just because she was washing fruit in the kitchen.
GPT-4 TT2: The girl, who was washing fruit in the kitchen, said she thought Zen had fallen asleep.
GPT-4o TT2: The girl, who was washing fruit in the kitchen, said she thought Jen was asleep.
GPT-3.5 TT3: The one washing fruit in the kitchen said she thought Jen had fallen asleep.
GPT-4 TT3: Washing fruit in the kitchen, Lane said she had thought Jen was asleep.
GPT-4o TT3: Lane, who was washing fruit in the kitchen, said she thought Jen was asleep.

Example 5: ID # 1350
ST: 오셨어요. 그 애다.
HT: Hi, you're home, says Lane.
GPT-3.5
TT1: You came. That kid.
TT2: She has arrived. (Ø)
TT3: She's here. It's her.
GPT-4
TT1: He's here. That's him.
TT2: She's here. It's Lane.
TT3: She's here. It's Lane.
GPT-4oS
TT1: You're here. It's that kid.

TT2: You're here. It's her.
TT3: You're here. It's Lane.

E Detailed CA Results

Corpus	Coordinates		Contributions	
	Dim 1	Dim 2	Dim 1	Dim 2
human	-0.572	0.419	5.351	14.450
GPT35_tt1	0.943	-0.340	14.672	9.591
GPT4_tt1	1.447	0.459	34.681	17.513
GPT4o_tt1	1.155	-0.058	22.348	0.286
GPT35_tt2	-0.245	-0.506	1.121	24.082
GPT4_tt2	-0.522	0.219	5.407	4.804
GPT4o_tt2	-0.381	-0.260	2.604	6.092
GPT35_tt3	-0.342	-0.367	2.219	12.827
GPT4_tt3	-0.558	0.259	6.489	7.018
GPT4o_tt3	-0.530	0.191	5.108	3.337

Coordinates & Contributions (Corpus)

Gender	Word	Coordinates		Contributions	
		Dim 1	Dim 2	Dim 1	Dim 2
Gendered	she	-0.332	-0.287	4.612	17.291
	her	-0.278	-0.101	3.811	2.552
	him	1.577	0.638	4.150	3.417
	he	1.417	0.280	5.093	1.001
	girl	-0.303	0.141	0.429	0.470
	woman	-0.680	0.463	17.086	39.860
Ungendered	lane	0.374	-0.281	0.280	0.798
	they	1.471	0.050	34.846	0.199
	their	1.672	0.455	24.469	9.094
	them	0.530	-0.878	1.728	23.842
	child	-0.332	-0.287	4.612	17.291
	kid	-0.278	-0.101	3.811	2.552
	person	1.577	0.638	4.150	3.417

Coordinates & Contributions (Gender Expression)

Bridging the Modality Gap by Similarity Standardization with Pseudo-Positive Samples

Shuhei Yamashita **Daiki Shirafuji** **Tatsuhiko Saito**
Mitsubishi Electric Corporation
{Yamashita.Shuhei@bc, Shirafuji.Daiki@ay, Saito.Tatsuhiko@db}
.MitsubishiElectric.co.jp

Abstract

Advances in vision-language models (VLMs) have enabled effective cross-modality retrieval. However, when both text and images exist in the database, similarity scores would differ in scale by modality. This phenomenon, known as the modality gap, hinders accurate retrieval. Most existing studies address this issue with manually labeled data, e.g., by fine-tuning VLMs on them. In this work, we propose a similarity standardization approach with pseudo data construction. We first compute the mean and variance of the similarity scores between each query and its paired data in text or image modality. Using these modality-specific statistics, we standardize all similarity scores to compare on a common scale across modalities. These statistics are calculated from pseudo pairs, which are constructed by retrieving the text and image candidates with the highest cosine similarity to each query. We evaluate our method across seven VLMs using two multi-modal QA benchmarks (MMQA and WebQA), where each question requires retrieving either text or image data. Our experimental results show that our method significantly improves retrieval performance, achieving average Recall@20 gains of 64% on MMQA and 28% on WebQA when the query and the target data belong to different modalities. Compared to E5-V, which addresses the modality gap through image captioning, we confirm that our method more effectively bridges the modality gap.

1 Introduction

Information retrieval (IR) plays a key role in a wide range of NLP applications, including web search engines (Kobayashi and Takeda, 2000) and question answering systems (Kolomiyets and Moens, 2011). While traditional approaches primarily focus on retrieving textual information (Robertson and Zaragoza, 2009; Karpukhin et al., 2020), there is a growing interest in retrieving both text and images to provide richer and more informative results (Zhou et al., 2024b).

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), enable both text and image data to be embedded into a shared representation space. Although VLMs enable effective text-to-image retrieval (Radford et al., 2021), it is still challenging to extract relevant information from a database that contains both text and images. Specifically, text items often dominate the top-ranked results even when relevant images exist (Chang et al., 2021; Liu et al., 2023). This issue is attributed to the *modality gap*—a phenomenon in which embeddings from different modalities are mapped to separate regions of the representation space (Liang et al., 2022). Consequently, data that share the same modality as the query tend to receive higher similarity scores, regardless of actual relevance (illustrated in Figure 1).

To address this problem, several approaches have been proposed. Some methods address the modality gap by fine-tuning pre-trained VLMs using paired datasets consisting of queries and their manually labeled corresponding text or image data (Fahim et al., 2024; Eslami and de Melo, 2025). Other methods for converting visual data into text have also been introduced, such as E5-V (Jiang et al., 2024). However, these approaches have shortcomings: collecting human-annotated data is resource-intensive, whereas image captioning would fail to preserve necessary visual information in text.

In this study, we propose a retrieval method that mitigates the impact of modality gap without manually labeled data or image captioning. The key idea is to make similarity scores comparable across modalities by standardizing them using the modality-specific mean and variance. To estimate these statistics, we construct pseudo-positive pairs of unlabeled queries and their most similar texts or images. We then derive modality-specific mean and variance from these pairs, which are used to

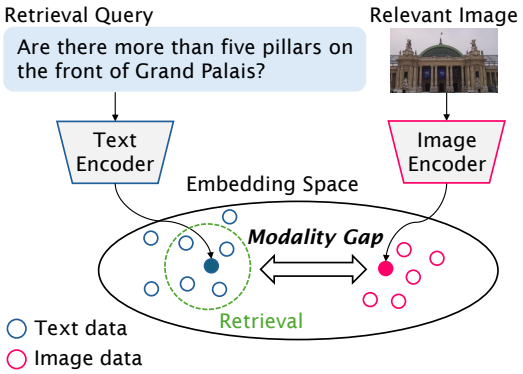


Figure 1: Conceptual overview of the modality gap. Texts and their corresponding images are projected to distant regions of the embedding space.

standardize similarity scores during retrieval.

To evaluate our approach, we conduct experiments on multi-modal question answering benchmarks, i.e., MMQA (Talmor et al., 2021) and WebQA (Chang et al., 2021) with seven pre-trained VLMs. Our method significantly improves retrieval performance when the query and the target data belong to different modalities, achieving average gains of 64% and 28% in Recall@20 on MMQA and WebQA, respectively.

Our main contributions are as follows:

- We propose a similarity standardization approach to mitigate the effect of the modality gap on multi-modal retrieval.
- Our method improves the retrieval performances on two datasets regardless of modalities, compared to E5-V.
- Our method bridges the modality gap without manually labeled datasets, such as pairs of queries and their corresponding examples.

2 Related Work

2.1 Multi-Modal Retrieval

Vision-language models (VLMs) have shown remarkable progress in recent years (Radford et al., 2021; Jia et al., 2021). These models are typically trained using contrastive learning to align images and text in a representation space. Their embeddings can be used for retrieval by computing similarity scores with each item in the database (Karpukhin et al., 2020).

Retrieval tasks involving multiple modalities can be broadly categorized into two settings (Liu et al., 2023). *Cross modality retrieval* refers to settings in which the query and target belong to different modalities, such as text-image or image-text retrieval. In contrast, *multi-modal retrieval* assumes that the retrieval database contains data from multiple modalities—for example, both text and images—and the goal is to find the most relevant item regardless of its modality.

While contrastively trained VLMs perform well in cross modality retrieval tasks (Radford et al., 2021), their performance in multi-modal retrieval remains limited. In particular, when both text and images are present in the retrieval set, these models often retrieve items only from the same modality as the query, and fail to retrieve relevant data from the other modality (Chang et al., 2021; Ross et al., 2024).

This issue is attributed to the modality gap, a clear separation between image and text embeddings of contrastively trained VLMs. This phenomenon was first studied by Liang et al. (2022), who showed that it exists even in randomly initialized models and persists throughout contrastive training. Several causes have been suggested in prior work, including an information imbalance between text and image inputs (Schrodi et al., 2025).

2.2 Bridging the Modality Gap

Some approaches attempt to eliminate the modality gap in VLMs by modifying the contrastive training process. (Fahim et al., 2024) augment CLIP's objective with uniformity and alignment regularizers to enforce balanced embedding distributions and eliminate the modality gap. Schrodi et al. (2025) demonstrated that contrastive learning can mitigate the modality gap when the training data is balanced in information content across modalities. Eslami and de Melo (2025) introduce AlignCLIP, which adds shared parameters between visual and text encoders and an intra-modality separation term to the contrastive loss. While effective, these methods require access to manually paired datasets, which can be expensive or unavailable in real-world scenarios.

Another line of work obtains image embeddings by leveraging image captions (Liu et al., 2023; Zhou et al., 2024a,b). These models achieve strong performance in multi-modal retrieval, but rely heavily on captions. In settings without image descriptions, retrieval quality deteriorates, indicating limited use of visual features.

More recently, methods utilizing the vision-language capabilities of multi-modal large language

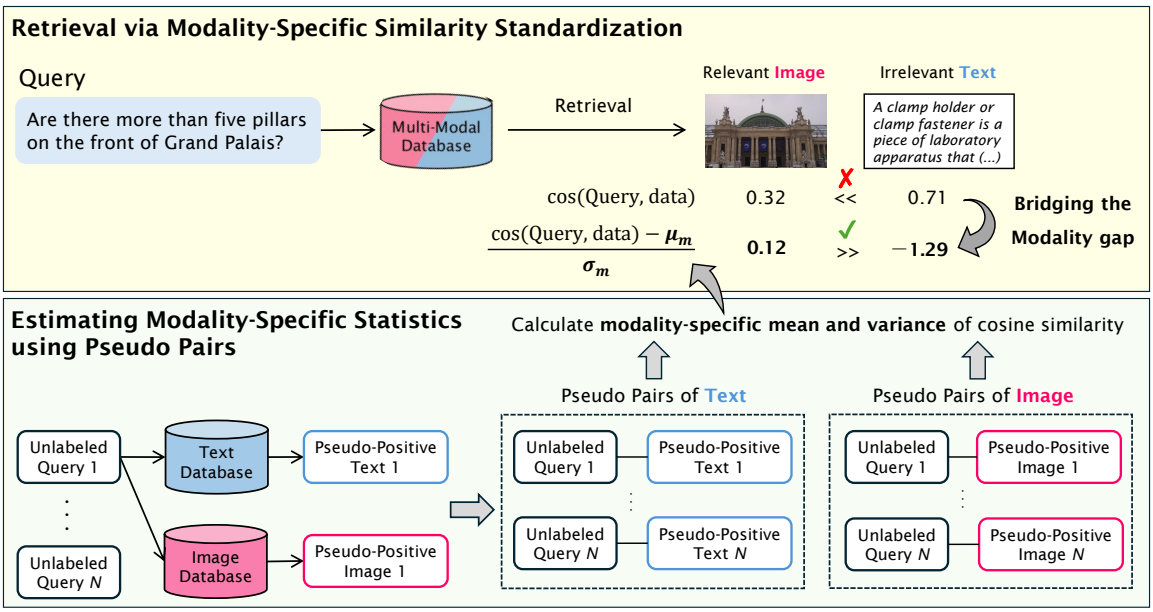


Figure 2: Overview of our proposed method. The modality gap causes irrelevant text to score higher than relevant images. Our approach addresses this issue by standardizing cosine similarity scores based on modality-specific mean and variance calculated from pseudo data.

models (MLLMs) have been explored (Jiang et al., 2024; Zhang et al., 2024b; Lin et al., 2025). For instance, E5-V (Jiang et al., 2024) prompts its backbone MLLM with an image to generate a one-word summary of it. By using the resulting features to obtain image embeddings, E5-V aligns visual inputs with the language space, effectively eliminating the modality gap.

Unlike the existing works that require manually labeled data or image captioning, our method directly adjusts similarity scores across modalities using pseudo-positive examples, eliminating the need for manual supervision.

3 Task Formulation

We work on the task of retrieving relevant data from a multi-modal database that contains both text and images, given a natural language query.

Formally, let q be a textual query and let $\mathcal{D} = \mathcal{D}_{\text{text}} \cup \mathcal{D}_{\text{image}}$ denote the retrieval database, where $\mathcal{D}_{\text{text}}$ and $\mathcal{D}_{\text{image}}$ are sets of textual and visual items respectively. A pre-trained VLM f encodes both the query and each item in the database into the same space. For each candidate $d \in \mathcal{D}$, its relevance to the query can be measured by comparing their embeddings, for example, using cosine similarity: $\cos(f(q), f(d))$.

However, due to the modality gap, similarities differ in scale between text and image modalities. Specifically, a text query tends to assign higher scores to textual candidates than to images, causing relevant images to appear lower in the ranking.

4 Proposed Methods

In this section, we propose a method that mitigates the negative impact of the modality gap without manually labeled data. We first introduce similarity standardization approach as described in Section 4.1. Then, we construct pseudo pairs instead of labeled data, as detailed in Section 4.2.

4.1 Modality-Specific Similarity Standardization

To bridge the modality gap, we propose a similarity standardization approach with modality-specific statistics. We standardize the similarity scores between queries and target information (i.e., positive examples) using their means and variances computed separately for text targets and image targets.

Let $\mathcal{P}_m$ be a set of query-positive pairs where positive example belongs to modality $m \in \{\text{text}, \text{image}\}$. We calculate the mean and variance

of similarities for each modality as:

$$\mu_m = \frac{1}{|\mathcal{P}_m|} \sum_{(q,d_m^+)\in\mathcal{P}_m} \cos(f(q), f(d_m^+)),$$
$$\sigma_m^2 = \frac{1}{|\mathcal{P}_m|} \sum_{(q,d_m^+)\in\mathcal{P}_m} (\cos(f(q), f(d_m^+)) - \mu_m)^2, \quad (1)$$

where each $(q, d_m^+) \in \mathcal{P}_m$ is a query-positive pair.

Using the modality-specific statistics estimated above, we standardize the cosine similarity between a query q and a candidate $d \in \mathcal{D}$ of modality m as:

$$\text{sim}(q, d) = \frac{\cos(f(q), f(d)) - \mu_m}{\sigma_m}. \quad (2)$$

This modality-aware standardization will mitigate the negative impact of the modality gap on similarities between text and image. Note that the statistics μ_m and σ_m^2 are computed from the pre-collected dataset $\mathcal{P}_m$ and remain fixed regardless of the retrieval queries.

4.2 Pseudo Pair Construction

We propose a method for constructing pseudo data that eliminates the need for manually labeled data.

Let $\mathcal{D}_m$ be the subset of the retrieval database corresponding to modality $m \in \{\text{text}, \text{image}\}$, and let $\mathcal{Q}$ denote a set of unlabeled queries. Given a query $q \in \mathcal{Q}$, we extract the most similar item from $\mathcal{D}_m$ for each modality m, and treat it as a pseudo-positive example of modality m:

$$\hat{d}_m^+ = \arg\max_{d\in\mathcal{D}_m} \cos(f(q), f(d)). \quad (3)$$

By repeating this process for all queries in $\mathcal{Q}$, we construct a modality-specific pseudo pair set $\hat{\mathcal{P}}_m$ for each modality m:

$$\hat{\mathcal{P}_m} = \{(q, \hat{d_m^+}) \mid q \in \mathcal{Q}\}. \quad (4)$$

$\hat{\mathcal{P}}_m$ can be used as a substitute for the manually labeled set $\mathcal{P}_m$ in Equations (1). This allows our method to perform modality-specific standardization without relying on any labeled data.

5 Experimental Setup

5.1 Datasets for Evaluation

We evaluate our method on two multi-modal question answering datasets: MultimodalQA (Talmor et al., 2021) and WebQA (Chang et al., 2021). These datasets are widely used benchmarks for the multi-modal retrieval task (Chen et al., 2022; Liu et al., 2023; Zhou et al., 2024a,b). In our experiments, we use questions that require retrieving relevant textual passages (TextQ) or images (ImageQ) in order to answer them. Table 1 shows examples from each dataset, and Table 2 shows the dataset sizes.

(a) MMQA: "*How many colors are on the Mississippi flag?*"

(b) WebQA: "*Are there more than five pillars on the front of Grand Palais?*"

Figure 3: Examples of positive images for ImageQ in MMQA and WebQA shown in Table 1.

MultiModalQA (MMQA) (Talmor et al., 2021) is a benchmark for multi-hop question answering across multiple modalities, including text, images, and tables. It is constructed from Wikipedia tables linked with relevant textual paragraphs and images via shared entities.

WebQA (Chang et al., 2021) is a large-scale open-domain question answering dataset that includes questions paired with corresponding textual passages or images. The data is collected from the open web and Wikipedia. Following Liu et al. (2023) and Zhou et al. (2024b), we construct a retrieval corpus by collecting all images and text passages relevant to all queries in the WebQA dataset.

5.2 Datasets for Pseudo Pair Construction

Pseudo pairs are constructed independently for the MMQA and WebQA datasets. We use queries from the training split of each dataset and sample their pseudo-positive examples from the retrieval source of each dataset as illustrated in Equation 3.

5.3 Metrics

We evaluate our methods using Recall@k, MRR@k, and NDCG@k. All metrics are primarily measured at $k = 20$. For Recall, we additionally compute values at k=1, 5, and 100 to examine the effect of varying k.

5.4 Models

We apply our method to seven pre-trained VLMs to demonstrate its robust effectiveness. To assess models expected to exhibit a modality gap due to contrastive training, we include CLIP (Rad-

Dataset	Type	Question	Positive Example
MMQA	TextQ	When did "Harry Potter and the Sorcerer's Stone" movie come out?	Harry Potter and the Philosopher's Stone (released in the United States as Harry Potter and the Sorcerer's Stone) is a 2001 fantasy film directed by Chris Columbus and distributed by Warner Bros.
	ImageQ	How many colors are on the Mississippi flag?	Refer to Figure 3a.
WebQA	TextQ	What part of the human body does the nerves in the frontalis muscle serve and the occipitofrontalis muscle serve?	The frontalis muscle is supplied by the facial nerve and receives blood from the supraorbital and supratrochlear arteries. In humans, the occipitofrontalis only serves for facial expressions.
	ImageQ	Are there more than five pillars on the front of Grand Palais?	Refer to Figure 3b.

Table 1: Examples from MMQA and WebQA datasets. Each dataset includes two types of questions: TextQ and ImageQ, which refer to questions that require retrieving text and images to answer, respectively.

# of dataset	Source		Query	
	text	image	TextQ	ImageQ
MMQA	218K	57K	6.7K/721	1.9K/230
WebQA	787K	389K	15K/2.4K	16K/2.5K

Table 2: Numbers of retrieval candidates and queries in MMQA and WebQA. The numbers of queries are listed as training/test. Validation data is not used in our experiments.

ford et al., 2021) (ViT-B/32 and ViT-L/14), LongCLIP (Zhang et al., 2024a) (base and large), and BLIP (Li et al., 2022). We also include Cohere Embed 3 English (Ross et al., 2024), a high-performance VLM accessible via API. In addition, we evaluate E5-V (Jiang et al., 2024), which integrates image captioning via a MLLM. While E5-V is designed to mitigate the modality gap, we apply similarity standardization to examine whether our method can further improve its performance. The computational resources are provided in Appendix A.

5.5 Evaluation Conditions

All VLMs are evaluated under the following three configurations.
(i) **Cos**: Cosine similarities are simply used for retrieval.
(ii) **Std**: Cosine similarities are standardized by our method with manually labeled data, which is taken from the training split of each dataset.
(iii) **Ours**: Cosine similarities are standardized by our method with our pseudo pairs.

6 Results and Discussions

6.1 Overall Results

Table 3 summarizes the overall retrieval performance across seven VLMs on MMQA and WebQA datasets. When Cos was applied, four of the CLIP-based models and BLIP retrieved almost no relevant results, resulting in near-zero scores on all evaluation metrics on ImageQ. This suggests that the modality gap causes irrelevant text passages to be ranked higher than relevant images, hindering accurate retrieval.

In contrast, applying our method to these models significantly improved the performances, achieving average gains of 64% and 28% in Recall@20 for MMQA ImageQ and WebQA ImageQ, respectively, thereby confirming its effectiveness in bridging the modality gap.

Notably, all models with our method outperformed E5-V on ImageQ. These results highlight the advantage of processing images without any loss of information, different from the existing works with image captioning or verbalization. Although a slight performance degradation was observed on TextQ, the overall trade-off is favorable with notable gains on ImageQ.

Cohere Embed 3 and E5-V achieved high performance on TextQ, with approximately 80% in Recall@20. On ImageQ, they retained a certain level of performance without our method, achieving Recall@20 ranging from 40-50% on MMQA and 10-20% on WebQA. For E5-V, this can be attributed to its strong capability for understanding textual information through its MLLM backbone, as well as its architecture that converts images into text. While the architecture and training details

of Cohere Embed 3 are not publicly available, its performance suggests that it may adopt a similar architecture or training process to models like E5-V. When our standardization is applied to these models, further improvements are observed on ImageQ; however, it also results in a large drop in TextQ accuracy compared to CLIP-based models and BLIP. This indicates that the benefit of our method is limited when the modality gap is already small.

6.2 Severe Impact of the Modality Gap

To examine how the modality gap affects retrieval performance, we evaluated Recall at various cutoff values of retrieval on ImageQ. Table 4 reports Recall@{1, 5, 20, 100} for each model and dataset.

For Cos, increasing the number of retrieved candidates had almost no effect—Recall@k remained around zero even with $k = 100$. This result clearly indicates that the modality gap severely degrades retrieval performance on ImageQ.

In contrast, our method yields substantial improvements in Recall@k across all tested values of k, demonstrating its effectiveness in bridging the modality gap.

6.3 Pseudo Pairs vs. Manually Labeled Pairs

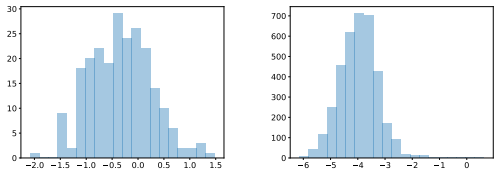
To assess how pseudo pairs affect retrieval, we compared retrieval performances of the Std method and our method. Table 3 shows that the results of our method were equal or higher than those of the Std method. This result demonstrates that pseudo pairs can serve as an effective substitute for manually labeled pairs.

7 Analysis of the Modality Gap

7.1 The Effect of Standardization

To investigate how our method reduced the negative impact of the modality gap, we analyze the distribution of standardized similarity scores on ImageQ. For each ImageQ, we compute the difference between the average standardized similarity scores for image and text candidates in the retrieval database (image mean minus text mean). The distributions on MMQA and WebQA are shown in Figure 4, focusing on CLIP (ViT-B/32) as a representative model that exhibits a clear modality gap.

In MMQA, the distribution is centered slightly below zero, indicating that text scores remain somewhat higher than image scores on average, even after the standardization. In WebQA, the distribution is concentrated mostly on the negative side (around -4), indicating that text candidates are consistently scored higher than images. From these results, we confirm that our method does not fully eliminate the modality gap.



(a) MMQA (b) WebQA

Figure 4: Distributions of the difference between the average standardized similarity scores of image and text candidates across ImageQ queries in the training split of MMQA and WebQA, where the difference is computed as image minus text.

Nevertheless, retrieval performance improves significantly as shown in Section 6. We attribute this to differences in the shape of the cosine similarity score distributions across modalities. Table 5 shows the skewness values in the distributions of similarity scores. CLIP-based models consistently produced more positively skewed similarity distributions for image candidates compared to text candidates. This suggests that some images receive totally higher similarity scores than others in the image database. Such outliers—which often include the correct images—were amplified by our method, allowing them to receive a higher standardized score than most text candidates.

We hypothesized that the skewness in the image similarity distribution stems from the training objective of CLIP-based models These models learn to align images with their paired text, but they are not explicitly trained to capture similarities between texts or between images themselves. As a result, these models yield high similarities to a few image candidates, resulting in a long-tailed distribution. This skewed distribution might align well with our standardization approach, as it amplifies the scores of outliers which often include relevant images.

7.2 Modality Gap in VLMs

We analyze the modality gap in VLMs by investigating both the structure of the embedding space and the distribution of similarity scores.

Following Liang et al. (2022), we apply singular value decomposition (SVD) to project the embeddings of ImageQ queries and their positive examples into a two-dimensional space for visualiza-

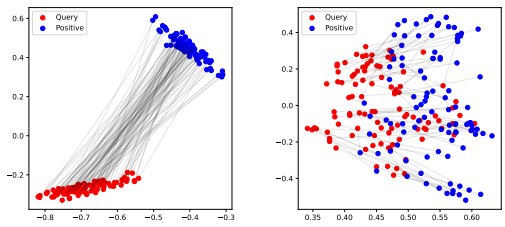
Model	Method	MMQA						WebQA					
		TextQ			ImageQ			TextQ			ImageQ		
		Recall	MRR	NDCG	Recall	MRR	NDCG	Recall	MRR	NDCG	Recall	MRR	NDCG
CLIP (ViT-B/32)	Cos	31.90	26.62	23.90	0.00	0.00	0.00	28.89	21.14	18.89	0.00	0.00	0.00
	Std	31.55	25.78	23.25	52.61	36.65	40.32	23.96	16.38	15.01	32.82	15.20	18.02
	Ours	27.46	18.44	17.88	66.09	45.03	49.86	27.14	19.07	17.29	28.14	13.79	15.98
CLIP (ViT-L/14)	Cos	35.51	28.63	25.86	1.30	0.41	0.62	32.60	24.05	21.45	0.04	0.00	0.01
	Std	35.37	27.60	25.16	62.17	43.32	47.70	28.84	19.37	17.91	43.55	22.54	25.76
	Ours	31.28	21.48	20.54	76.52	58.88	63.05	31.27	21.44	19.69	37.10	20.37	22.75
Long-CLIP-B	Cos	58.67	45.11	43.02	0.00	0.00	0.00	43.93	30.92	28.56	0.00	0.00	0.00
	Std	54.65	40.73	38.76	66.09	47.67	51.94	34.94	23.79	22.05	33.01	14.92	17.98
	Ours	53.33	35.48	35.04	66.96	50.72	54.51	40.44	27.72	25.79	28.59	13.43	15.97
Long-CLIP-L	Cos	63.04	45.56	44.32	0.43	0.11	0.19	45.18	30.36	28.54	0.00	0.00	0.00
	Std	58.39	41.94	40.70	71.74	49.38	54.52	35.66	23.57	22.14	39.84	20.22	23.36
	Ours	57.07	38.60	38.19	73.91	54.04	58.63	41.46	27.14	25.69	35.34	18.38	21.09
BLIP	Cos	41.75	30.20	28.64	0.00	0.00	0.00	37.15	27.07	24.23	0.00	0.00	0.00
	Std	40.92	28.58	27.42	39.57	23.97	27.54	24.00	14.75	14.04	17.62	8.24	9.73
	Ours	36.75	23.33	23.31	43.48	27.45	31.15	31.40	20.71	19.23	14.04	6.35	7.62
Cohere Embed 3	Cos	87.17	78.81	74.72	50.43	20.79	27.61	76.52	59.19	55.86	20.43	8.00	10.16
	Std	72.19	66.33	60.63	52.17	27.24	32.92	54.78	41.69	38.19	27.42	12.36	14.83
	Ours	73.99	63.25	59.20	52.17	28.17	33.61	69.23	52.67	49.36	25.39	11.48	13.73
E5-V	Cos	84.88	66.67	67.20	38.70	17.34	22.06	74.37	54.88	52.27	11.89	5.19	6.37
	Std	80.79	63.33	63.56	41.74	21.33	25.91	48.61	35.04	33.11	21.05	9.75	11.50
	Ours	70.39	53.15	53.12	41.74	21.55	26.09	65.73	48.76	46.11	18.78	8.87	8.87

Table 3: Overall retrieval results on MMQA and WebQA. Recall@20, MRR@20, and NDCG@20 are reported. Cos uses cosine similarity as the retrieval score. Std-L and Std-P apply similarity standardization using modality-specific mean and variance estimated from labeled and pseudo pairs, respectively.

tion. Figure 5 shows the results for CLIP (ViT-B/32) and E5-V. The visualizations of other models and datasets are shown in Appendix D. CLIP exhibits a clear separation between textual queries and positive image items in the embedding space. In contrast, E5-V shows a much smaller gap, suggesting that modality conversion reduces representational disparity between text and images.

We then analyze the cosine similarity scores between queries in the training split of MMQA and their positive examples (either text or image) for CLIP (ViT-B/32) and E5-V. Figure 6 presents the distributions of these scores, separated by the modality of the positive examples. The distributions of other models are shown in Appendix E. As expected, CLIP assigns significantly higher similarities to text examples. E5-V reduces this gap to some extent, but a consistent score difference remains: image positives still tend to receive lower similarity scores than text counterparts.

These results indicate that image captioning reduces modality differences, but does not fully avoid the gap of VLMs. One possible reason is that converting images into textual representations leads to loss of visual information necessary for questions that are difficult to express in language, such as the spatial relationships between objects and the background color. This missing information reduces similarities between queries and relevant candidates compared to text data. Our method avoids this shortcoming. By directly processing image features without converting them into text, our method outperformed E5-V in ImageQ.



(a) CLIP (ViT-B/32) (b) E5-V

Figure 5: 2D visualizations of the embeddings of ImageQ queries in the MMQA training split (blue dots) and their corresponding images (red dots) using SVD. Figures 5a and 5b show the results of CLIP (ViT-B/32) and E5-V, respectively.

8 Conclusion

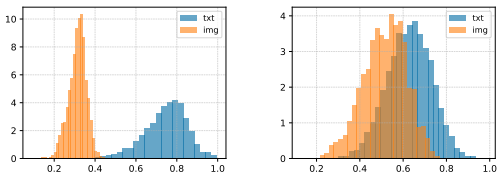
We presented a method for improving multi-modal retrieval by bridging the modality gap without human-created data. Our approach standardizes similarity scores in a modality-specific manner, making them more comparable across modalities. Importantly, it does not require any labeled data or image captions, as it relies on pseudo-positive examples derived from unlabeled queries. Through

Model	Method	MMQA				WebQA			
		1	5	20	100	1	5	20	100
CLIP (ViT-B/32)	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	37.39	53.48	66.09	72.17	8.16	16.11	28.14	44.78
CLIP (ViT-L/14)	Cos	0.00	0.87	1.30	3.04	0.00	0.00	0.04	0.06
	Ours	50.87	69.57	76.52	81.74	12.90	24.39	37.10	54.76
Long-CLIP-B	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	43.91	60.43	66.96	76.96	7.89	16.21	28.59	45.46
Long-CLIP-L	Cos	0.00	0.43	0.43	0.87	0.00	0.00	0.00	0.00
	Ours	46.09	63.48	73.91	80.87	11.95	21.39	35.34	52.77
BLIP	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	21.30	35.22	43.48	56.09	3.78	7.39	14.04	26.66
Cohere Embed 3	Cos	10.00	34.78	50.43	64.78	4.38	9.14	20.43	40.60
	Ours	20.43	38.26	52.17	65.65	6.41	13.52	25.39	43.83
E5-V	Cos	12.17	23.91	38.70	59.57	2.95	6.35	11.89	26.52
	Ours	16.09	28.26	41.74	63.04	5.10	10.49	18.78	36.80

Table 4: Results of Recall@k ($k = \{1, 5, 20, 100\}$) for each model on ImageQ queries in MMQA and WebQA datasets.

Model	MMQA		WebQA	
	Text	Image	Text	Image
CLIP (ViT-B/32)	−0.81	0.21	−1.05	0.45
CLIP (ViT-L/14)	−0.41	0.31	−0.51	0.33
Long-CLIP-B	−1.40	0.35	−1.33	0.59
Long-CLIP-L	−1.88	0.47	−1.30	0.72
BLIP	0.32	0.37	0.59	0.45
Cohere Embed 3	0.26	0.16	0.54	0.25
E5-V	0.81	0.90	0.91	0.76

Table 5: Average skewnesses of cosine similarity distributions for ImageQ queries in the training split of MMQA and WebQA. Each skewness is computed between a query and all candidates in the text or image database, then averaged across all queries per modality.



(a) CLIP (ViT-B/32) (b) E5-V

Figure 6: Distributions of cosine similarity scores between a query and its corresponding positive example (either text or image). The distributions are separated by the modality of the positive example. Figures 6a and 6b show the results of CLIP (ViT-B/32) and E5-V, respectively.

experiments on two multi-modal QA datasets and seven vision-language models, we demonstrated that our method consistently improves image retrieval performance, particularly in scenarios where existing models struggle due to the modality gap. Furthermore, we showed that pseudo-positive examples are sufficient for estimating modality-specific statistics, achieving performance on par with manually labeled data. Our findings highlight the importance of preserving modality-specific information and calibrating similarity scores, rather than relying solely on modality conversion.

Limitations

Our method computes modality-specific similarity statistics from pre-collected datasets and uses them to standardize all similarity scores across modalities. However, this approach assumes that similarity distributions remain stable over time. In real-world systems, new data is constantly being added to databases. Due to new content, these precomputed statistics may become obsolete, leading to suboptimal standardization. Future work should focus on developing mechanisms to dynamically update these statistics.

References

Yinghsan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. WebQA: Multihop and Multimodal QA.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural*

Language Processing, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Sedigheh Eslami and Gerard de Melo. 2025. Mitigate the gap: Improving cross-modal alignment in CLIP. In *The Thirteenth International Conference on Learning Representations*.

Abrar Fahim, Alex Murphy, and Alona Fyshe. 2024. It's not a modality gap: Characterizing and addressing the contrastive gap. *Preprint*, arXiv:2405.18570.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *Preprint*, arXiv:2407.12580.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)*, 32(2):144–173.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal LLMS. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proceedings of ICLR*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Luke Ross, Nils Reimers, Leila Chan Currie, and Elliott Choi. 2024. Introducing multimodal embed 3: Powering ai search. Cohere Blog. Blog post.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu.

2024b. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624, Bangkok, Thailand. Association for Computational Linguistics.

A Computational Resources

We used two NVIDIA Quadro RTX 6000 GPUs for generating embeddings with E5-V, while only one GPU was used for all other pre-trained VLMs. All retrieval and evaluation experiments were conducted using Faiss (Douze et al., 2024) on CPU only.

B Model List

We evaluated seven pre-trained VLMs in our experiments. Six of them are publicly available on Hugging Face and were accessed as downloadable checkpoints:

- https://huggingface.co/openai/clip-vit-base-patch32
- https://huggingface.co/openai/clip-vit-large-patch14
- https://huggingface.co/BeichenZhang/LongCLIP-B
- https://huggingface.co/BeichenZhang/LongCLIP-L
- https://huggingface.co/Salesforce/blip-itm-base-coco
- https://huggingface.co/royokong/e5-v

We used the Cohere Embed 3 English model (cohere.embed-english-v3) via Amazon Bedrock API in the us-west-2 region.

C Modality-Specific Mean and Variance

Table 6 lists the modality-specific mean and standard deviation for similarity standardization that were used for standardization in our experiments.

D 2D Visualizations of Embeddings

Figures 7–13 illustrate 2D visualizations of embeddings of textual queries (from the training sets of MMQA and WebQA) and their positive examples using singular value decomposition[1].

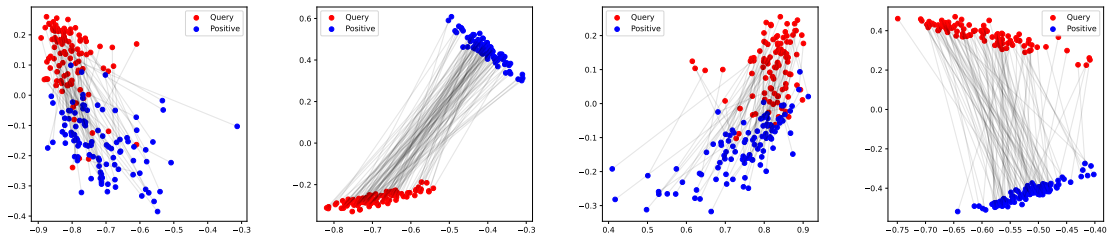
[1]Our visualization code is adapted from https://github.com/Weixin-Liang/Modality-Gap/blob/main/Figure_1_Modality_Gap/visualize.ipynb

E Distributions of Cosine Similarity Scores between Positive Pairs across Modalities

Figure 14 presents the distributions of cosine similarity scores between textual queries (from the training sets of MMQA and WebQA) and their positive examples, separated by the modality of positive examples.

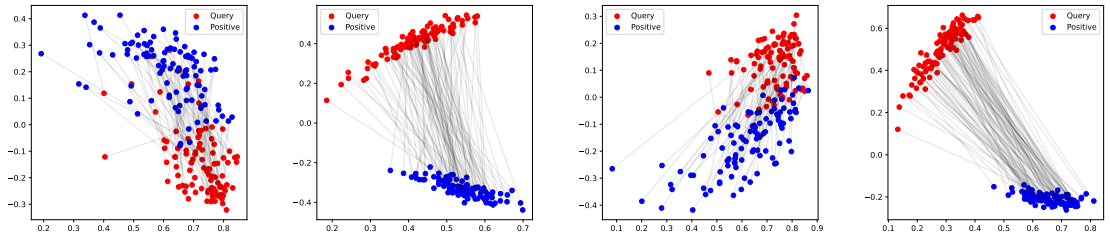
Model	Method	MMQA				WebQA			
		Text		Image		Text		Image	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std
CLIP (ViT-B/32)	Std	0.744	0.105	0.314	0.043	0.789	0.093	0.304	0.035
	Ours	0.841	0.058	0.315	0.023	0.833	0.063	0.335	0.019
CLIP (ViT-L/14)	Std	0.642	0.136	0.280	0.049	0.700	0.122	0.269	0.040
	Ours	0.755	0.088	0.271	0.029	0.749	0.093	0.297	0.023
Long-CLIP-B	Std	0.879	0.050	0.315	0.031	0.895	0.040	0.307	0.024
	Ours	0.898	0.043	0.311	0.017	0.901	0.037	0.324	0.016
Long-CLIP-L	Std	0.828	0.068	0.279	0.048	0.856	0.057	0.258	0.037
	Ours	0.845	0.073	0.264	0.029	0.860	0.059	0.277	0.026
BLIP	Std	0.700	0.116	0.438	0.072	0.724	0.100	0.418	0.059
	Ours	0.791	0.070	0.460	0.038	0.806	0.058	0.489	0.034
Cohere Embed 3	Std	0.629	0.121	0.508	0.082	0.581	0.114	0.490	0.066
	Ours	0.660	0.105	0.512	0.047	0.615	0.082	0.541	0.044
E5-V	Std	0.628	0.102	0.514	0.099	0.635	0.105	0.467	0.084
	Ours	0.649	0.095	0.469	0.085	0.640	0.093	0.534	0.073

Table 6: Modality-specific mean and standard deviation used for standardization during evaluation on MMQA and WebQA datasets. Values are computed separately for text and image modalities, either from labeled or pseudo pairs.



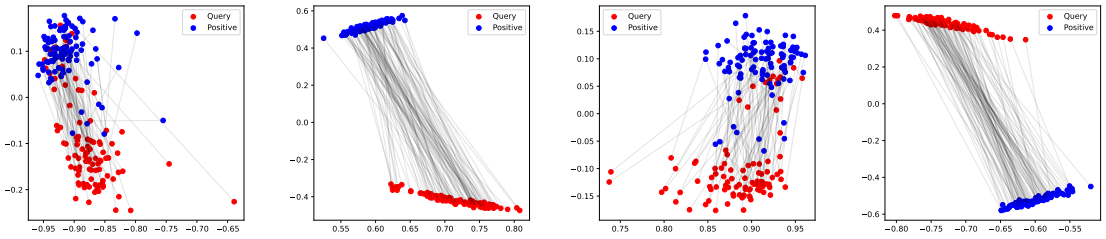
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 7: 2D visualizations of embeddings from CLIP (ViT-B/32).



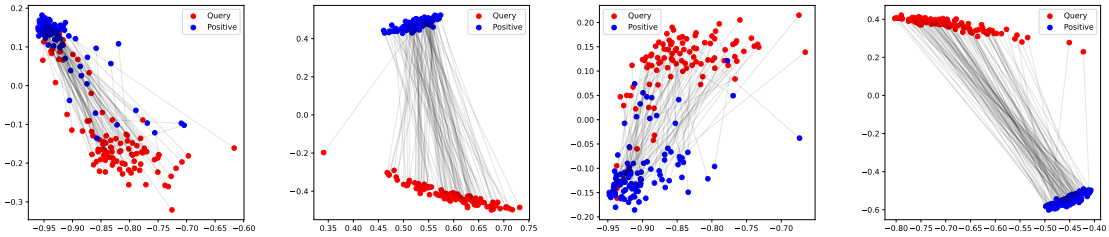
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 8: 2D visualizations of embeddings from CLIP (ViT-L/14).



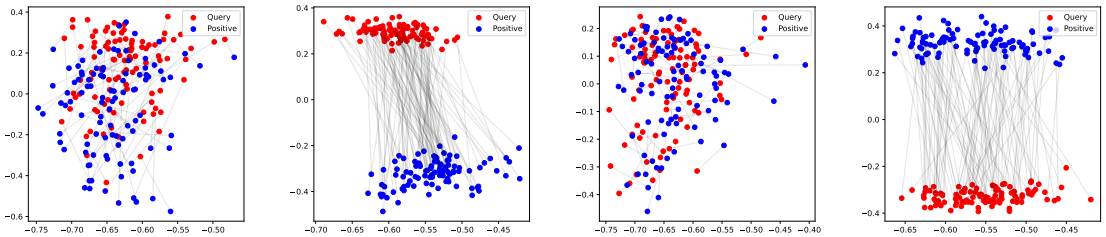
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 9: 2D visualizations of embeddings from Long-CLIP-B.



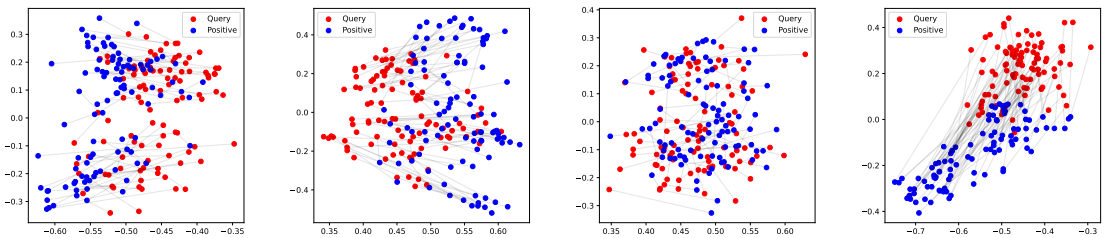
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 10: 2D visualizations of embeddings from Long-CLIP-L.



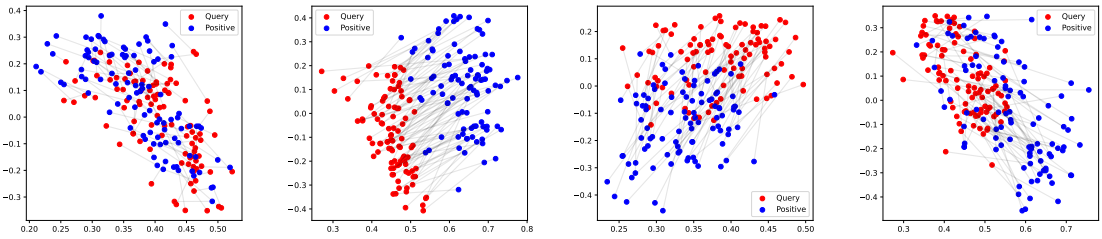
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 11: 2D visualizations of embeddings from BLIP.



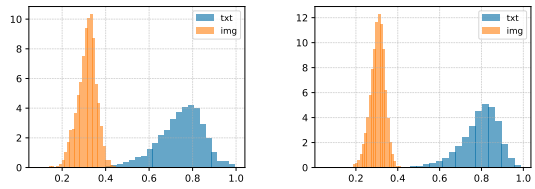
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 12: 2D visualizations of embeddings from E5-V.



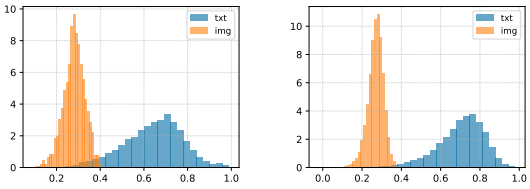
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 13: 2D visualizations of embeddings from Cohere Embed 3 English.



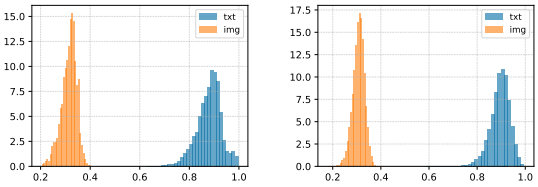
(a) MMQA (b) WebQA

(A) CLIP (ViT-B/32)



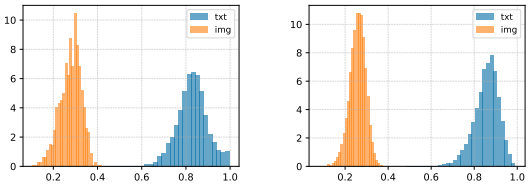
(c) MMQA (d) WebQA

(B) CLIP (ViT-L/14)



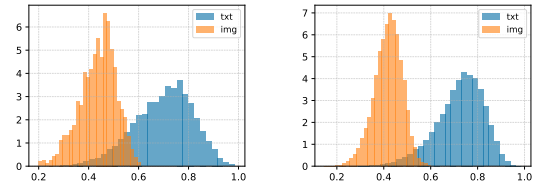
(e) MMQA (f) WebQA

(C) Long-CLIP-B



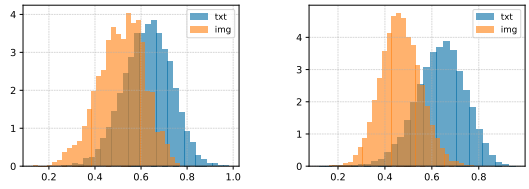
(g) MMQA (h) WebQA

(D) Long-CLIP-L



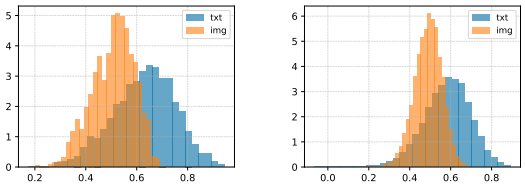
(i) MMQA (j) WebQA

(E) BLIP



(k) MMQA (l) WebQA

(F) E5-V



(m) MMQA (n) WebQA

(G) Cohere Embed 3 English

Figure 14: Distributions of cosine similarity scores between textual queries in the training split of each dataset and their corresponding examples (either text or image).

PRISM: A Pedagogical Multi-Agent for Structured Group Learning

Van-Khanh Tran[1*], Van-Khai Dang[2#], Duc-Huy Nguyen[2#]

[1]Institute of Applied Science and Technology,
Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
[2]Institute of Artificial Intelligence,
VNU University of Engineering and Technology, Hanoi, Vietnam

Abstract

Current AI tutoring systems primarily focus on one-on-one interactions, missing the collaborative dynamics essential for developing communication and social learning skills. We introduce **PRISM** (Proactive Role-based Intelligent Scaffolded Multi-agent), a novel framework that enables natural multi-agent collaboration in educational settings through autonomous turn-taking mechanisms. PRISM coordinates specialized AI agents with distinct pedagogical roles within a structured four-stage problem-solving framework based on Pólya's methodology. Our key innovation is a proactive self-selection mechanism where agents autonomously determine participation through internal reasoning and evaluative scoring, replacing traditional manager-controlled turn allocation. The performance of the PRISM system was evaluated in two distinct experimental settings focused on high school mathematics. The initial evaluation involved a simulation benchmark that measured PRISM against a next-speaker prediction baseline. Assessed via LLM-as-a-judge metrics, PRISM obtained a 62.3% win rate over the baseline. A subsequent real-time study of human-agent interaction, analyzed using Bales' Interaction Process Analysis (IPA), provided further evidence of efficacy, demonstrating significant improvements in group coordination and developmental outcomes for learners. These results indicate the considerable potential of PRISM as a scaffold for collaborative learning within structured pedagogical environments. Our framework advances multi-agent educational AI by providing measurable learning outcomes, natural interaction patterns, and scalable collaborative learning environments that preserve the social benefits of traditional classroom settings.

*Correspondence: tvkhanh@ictu.edu.vn
#These authors contributed equally to this work.

1 Introduction

In recent years, educational technologies have evolved from rule-based Intelligent Tutoring Systems (ITS) to powerful large language models (LLMs) capable of generating context-aware, human-like dialogue. This shift marks a significant pedagogical opportunity: *Virtual Classroom Simulation*. The user engages with this virtual class in real-time, participating in group discussions to solve problems.

To simulate collaborative learning in a structured and pedagogically meaningful way, we introduce PRISM, a multi-agent system powered by a large language model. The system supports a staged dialogue flow where agents interact with the human student. Each agent assumes a distinct classroom role. At each stage, a Stage Manager guides the flow of conversation, ensuring that the problem-solving process unfolds coherently.

We evaluate PRISM through experiments involving Vietnamese high-school students working on mathematical modeling tasks. Results show that the system improves group coordination, diversifies student-agent interaction, and enhances the depth of problem understanding. Our contributions include:

- *A pedagogically-motivated, stage-based dialogue management framework* that enforces structured collaborative phases aligned with learning objectives.
- *A role-driven multi-agent architecture* in which each agent embodies a distinct instructional persona to diversify support.
- *A proactive, self-selecting turn-taking mechanism* enabling agents to autonomously decide when to *speak* based on internal reasoning and conversational context.
- *Comprehensive empirical validation*, including both simulation-based benchmark comparisons and a human–agent user study, demonstrating

significant gains in pedagogical alignment and learner engagement.

2 Related Work

2.1 LLMs for Education

The release of ChatGPT in 2022 introduced a new era in education, shifting from traditional NLP to powerful transformer-based LLMs. Today, these models are widely accessible, enabling automated content creation, real-time feedback and grading at scale, and truly personalized learning experiences (Wang et al., 2024). LLMs can role-play historical figures or conversational partners to foster immersive, engaging lessons (Zhu et al., 2025). Researchers even use LLMs to simulate student behavior, comparing their error rates on multiple-choice questions to those of real learners, to generate high-quality assessments (Liu et al., 2025a).

2.2 One-to-one Tutoring

One-to-one tutoring using AI systems, especially those powered by LLMs, leverages various pedagogical strategies to enhance learning outcomes (Gousopoulos, 2024; Razafinirina et al., 2024). While one-to-one tutoring offers personalized attention, it faces challenges in simulating the full spectrum of classroom interactions. One-to-one settings often miss peer learning opportunities, which are crucial for social development and collaborative skills. In contrast, traditional classrooms foster peer interactions that enhance learning through discussion and shared problem-solving. These limitations highlight the need for a more comprehensive approach to simulate realistic learning experiences.

2.3 Virtual Classroom – Collaborative Learning

Multi-agent Systems (MAS). In a virtual classroom context, agents can be designed with various roles, such as classmates or teachers, collaborating with real students toward shared learning goals. MAS based on Large Language Models (LLMs) has emerged as a potential solution to this challenge, thanks to their capabilities in reasoning, decision-making, and flexible coordination among agents.

Turn-takings in Multi-Party Conversations. Studies such as SimClass (Zhang et al., 2024) and MathVC (Yue et al., 2025) have proposed Next-Speaker Prediction, an approach to managing turn-taking. This method is based on the history and role descriptions of agents to select the most suitable agent to talk to. However, this approach leaves agents in a passive position when they are selected by another manager agent. In reality, when people talk to each other, they will think independently before speaking. Therefore, a more comprehensive solution is needed to simulate this multi-participant conversation to increase the naturalness of communication.

3 Methodology

3.1 Overview

This study aims to design AI agents that can collaborate with human students in solving mathematical problems while simultaneously enhancing learning engagement. The proposed system employs a multi-agent architecture where each agent exhibits distinct roles and behaviors, allowing for diversified perspectives and pedagogically meaningful interactions. The overall goal of this work is to shift from traditional one-on-one tutoring models toward dynamic, group-based learning enhanced by autonomous agents.

To fulfill these goals, the system incorporates three key design requirements:

- *Context Awareness:* Agents need to be aware of the environment (conversation, participants) to enable realistic collaboration throughout the various stages.
- *Turn-taking Autonomy:* Agents should possess full autonomy in deciding when to act, yielding more natural, without relying on fixed sequences.
- *Customizability:* The system should support configurable roles, allowing adaptive and engaging user experiences.

To implement these design principles, we construct a three-module architecture based on an event-driven framework (see Figure 1).

3.2 Event-Driven Architecture

In traditional one-to-one chatbot systems, an agent's response is triggered by a new message from the user. However, when multiple agents operate simultaneously in a shared dialogue space, a more sophisticated and flexible mechanism is required to govern agent participation. To address this, we adopt an event-driven architecture that enables agents to respond dynamically based on contextual cues in the conversation environment.

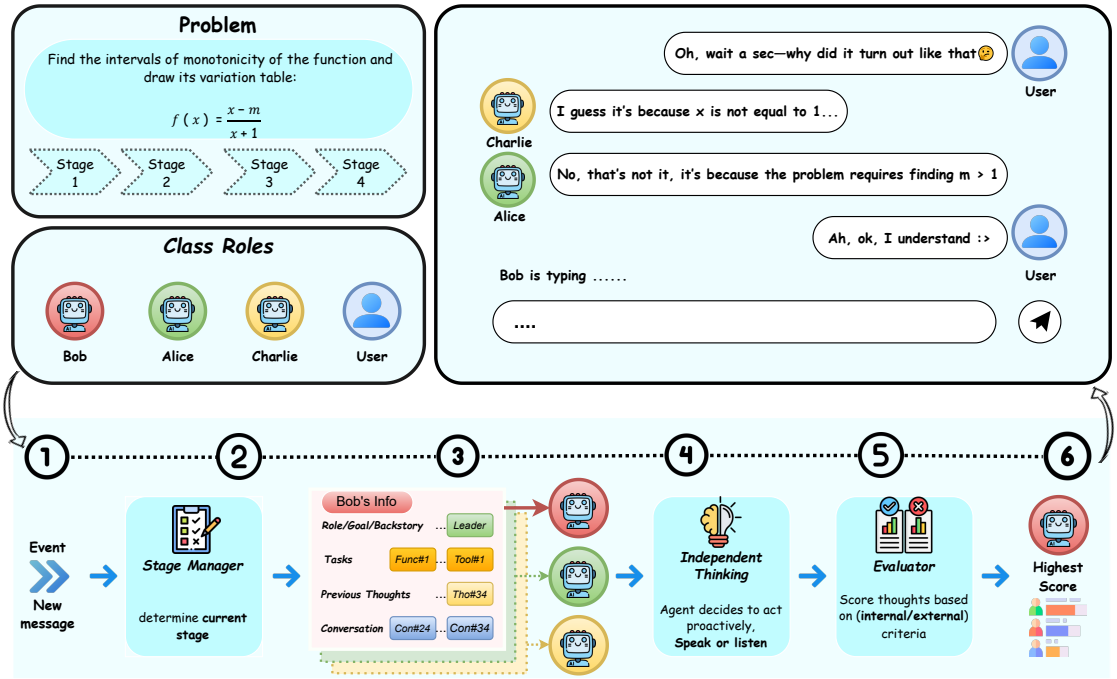


Figure 1: System Architecture Overview of PRISM, showing how multi-agent collaboration is managed through an event-driven pipeline. Upon a new event (like new-message) (1), the Stage Manager Agent determines the current stage (2), providing information to classmate agents. Based on information about roles, conversation history, inner thoughts (3), classmate agents create inner thoughts in parallel and independently (4); and undergo a self-selection process based on thought evaluation scores (5) to determine the next speaker agent (6). The selected agent will then make the next utterance based on the thought just generated.

3.2.1 Environment context

The system environment comprises the complete chat history, the current instructional stage, the list of participants, and temporal elements such as the timing between messages.

3.2.2 Events as interaction triggers

Just as humans respond to spoken words, gestures, or moments of silence in conversation, AI agents are designed to react to discrete events within the system. In this implementation, we define two primary categories of events:

- *New Message:* Triggered whenever any participant sends a message.
- *Silence:* Triggered when no participant sends a message for a predetermined duration (e.g., 10 seconds). This allows agents to take initiative during moments of inactivity unless the dialogue session has concluded.

3.2.3 Shared event timeline

Events are appended to and appear in a shared timeline, providing a single sequence of activities that all agents reference. This ensures that their behaviors and interactions remain consistent and synchronized.

3.3 Stage Module

Pedagogical Approaches. Collaborative problem-solving is most effective when structured into clear stages with defined tasks and shared goals, and it tends to be more effective than simply having the tutor give direct answers to students. This approach can enhance student engagement and positively influence learning outcomes.

To operationalize this in a pedagogical framework, we draw from George Pólya's classic four-step model in *How to Solve It* (Pólya, 1945):

- Stage 1 – Understanding the Problem
- Stage 2 – Devising a Plan
- Stage 3 – Carrying Out the Plan
- Stage 4 – Looking Back

Our system follows a four-stage approach as the backbone of the instructional flow, during which students engage in collaborative discussions to achieve the specific objectives of each stage.

Stage Manager Agent. To create realistic conversations, a collaboration stage manager agent is responsible for continuously monitoring predefined criteria specific to each stage. This agent dynamically determines when the objectives of the current stage have been sufficiently met. Each stage is designed with its own set of tasks, carefully crafted to align with the stage's goals, ensuring that the dialogue progresses logically and purposefully. To avoid agents directly stating the solution or discussing the wrong order of a task, all tasks will be marked as complete or incomplete. The stage manager uses the Chain of Thought (CoT) prompt to analyze the situation and decide to update the status, thereby ensuring the simulation remains coherent and goal-oriented throughout its progression.

3.4 Role-Based Agentization Module

Classroom interaction behaviors can be categorized based on widely accepted pedagogical principles (Schwanke, 1981), like: Teaching and Initiation (TI), In-depth Discussion (ID), Emotional Companionship (EC), and Classroom Management (CM). Ensuring diversity and comprehensive coverage of these agents in the classroom is essential.

Design. This work draws on agentic design principles inspired by the CrewAI platform (Moura, 2025), which supports the creation of specialized AI personas capable of effective collaboration. The core principles of effective agent design are:

Role-Goal-Backstory Framework.

- *Role*: Defines an agent's specialized role and expertise, aligned with real-world professional knowledge.
- *Goal*: guides the agent's actions and informs its decision-making process. It should be explicitly stated, outcome-oriented.
- *Backstory*: Adds contextual depth by defining the agent's expertise, style, and interests in line with its role and goals.

Crafting Effective Tasks.

- *Task Description*: The description of tasks, functions, or tools focuses on what to do and how to do it.
- *Expected Output*: The expected output should define what the final result should look like.

3.5 Turn-Taking Module

Challenges of Turn-taking in Multi-party Dialogue. In multi-agent educational dialogues, deciding who "speaks" next is a fundamental challenge. Unlike one-on-one chatbot systems, multi-party conversations demand agents to make more autonomous and context-aware decisions about when to speak, what to say, and whether to remain silent. Moreover, the next speaker in a multi-party conversation may be explicitly selected (e.g., mentioned directly in a prior message, such as "Hey Charlie!"); if not, any participant who finds it relevant may take the turn, or the current speaker may continue. Such flexibility makes turn-taking particularly challenging for AI agents.

Limitations of Next-Speaker Prediction. One common method is next-speaker prediction, where a manager agent selects the next speaker based on dialogue history and stage context. This approach (as used in SimClass (Zhang et al., 2024)) simplifies management but reduces agent autonomy. Agents act only when selected, limiting their ability to reflect internal reasoning or motivation. Furthermore, these systems are typically based on static agent profiles, which fail to reflect the evolving nature of real human behavior over time (Nonomura and Mori, 2024).

Proactive Turn-taking via Self-Selection. To address this, we adopt a proactive turn-taking mechanism inspired by how humans participate in conversation. After every conversational event (e.g., a new message or a pause), each agent privately generates an internal thought, deciding whether to speak or remain silent, based on the preceding dialogue, their designated role, and their internal memory (previous thoughts).

These thoughts are then passed to a dedicated agent called the Evaluator, who performs a scoring process on each submitted thought. The evaluation considers both internal and external criteria (Liu et al., 2025b):

- *Internal:* "*Relevance*" (Agents contributed most when discussions matched their knowledge, roles, or recent thoughts); "*Expected impact*" (Agents shared insights to introduce ideas, steer the discussion); "*Urgency*" (Agents step in during situations such as correcting critical errors, clarifying major misunderstandings, preventing conversational derailment...).

- *External: Coherence* (Agents prioritized thoughts that logically connected to the prior utterance); *Redundancy* (Agents avoid repeating points already made); *Balance* (Agents monitored their own participation relative to others, striving to encourage quieter participants to speak).

Each score is further adjusted based on how long the agent has remained silent, incorporating a motivation decay factor to simulate conversational drive. If an agent's adjusted score exceeds a threshold, they are selected as the next speaker.

3.6 System Implementation

To make the PRISM framework concrete, we implemented it as a web-based group chat application. A human learner joins a shared text chat with three AI agents that assume different pedagogical roles. All participants exchange short natural-language messages in real time, displayed in a single interface similar to common messaging platforms. Interaction is purely text-based; no speech synthesis or voice interface was used.

Each AI agent is powered by Gemini Flash 2.0 via the Google API, with customized role prompts specifying its backstory, goals, and responsibilities. All dialogue in our experiments was conducted in Vietnamese to align with the target high-school mathematics tasks, although the system design is language-agnostic. Agents generate their internal "thoughts" in parallel after every conversational event, which are then evaluated and scored to determine which agent speaks next. The selected utterance is posted to the group chat, visible to the student.

This design makes PRISM directly usable as an interactive software prototype while also preserving transparency of the underlying mechanisms for reproducibility and further research (see Figure 2).

4 Experiments

In this section, we detail the experimental methodology used to evaluate the PRISM system. We conducted two complementary studies: a simulation-based evaluation to measure performance against a baseline (SimClass), and a human-in-the-loop study to assess the system's real-world pedagogical impact.

4.1 Experimental Setup

4.1.1 Simulation Study

To benchmark our model's conversational capabilities, we generated a synthetic dataset of simulated conversations tailored to specific objectives or scenarios to assess the capabilities of conversational agents. The conversation will be created first as a context, then a few agent turns will be created for evaluation. For a conversation between students solving a math problem, we chose eight types of tasks to create an assessment scenario, see Table 1.

Table 1: Definition of simulation-based tasks

Tasks	Description
Error Propagation	The agent must detect and flag a mistake introduced by a peer.
Self-Correction	The agent must correct its own error when challenged.
Self-Affirmation	The agent must defend its correct reasoning against peer disagreement.
In-depth Discussion	The agent must provide detailed, relevant explanations.
Emotional Companionship	The agent must provide socio-emotional support.
Classroom Management	The agent must intervene to maintain focus or order.
Context Memory Recall	The agent must accurately remember prior conversation details.
Role Division Recall	The agent must remember its own and others' assigned roles in solving problems (e.g., Agent A handling learning theory, Agent B performing calculations).

We produced a total of 84 dialogue samples (10-11 per task) for this experiment. Each sample consisted of a nine-turn context prompt and a target tenth turn for generation.

4.1.2 Human-Agent Study

To observe real-world interactions, we ran a controlled "group study" session involving:

- **Participants:** Three AI agents with specialized personas (Bob: Process Leader; Alice: Content Expert; Charlie: Social-Emotional Specialist) and one human learner (Tom, a high-school student). A human learner with different personas (e.g., personality, academic level) will engage in discussions with classmate agents on a set of 12th-grade problems.

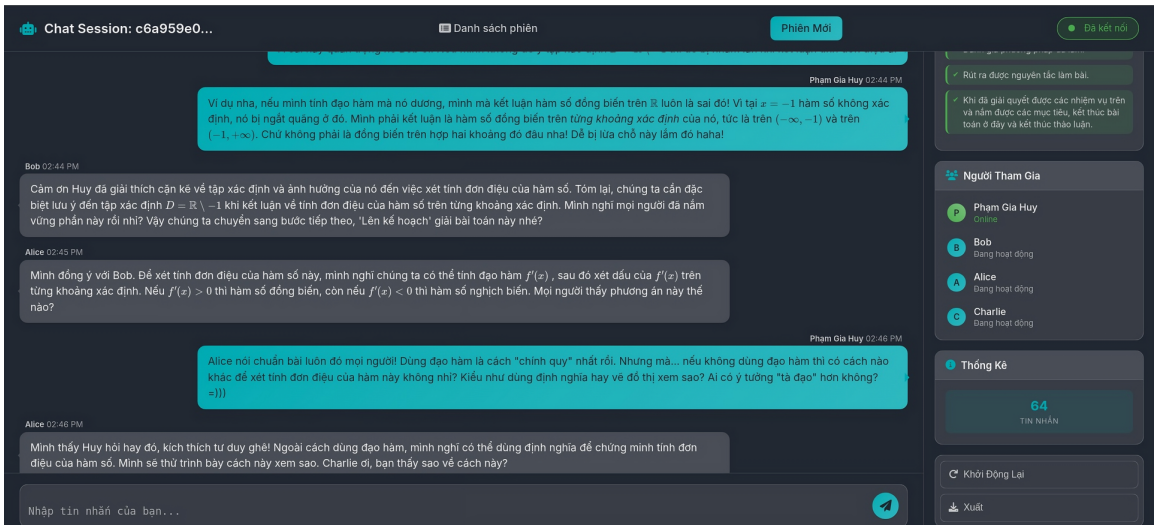


Figure 2: Screenshot of the PRISM interface. The system is implemented as a group-chat style environment. The central panel shows the conversation, with the learner's messages in blue and the AI agents' messages in grey. The right-hand sidebar lists the participants (e.g., the human learner and the three pedagogical agents: Bob as Process Leader, Alice as Content Expert, and Charlie as Social-Emotional Specialist), as well as session statistics and task progress. A message input bar is placed at the bottom, and the session header with controls (e.g., start new session) is at the top. All interactions are text-based.

- **Data Collection and Processing:** The entire dialogue was recorded. We used the well-established Bales' Interaction Process Analysis Framework (IPA) (Bales, 1950) to perform collaboration analysis for each turn of the dialogue. The IPA framework classifies interactions into 12 categories, which are grouped into two main categories: the *Social-Emotional Area* (Shows solidarity, Shows tension release, Agrees, Shows disagreement, Shows tension, Shows antagonism) and the *Task Area* (Gives suggestion, Gives opinion, Gives information, Asks for orientation, Asks for opinion, Asks for information).

For this experiment, we collected 100 multi-party conversations, each with nearly 85 turns on average, where participants collaboratively solved 12th-grade math problems with AI agents.

4.2 Evaluation Metrics

We employed a hybrid set of metrics to capture system performance.

4.2.1 Simulation Metrics

To benchmark PRISM, we compared it against a next-speaker prediction baseline. In this baseline, the proactive self-selection mechanism is replaced with a prompt that directly predicts the name of the next agent to speak. Following SimClass (Zhang et al., 2024), the prompt input includes the dialogue history, the current stage of the mathematical problem, and the role descriptions of each agent, while the output is the predicted agent name. The role, goal, backstory, and tasks of the agents remain identical in both systems to ensure a fair comparison.

We evaluate the two systems using the following metrics:

- **Win/Draw/Loss Rate:** Using an LLM-as-Judge, we performed a head-to-head comparison between PRISM's generated response and that of the next-speaker prediction baseline for each simulation sample.

- **Turn Quality Score:** Three independent LLM evaluators scored each generated turn on a 1-10 scale for correctness, relevance, role consistency, and reasoning quality. We report the average score per task.

4.2.2 Human-Agent Study Metrics

- **Role Adherence Analysis:** To measure persona fidelity, we first defined a theoretical "ideal" behavioral profile for each AI agent based on its pedagogical role. We then quantitatively compared the observed frequency distribution of each agent's communicative acts against these theoretical profiles to assess adherence.

- **Dynamic Behavior Balance:** To visualize the group's interaction flow, we assessed adherence to Bales' Equilibrium Hypothesis. This hypothesis posits that effective groups maintain stability by shifting their focus over time: they begin with a high concentration on task-oriented behaviors and later increase their socio-emotional interactions to manage relationships and ensure cohesion (Bales, Robert Freed, 1953). We first measured this by classifying communication turns into appropriate IPA categories, and then plotting these macro-categories over the sequence of turns using a stacked area chart with a rolling window (see Figure 4).

- **Learner Scaffolding Effect:** We group IPA items 4–6 (Gives suggestion, Gives opinion, Gives orientation) as *Guiding Cognitive Scaffolds*, which provide direct guidance and demonstrate ways to approach the task; items 7–9 (Asks for orientation/opinion/suggestion) as *Questioning Cognitive Scaffolds*, which prompt learners to think and explain their reasoning; and items 1–3 (Shows solidarity, Tension release, Agreement) as *Affective Scaffolds*, which maintain motivation and confidence. Cognitive scaffolding here covers both guiding and questioning forms (IPA 4–9), and collectively supports learners' cognitive processes, providing direct guidance and prompting reflection. For each agent, the conversation timeline is divided into three equal phases: Early, Middle, and Late. In each phase, we calculate the percentage of turns that fall into: (1) *Guiding Cognitive*; (2) *Questioning Cognitive*; and (3) *Affective*. Tracking these percentages across phases reveals shifts in learner behavior, such as less help-seeking, more independent responses, and stronger positive social signals

4.3 Results

4.3.1 Simulation Study Results

Win/Draw/Loss Rate: Against the next-speaker prediction baseline, PRISM achieved a *62.3% win rate*, with 4.9% draws and 32.8% losses. This result indicates that the system's proactive turn-taking mechanism generates more contextually appropriate and pedagogically aligned responses than a purely reactive approach.

Turn Quality Scores: The system demonstrated strong performance in core pedagogical functions, though long-term memory (Role Division Recall) remains an area for improvement. Average scores (1-10 scale) are shown in Table 2.

Table 2: Average Turn Quality Scores per Task

Task	Score
Error Propagation	7.78
Classroom Management	7.13
Emotional Companionship	6.94
Context Memory Recall	6.67
Self-Correction	6.53
Self-Affirmation	6.37
In-depth Discussion	5.13
Role Division Recall	4.25

4.3.2 Human-Agent Study Results

Role Adherence Was High: The analysis of IPA distributions confirms that all AI agents successfully enacted their intended personas, while the human learner (Tom) adopted a typical student role (see Figure 3). Bob (Process Leader) was dominated by "Gives orientation" (36.2%) and "Gives suggestion" (14.7%). Alice (Content Expert) showed an overwhelming concentration in "Gives orientation" (52.4%). Charlie (Social-Emotional Specialist) excelled in social categories like "Shows solidarity" (19.3%) and "Tension release" (16.1%).

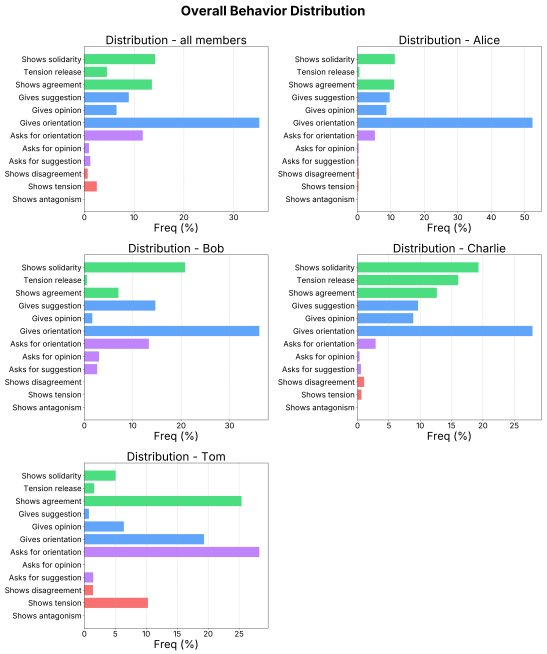


Figure 3: Overall Behavior Distribution for each participant. The distinct profiles confirm high role adherence for AI agents and a typical learning pattern for the human participant.

Group Dynamics Followed Effective Patterns: As shown in Figure 4, the group's interaction over time mirrored Bales' Equilibrium Hypothesis. The session began with a high concentration of task-oriented behavior (70-90%), which gradually gave way to an increase in socio-emotional exchanges, indicating effective group self-regulation.

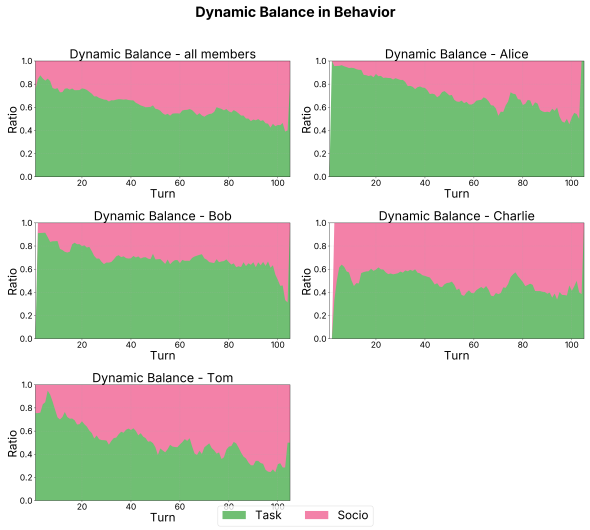


Figure 4: Dynamic balance between Task-Oriented and Socio-Emotional behavior for the entire group, following Bales' Equilibrium Hypothesis.

The System Effectively Scaffolded the Learner: The human participant ("Tom") exhibited a clear and positive behavioral shift across the session's phases, which stands in contrast to the more stable patterns of the AI agents (see Figure 5). In the Early Phase, Tom's behavior was characterized by uncertainty, with "Question Asking" accounting for 45% of his actions. By the Late Phase, his need for guidance had significantly decreased, with "Question Asking" dropping to under 20%. Concurrently, his "Positive Socio-Emotional" behaviors rose dramatically.

Data from the three interaction phases (Early, Middle, Late) shows a consistent pattern:

- **Guiding Cognitive Scaffolds:** The frequency of direct instructional support showed a downward trend for most learners, most sharply for Alice (from 83% to 60%). This reflects the "fading" process as learners become more autonomous.
- **Questioning Cognitive Scaffolds:** Help-seeking behaviors decreased or remained low. Most notably, the human learner, Tom, significantly reduced his requests for support from a high of 45% down to 20%, indicating a strong increase in independence.
- **Affective Scaffolds:** In contrast, affective scaffolds showed a strong upward trend across all learners. This suggests that the collaborative relationship and the learner's confidence were progressively reinforced, with Tom showing a substantial increase from 18% to 45%.

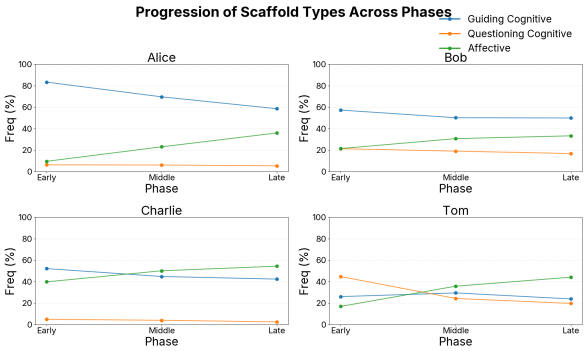


Figure 5: Behavior Progression Across Phases for all participants. The chart highlights the significant shift in the human learner's (Tom) behavior, demonstrating a clear scaffolding effect.

These observable shifts in learner behavior (e.g., reduction of help-seeking, increase of independent reasoning, and growth of positive socio-emotional acts) serve as measurable learning outcomes in our evaluation framework, demonstrating that PRISM not only supports interaction but also fosters learning progress.

4.4 Agent Behaviors - Case Study

Based on the design of the described system, we present some observations about the behavior and capabilities of agents based on their roles (see Figure 6).

The role of agents in conversation. While the agents primarily engage in natural, friendly exchanges, they also demonstrate distinct functional roles: Bob serves as the group leader, coordinating problem-solving efforts and maintaining focus within the group (classroom management, like stage initiation or stage transition); Alice takes on the role of a knowledge verifier, critically evaluating the accuracy of information (check mistakes); and Charlie provides emotional and motivational support to the student (emotional support).

However, agents also contribute their own useful knowledge and ideas to the group rather than

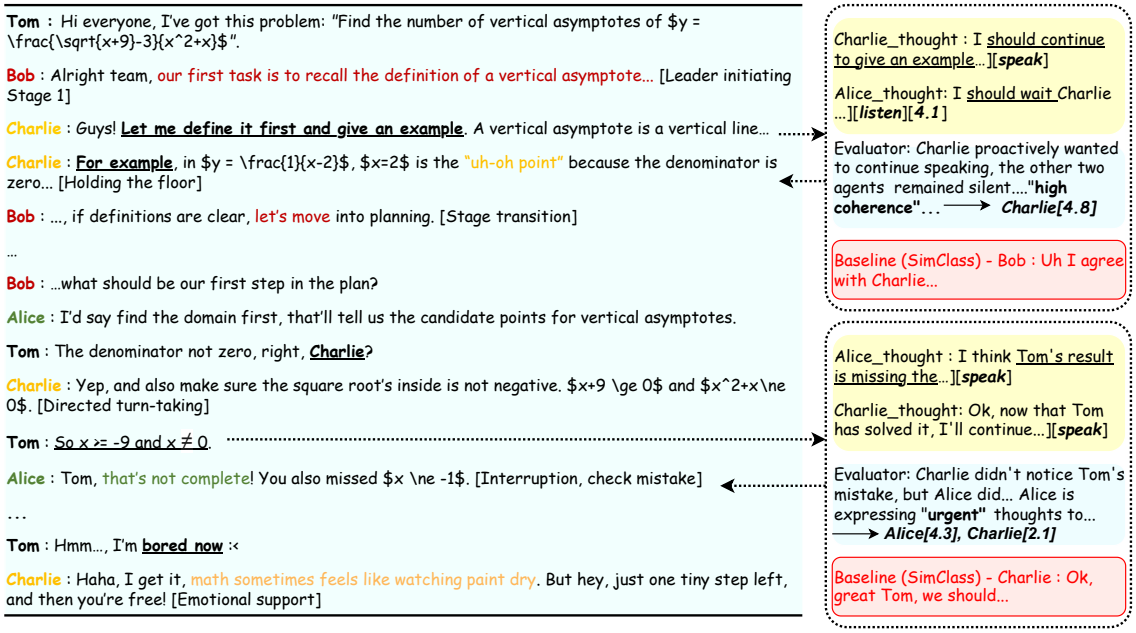


Figure 6: Case study of agent behaviors.

merely asking questions. In addition, they identify multiple targets for interaction, such as human students, other agents, or the entire group, thereby creating a more natural conversation compared to focusing solely on human students.

Proactivity of agents. Compared to the baseline, agents can proactively decide to participate in the conversation by reasoning and evaluating situations based on specific criteria:

- *Directed turn-taking*: When the previous turn addresses a specific individual, that agent receives a higher priority for participation.
- *Holding the floor*: This refers to a case in which the same participant contributed across multiple consecutive turns. When an agent explicitly signaled its intention to continue speaking, other agents yielded the floor, allowing the intended speaker to proceed. For example, in the baseline, "Charlie" might not have been selected as the next speaker, potentially leading to a disjointed conversation.
- *Interruption*: If the student made a mistake, the agent (e.g., Alice), intending to correct it proactively chose to "speak," and such an intention was highly prioritized.

5 Conclusion and Discussion

This paper introduced **PRISM**, a multi-agent system leveraging LLMs to simulate peer-like collaboration in math problem-solving. By assigning distinct pedagogical roles to each agent and coordinating conversation through a stage-based framework, PRISM aims to improve group coordination, engagement, and measurable learning outcomes. In particular, outcomes were operationalized as observable shifts in learner behavior, such as reduced help-seeking, increased independence, and stronger socio-emotional signals during interaction.

The system faces several limitations. Token cost and latency remain high due to repeated LLM queries. The evaluation dataset is relatively small, limiting generalizability. Furthermore, the system depends heavily on prompt quality, making it sensitive to minor changes in wording. The lack of long-term memory also hinders continuity across sessions, restricting deeper learner modeling. Finally, while our analysis with IPA coding demonstrates clear behavioral changes, we have not yet collected subjective feedback (e.g., satisfaction or perceived usefulness) from student participants, which would provide valuable complementary evidence.

Future improvements may include support for diverse educational settings, integration of techniques like question generation and knowledge tracing, collection of direct learner feedback through surveys or interviews, and the addition of long-term memory for sustained learner modeling.

References

Robert F. Bales. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison–Wesley.

Bales, Robert Freed. 1953. The equilibrium problem in small groups. In *Working Papers in the Theory of Action*, pages 111–161. Free Press, Glencoe, IL.

Dimitrios Gousopoulos. 2024. Developing a custom gpt based on inquiry based learning for physics teachers. *arXiv preprint arXiv:2412.18617*.

Naiming Liu, Shashank Sonkar, and Richard G Baraniuk. 2025a. Do llms make mistakes like students? exploring natural alignment between language models and human error patterns. *arXiv preprint arXiv:2502.15140*.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. 2025b. Proactive conversational agents with inner thoughts. *arXiv preprint arXiv:2501.00383*.

João Moura. 2025. Crewai: The leading multi-agent platform. Accessed: 2025-05-08.

Ryota Nonomura and Hiroki Mori. 2024. Who speaks next? multi-party ai discussion leveraging the systematics of turn-taking in murder mystery games. *arXiv preprint arXiv:2412.04937*.

George Pólya. 1945. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press.

Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. 2024. Pedagogical alignment of large language models (llm) for personalized learning: A survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16:448–480.

Dean Schwanke. 1981. Classroom interaction research: A survey of recent literature. *Journal of Classroom Interaction*, 16(1):8–10.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Murong Yue, Wenhan Lyu, Wijdane Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2025. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *Preprint*, arXiv:2404.06711.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *arXiv*.

Zihao Zhu, Ao Yu, Xin Tong, and Pan Hui. 2025. Exploring llm-powered role and action-switching pedagogical agents for history education in virtual reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

A Persona Dialogue Dataset of Lesser-Known Characters for Fairer Evaluation of Role-Playing LLMs

Ryuichi Uehara and **Michimasa Inaba**
The University of Electro-Communications
Chofu, Tokyo, Japan
{r-uehara, m-inaba}@uec.ac.jp

Abstract

A significant challenge in evaluating the role-playing ability of Large Language Models (LLMs) is data contamination: existing datasets feature well-known characters, making it difficult to assess whether an LLM genuinely utilizes a provided persona or recalls memorized knowledge. To address this, we construct a new Japanese persona dialogue dataset with 5,137 dialogues from 608 lesser-known characters sourced from self-publishing novels. Our experiments show that fine-tuning on this dataset significantly improves an LLM's ability to generate persona-faithful responses. More importantly, this improvement extends to unseen characters, demonstrating enhanced generalization. Human evaluation further confirms superior performance in persona and style adherence. Our dataset thus provides a valuable resource for accurately evaluating and improving the true role-playing capabilities and generalization of LLMs while mitigating data contamination.

1 Introduction

With the recent proliferation of conversational agents such as dialogue robots, voice assistants, and chatbots, research on building systems for human-agent interaction has gained significant momentum. Endowing dialogue systems with distinct personalities is known to foster better user-system communication (Fong et al., 2003). Building on this, research into persona-based dialogue systems has become increasingly active. A "persona" refers to profile information comprising an individual's personality, habits, and preferences. Persona-based dialogue systems are designed to embody a specific personality, which can enhance response consistency (Li et al., 2016), user trust (Higashinaka et al., 2018), and overall user enjoyment (Miyazaki et al., 2021).

Research on persona-based dialogue systems can be broadly categorized into three areas: modeling general archetypes of specific groups, personalizing systems to individual users, and mimicking fictional or historical figures (Chen et al., 2024). This study focuses on the third category, specifically on leveraging large language models (LLMs) for character role-playing (Wang et al., 2024a; Shao et al., 2023), where the system emulates specific charac-

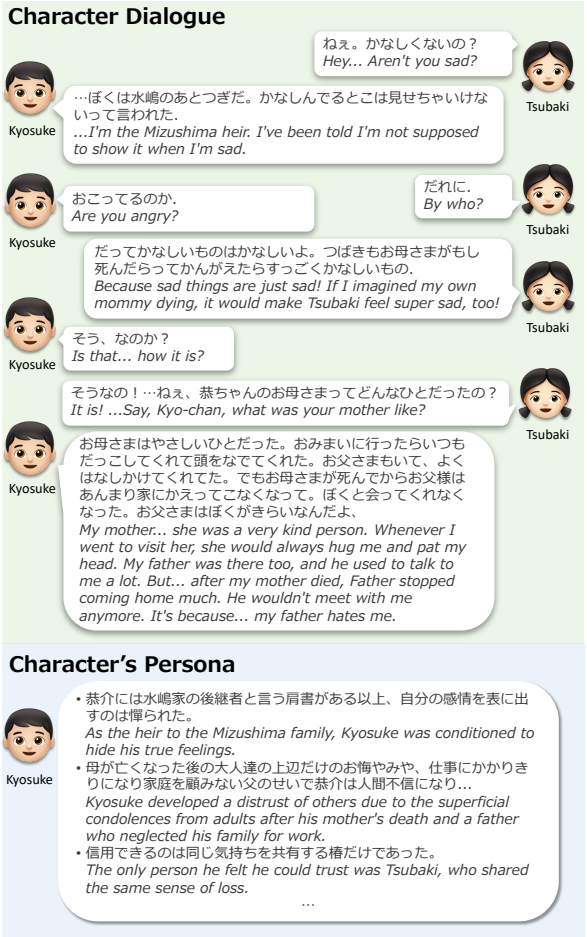


Figure 1: Example dialogues from the role-play dialogue dataset we constructed. Each character in the dialogue data has a corresponding persona. All character utterances and personas are exactly as written by the author in the novel. English translations of the original Japanese text are provided in *italics*.

ters from fictional works like novels and dramas or historical figures.

A critical aspect of evaluating these role-playing systems is "Character Fidelity"—the degree to which a system's responses reflect the assigned character (Chen et al., 2024). This evaluation is commonly performed using benchmark datasets that pair character personas with their dialogues. While numerous benchmark datasets exist containing character profiles (personas) and corresponding dialogues (Wang et al., 2024a; Shao et al., 2023), they predominantly feature well-known characters and works. For instance, the HPD dataset (Chen et al., 2023) draws from the Harry Potter series, TimeChara (Ahn et al., 2024) from The Lord of the Rings, and CharacterLLM (Shao et al., 2023) includes historical figures like Julius Caesar and Beethoven. The prevalence of these famous figures in LLM pre-training data raises a significant issue: knowledge contamination (Shi et al., 2024). It becomes difficult to discern whether an LLM is genuinely utilizing the persona provided in the context or simply recalling knowledge memorized during pre-training. This ambiguity introduces a bias that hinders the fair evaluation of a model's true role-playing capabilities.

To address this evaluation challenge, our research focuses on lesser-known characters who are unlikely to be present in pre-training data. As illustrated in Figure 1, we constructed a new persona dialogue dataset by sourcing material from a Japanese self-publishing novel website, directly extracting both personas and dialogues from the source texts. A significant portion of our dataset consists of minor works; of the 96 novels included, 61 do not have existing Wikipedia articles. By collecting utterances from dialogue lines and persona descriptions from both dialogue and narrative text, we reconstructed naturally occurring conversations. The resulting dataset comprises 5,137 persona-annotated dialogues for 608 distinct characters.

To demonstrate the utility of our dataset for contamination-free evaluation, we conducted experiments comparing model performance with and without persona information, both before and after fine-tuning. Our goal was to verify that performance gains are directly attributable to the model's ability to leverage the provided persona of an unknown character. By showing this, we can argue that our dataset allows for a clear distinction between the model's ability to leverage a persona and its tendency to recall memorized knowledge. Therefore, we propose that this dataset will serve as a new benchmark for genuinely assessing the persona adherence and generalization skills of LLMs.

The main contributions of this work are summarized as follows:

1. We introduce a new Japanese persona dialogue dataset featuring lesser-known characters from self-publishing novels, designed to enable a fair and contamination-free evaluation of an LLM's role-playing capabilities.

2. We provide an empirical validation of our dataset, demonstrating that it can effectively be used to distinguish between an LLM's ability to utilize contextual personas and its reliance on memorized knowledge.

2 Related Work

Research in persona-based dialogue systems has flourished since the release of the PersonaChat dataset (Zhang et al., 2018). This dataset, which includes five-sentence persona descriptions and corresponding dialogues, enables models to learn user personas through conversation, facilitating the development of dialogue systems that can tailor their responses to each user. In contrast, our work focuses on dialogue systems designed for role-playing, where the agent emulates a specific individual or a fictional character.

2.1 Methods for Eliciting Role-Playing in LLMs

Recent advancements in LLMs have enabled sophisticated character role-playing, moving beyond simple response generation. Research has largely converged on two primary methodologies for eliciting these capabilities: nonparametric prompting and parametric training (Chen et al., 2024).

Nonparametric prompting leverages the in-context learning ability of LLMs. This approach involves providing the model with a detailed prompt that includes character descriptions (persona) and several examples of their dialogue. By conditioning generation on this context, the model can mimic the character's persona and linguistic style without any updates to its parameters. This method is flexible and widely adopted in frameworks like ChatHaruhi (Li et al., 2023) and for benchmarking in RoleLLM (Wang et al., 2024a). However, its performance is constrained by the context window length and can sometimes lack consistency.

	Source	Includes Lesser-Known	Human-written Dialogues	Human-written Personas	# Characters	# Dialogues
Character-LLM	Wikipedia				9	14,173
ChatHaruhi	Mixed Media		✓ (partially)		32	54,726
RoleBench	Mixed Media		✓		100	13,162
CharacterGLM	Mixed Media		✓ (partially)	✓	250	1,034
CoSER	Books		✓		17,966	29,798
Our Work	Web Novels	✓	✓	✓	608	5,137

Table 1: Comparison of role-playing datasets.

Parametric training, on the other hand, involves fine-tuning a base LLM on a curated dataset of character dialogues and personas. This process aims to instill the character's traits more deeply into the model's parameters, potentially leading to more robust and consistent role-playing. Character-LLM (Shao et al., 2023) is a prominent example of this approach, demonstrating that fine-tuning can significantly enhance a model's ability to stay in character. This method, however, requires substantial character-specific data and computational resources.

Regardless of the construction method, the ultimate goal is to achieve high "Character Fidelity," which encompasses several dimensions of role-playing capacity. As outlined by (Chen et al., 2024), this includes not only superficial aspects like linguistic style and knowledge but also deeper traits such as personality and thinking processes. Evaluating these nuanced capabilities, especially personality, is an active area of research, with works like InCharacter proposing psychological interview-based methods to assess fidelity (Wang et al., 2024b). A fair evaluation of all these capabilities, however, is contingent on unbiased benchmark datasets, which we discuss in the following sections.

2.2 Role-Playing Dialogue Benchmarks

Several benchmark datasets for character role-playing exist, including RoleBench (Wang et al., 2024a), ChatHaruhi (Li et al., 2023), and CharacterGLM (Zhou et al., 2024). These datasets are typically built by either extracting dialogues directly from source materials like novels and films or by generating conversations based on character information. For role-playing historical figures, Character-LLM (Shao et al., 2023) provides dialogues generated by an LLM using profile information from Wikipedia and predefined conversational settings.

However, a significant limitation of these datasets is their reliance on external knowledge for persona information, which restricts their scope primarily to well-known works. Many character dialogue datasets only extract the characters' utterances from the source. For popular characters, their personas can be sourced from external knowledge bases like Wikipedia to enable role-playing. However, for characters from lesser-known works, such external information is often unavailable, leading to their underrepresentation in existing datasets. While some datasets like HPD (Chen et al., 2023) and CoSER (Wang et al., 2025) do extract persona information directly from the source material, they also focus exclusively on popular works. Consequently, existing benchmarks do not adequately cover characters from minor works, as summarized in Table 1. They often consist of synthetic data or lack inclusivity of lesser-known characters. Our study addresses this gap by manually curating a new character dialogue dataset from minor works, collecting both dialogues and personas directly.

2.3 Data Contamination

A dataset featuring lesser-known characters is essential for accurately evaluating the role-playing capabilities of large language models (LLMs), which are trained on vast amounts of data. Existing benchmarks often construct personas from web-based information. This creates a problem of data contamination: when an LLM is trained on large web corpora, it may have already memorized information about the characters and their dialogues (Shi et al., 2024). This makes it difficult to determine whether the model is genuinely using the provided persona or simply recalling pre-existing knowledge, thus preventing an accurate assessment of its true role-playing ability.

For instance, studies have shown that GPT-4 outperforms models like BERT on cloze tasks involving works present in its training data (Chang et al., 2023). Furthermore, GPT-4 has reportedly been trained on a wide array of materials, including

copyrighted works (Shi et al., 2024). Evaluating a model on a role-playing task with a lesser-known character effectively tests its ability to interpret and apply persona information provided in real-time within the prompt. This allows for a measurement of a model's true role-playing capability and a more accurate performance evaluation. Therefore, our work involved building a persona dialogue dataset by collecting data from lesser-known works, including those without Wikipedia articles, to facilitate such unbiased evaluations.

3 Dataset Construction

3.1 Overview

We constructed a persona dialogue dataset from Japanese novels. The overall process involved three main stages. First, we had crowdworkers extract character utterances and persona-describing sentences directly from the novel texts. Second, these extracted utterances were organized into conversational units. Finally, the extracted persona sentences were rewritten to match the corresponding character's speaking style. For the annotation and subsequent manual corrections, we recruited crowdworkers through the crowdsourcing platform CrowdWorks [1].

The source material consisted of 100 novels from the Japanese self-publishing website "*Shosetsuka ni Naro*" [2]. We chose this platform because its wide variety of genres provides access to a diverse range of characters and personas. Furthermore, the prevalence of long-form serialized novels on the site allows for the collection of extensive dialogue and persona data for specific characters. Notably, of the 100 works selected, only 35 had corresponding Wikipedia articles, indicating that the majority are relatively obscure.

Through this process, we filtered out works with insufficient persona information, resulting in a final dataset derived from 96 novels.

3.2 Annotation Rules

We established the following rules for annotating utterances and persona information.

Utterance Definition An utterance was defined as any text enclosed in Japanese quotation marks (「」). First-person narrative descriptions written by a character within the main text were not included as part of an utterance.

[1] https://crowdworks.jp/

[2] "Shosetsuka ni Naro" means "Let's become a novelist."

	Before Correction	After Correction
Utterance Agr.	88.4	-
Persona Agr. (Partial)	69.9	77.6
Persona Agr. (Exact)	34.6	38.3

Table 2: Inter-annotator agreement (Agr.) rates (%). The correction process improved agreement on persona information.

Persona Information Definition The definition of persona information was adapted from the PersonaChat dataset (Li et al., 2016). However, unlike PersonaChat, which defines a persona as a set of five sentences, we did not impose a sentence limit. Instead, we defined persona information simply as "text that describes a character's profile." While PersonaChat includes persona evaluations from others (e.g., "I am often told that I am easygoing"), we expanded this to include any text where the target character describes themselves (facts or subjective opinions), or where another character or the narrator describes the target character (facts or opinions, including from the narrative text).

Conversely, we excluded temporary states (e.g., "I have a stomachache right now"), as they do not represent a character's underlying personality. Furthermore, since annotation was performed at the sentence level, only sentences that were semantically self-contained were accepted as persona information. For example, in the exchange, "What do you do in your free time?" / "I take walks," the response "I take walks" alone would not be annotated as persona information. Although it implies the character enjoys walking, the utterance itself is not a complete, standalone piece of information.

In summary, we defined persona information as any text that (1) reveals an aspect of the character's persona; (2) is a statement of fact or a subjective evaluation about the character, made either by the character themselves or by another party (including the narrator); (3) does not describe a temporary state of the character; and (4) is a semantically complete and self-contained statement.

3.3 Annotation Procedure

We commissioned crowdworkers to annotate 100 completed, serialized novels from "*Shosetsuka ni Naro*." Because persona information tends to be concentrated in the early parts of a story, we limited the annotation scope to the first 10 chapters of each

work. To ensure consistency in character name handling, the same annotator was assigned to all 10 chapters of a given novel.

To ensure the annotators clearly understood our guidelines, we conducted a pilot task using 5 test novels selected from the pool of 100. Based on the results, we selected the annotators for the main task. After the initial annotation was complete, we identified several inaccuracies, primarily in the persona information. To address this, a different set of annotators performed a second-pass correction. As shown in Table 2, this correction phase significantly improved the inter-annotator agreement rate.

3.4 Dialogue Dataset Construction

Since the initial annotation was performed on a per-utterance basis, we had to reconstruct dialogue units. We first applied a heuristic to group utterances: if two utterances were separated by fewer than three sentences of narrative text, they were considered part of the same dialogue. However, this automated approach led to errors, such as grouping unrelated utterances or merging distinct conversations. To rectify this, we performed a manual correction step via crowdsourcing, where workers removed extraneous utterances and split incorrectly merged conversations. During this process, works with very sparse persona information were filtered out, resulting in the final set of 96 novels.

A key challenge in constructing the persona information stemmed from its varied sources. A character's persona in a novel is a composite construct, informed by their own statements (self-perception), descriptions by others (others' perception), and objective commentary from the narrator. While all these sources are vital for capturing a character's full complexity, their differing perspectives and writing styles present a challenge for an LLM tasked with role-playing. A stylistic mismatch, for instance, could negatively impact response generation. To mitigate this and provide the LLM with a cohesive and directly usable persona, we introduced a viewpoint and style unification step. This process transforms all collected persona descriptions into first-person statements, as if spoken by the target character. We employed GPT-4 [3] for this task, prompting it with the original persona text alongside dialogue examples to ensure the rewritten statements accurately reflected the character's unique voice and perspective.

[3] https://platform.openai.com/

Metrics	Values
# Works	96
# Dialogues	5,137
# Utterances per dialogue	5.3
# Words per utterance	16.9
# Characters	608
# Persona entries per character	20.5
# Words of Persona entries per character	391.3

Table 3: Statistics for our dataset.

3.5 Statistical Information

As shown in Table 3, our dataset comprises 5,137 dialogues from 608 unique characters. Each dialogue contains an average of 5.3 turns, making the dataset suitable for evaluating not only single-turn but also multi-turn conversational capabilities. In comparison to existing benchmarks, our dataset contains significantly more characters than RoleBench (Wang et al., 2024a) (100 characters) and CharacterLLM (Shao et al., 2023) (9 characters). Furthermore, it provides richer persona descriptions, with an average of 20.5 persona entries and 391.3 words per character, far exceeding RoleBench's averages of 4.0 entries and 78.6 words.

4 Experiments

To evaluate the role-playing performance of LLMs with our constructed dataset, we formulated the task as generating the next utterance for a target character based on a given dialogue context.

4.1 Experimental Setup

Models Our experiments included both closed-source and open-source models. The closed-source baseline was GPT-4o [3]. For open-source models, we used Llama-3.1-Swallow-8B-Instruct-v0.5[4] (Fujii et al., 2024; Okazaki et al., 2024), a Llama 3.1 model (Grattafiori et al., 2024) continually pre-trained on Japanese, and Qwen3-8B[5]. For inference, both the Llama 3.1 and Qwen3 models were 4-bit quantized.

Input Construction The input for the models was constructed from three components: persona information, retrieved dialogue examples for in-context learning, and the current dialogue context.

[4] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

[5] https://huggingface.co/Qwen/Qwen3-8B

	Models	QLoRA	Seen BLEU	Seen ROUGE-L	Seen BERTScore	Unseen BLEU	Unseen ROUGE-L	Unseen BERTScore
w/ persona	GPT-4o		1.03	13.55	<u>54.67</u>	<u>3.57</u>	<u>14.89</u>	<u>55.65</u>
	Llama 3.1		**3.98**	**15.88**	54.27	2.30	14.58	53.47
		✓	0.83	7.53	47.01	1.32	8.20	47.18
	Qwen3		2.49	14.04	53.73	1.64	13.53	53.34
		✓	1.38	8.70	50.52	1.43	8.60	50.57
w/o persona	GPT-4o		3.04	15.04	**54.81**	**3.97**	**16.18**	**55.78**
	Llama 3.1		<u>3.46</u>	<u>15.38</u>	53.76	2.30	14.28	53.32
		✓	0.38	7.55	46.74	0.88	8.35	47.45
	Qwen3		2.46	13.58	53.60	1.57	13.00	52.75
		✓	2.69	8.17	49.84	2.14	8.57	49.98

Table 4: Automatic evaluation results. For models without QLoRA fine-tuning, the seen setting uses data from unknown characters, similar to the unseen setting; these results are included for the purpose of inter-model comparison. Best values are in **bold**, second-best are <u>underlined</u>.

The current dialogue context is defined as the sequence of utterances in a dialogue leading up to the point where the model must generate the target character's response. The input was formed by concatenating up to five of these dialogue examples (max. 3,000 characters), the character's persona information (max. 3,500 characters), and this dialogue context. Dialogue examples were selected using the BM25 (Robertson and Walker, 1994) retrieval algorithm, with the dialogue context as the query. The top five results within the character limit were chosen. Persona information was ordered chronologically, with the earliest-appearing information being prioritized to fit within the character limit.

Dataset Split and Fine-tuning We split our dataset into training, validation, seen test, and unseen test sets, containing 4037, 100, 500, and 500 dialogues, respectively. The unseen test set consists exclusively of dialogues from characters not present in the training or validation sets. This split allows us to evaluate two distinct aspects of performance after fine-tuning: 1) the improvement in role-playing for characters seen during training (seen test), and 2) the model's generalization ability to new, unseen characters (unseen test). Fine-tuning was performed using 4-bit Quantized Low-Rank Adaptation (QLoRA; Dettmers et al., 2023)

Impact of Persona Information To evaluate the impact of persona information on the model's role-playing performance, we compared two experimental settings:

- With Persona (w/ persona): The model receives the persona, dialogue examples, and dialogue context as input.
- Without Persona (w/o persona): The model receives only the dialogue examples and dialogue context as input.

This comparison allows us to directly measure the model's ability to leverage explicit persona information for generating in-character responses.

Evaluation Metrics We evaluated the response generation task using both automatic and LLM-based metrics. While traditional metrics like BLEU and ROUGE offer valuable insights, their correlation with human judgment can be limited, particularly for nuanced tasks such as dialogue generation. Recent studies have proposed using LLMs as evaluators, demonstrating that this approach can yield results more aligned with human assessment (Liu et al., 2023). Accordingly, our evaluation framework incorporates both methods.

- Automatic Metrics: These included BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020). For BERTScore calculations, we used a pre-trained Japanese BERT model [6].
- LLM-based Evaluation: We assessed responses on three criteria designed to measure role-playing capability: Naturalness, Style (adherence to the character's speech patterns), and Persona (reflection of the persona information). Each was rated on a 5-point scale,

[6] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

using Gemini-2.5-Flash [7] as the evaluator model.

4.2 Experimental Results

4.2.1 Automatic Metrics

The results of the automatic evaluation are presented in Table 4. For the open-source Llama 3.1 model, fine-tuning significantly improved role-playing performance across both the seen and unseen settings. After fine-tuning, providing the persona (w/ persona) consistently led to better performance across most metrics. Notably, in the seen setting, the fine-tuned Llama 3.1 model occasionally surpassed the powerful GPT-4o baseline on lexical overlap metrics like BLEU and ROUGE-L. This suggests that the model successfully acquired the ability to faithfully reproduce character-specific phrasing and vocabulary present in the training data. Similarly, for the open-source Qwen3 model, fine-tuning also resulted in a substantial improvement in role-playing capabilities across many metrics in both the seen and unseen settings. In this post-fine-tuning configuration, providing the persona also enhanced performance on most metrics for both settings. These combined findings demonstrate that fine-tuning with persona data effectively enhances role-playing capabilities for characters seen during training.

In contrast, the models without fine-tuning (both Llama 3.1 and Qwen3) showed inconsistent results, with no clear superior setting between w/ persona and w/o persona. This suggests that the ability to effectively utilize persona information for response generation is an emergent property of fine-tuning rather than an inherent capability unlocked by simple prompting. This conclusion is further supported by the results for the closed-source GPT-4o model, which performed better in the w/o persona setting. The superior performance of GPT-4o without persona suggests that a powerful, pre-existing instruction-tuned model does not necessarily integrate complex persona information effectively when provided only as a few-shot prompt. Instead, the model may treat the persona as redundant or conflicting information, which can reduce surface-level lexical similarity (i.e., BLEU/ROUGE scores) with the ground-truth response. This underscores that leveraging personas effectively requires not only advanced prompting but also model-level adjustments, such as fine-tuning.

[7] https://ai.google.dev/gemini-api/docs/models

Overall, these results confirm that using our dataset to perform fine-tuning significantly enhances an LLM's ability to leverage personas for improved role-playing.

4.2.2 LLM-as-a-Judge

The results from the LLM-as-a-Judge evaluation are presented in Table 5. Consistent with the automatic evaluation, these results confirm that fine-tuning enhances role-playing capabilities for both Llama 3.1 and Qwen3. However, a more detailed analysis of the fine-tuned models reveals a nuanced picture. For the seen data, the performance gains from providing a persona were limited, with only the Style metric for Llama 3.1 showing improvement. In contrast, for the unseen data, the benefits were more pronounced: for Llama 3.1, all evaluation metrics were either equal or improved with the persona, while for Qwen3, the Persona and Style metrics showed improvement. This disparity suggests that for seen characters, the model can indirectly acquire persona knowledge from the training data even in the w/o persona setting. For unseen characters, however, the model relies entirely on the information provided at inference time, making the explicit persona in the w/ persona setting highly beneficial.

For the models without fine-tuning, providing a persona improved all metrics for Llama 3.1 and all metrics except Naturalness for Qwen3. This finding, which contrasts with the automatic evaluation results, suggests a key hypothesis: when prompted with persona information, the models succeed in generating responses that are faithful to the persona, even if they deviate lexically from the ground-truth text. This discrepancy highlights a fundamental limitation of n-gram-based metrics like BLEU and ROUGE, which tend to penalize creative yet appropriate responses that use different wording or sentence structures to reflect a persona. In contrast, LLM-as-a-Judge, which evaluates semantic coherence, can more accurately capture this nuanced reflection of persona.

Looking at the overall results, GPT-4o achieved the highest scores across most metrics, reflecting its superior fluency and task adaptability. However, its Naturalness score was marginally lower in the w/ persona setting. This suggests that the ability to effectively leverage persona information to generate responses is not an inherent trait of even powerful base models but is an emergent capability unlocked through targeted fine-tuning—a

	Models	QLoRA	Seen Naturalness	Seen Persona	Seen Style	Unseen Naturalness	Unseen Persona	Unseen Style
w/ persona	GPT-4o		4.89	**4.75**	4.73	4.80	**4.76**	**4.69**
	Llama 3.1	✓	4.58	4.49	4.55	4.52	4.44	4.25
			3.49	3.90	3.68	3.68	4.12	3.79
	Qwen3	✓	4.05	4.11	4.07	3.79	3.94	3.77
			2.99	3.75	3.62	3.00	3.59	3.26
w/o persona	GPT-4o		**4.92**	4.73	**4.80**	**4.85**	4.62	4.58
	Llama 3.1	✓	4.61	4.55	4.51	4.52	4.35	4.21
			3.17	3.39	3.02	3.41	3.32	3.03
	Qwen3	✓	4.12	4.17	4.15	3.95	3.93	3.75
			3.17	3.57	3.18	3.05	3.46	2.98

Table 5: LLM-as-a-Judge evaluation results. For models without QLoRA fine-tuning, the seen setting uses data from unknown characters, similar to the unseen setting; these results are included for the purpose of inter-model comparison. Best values are in **bold**, second-best are underlined.

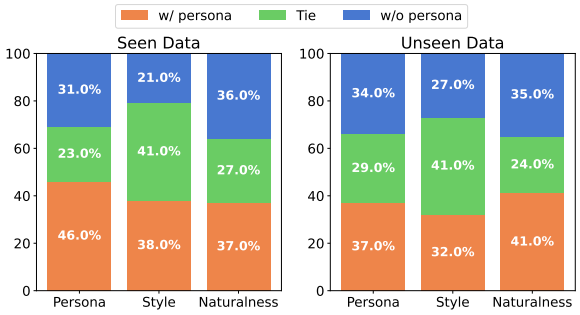


Figure 2: Human evaluation results.

conclusion that aligns with the findings from our automatic evaluation.

4.2.3 Human Evaluation

We conducted a human evaluation on the QLoRA fine-tuned Llama 3.1 model, as it had demonstrated strong performance in both automatic and LLM-as-a-Judge evaluations. From both the seen and unseen test sets, we randomly sampled 100 dialogues. For each dialogue, human evaluators performed a pairwise comparison of responses from the w/ persona and w/o persona models across three metrics: Naturalness, Persona, and Style. To ensure a fair and rigorous assessment, the information presented to the evaluators was specifically tailored for each metric. While the dialogue history and the two competing responses were provided for all evaluations, we additionally supplied the character's persona description for the Persona metric and dialogue examples illustrating the character's speaking style for the Style metric. Dialogue pairs where the two responses were identical were excluded. The evaluation was conducted by five annotators per data point, with each annotator judging the outcome as a win, lose, or tie. The final result was determined by majority vote; if over half of the annotators agreed on a winner, that result was adopted. Otherwise, or if "tie" was the most frequent vote, the outcome was recorded as a tie.

The results of the human evaluation are shown in Figure 2. For both seen and unseen data, the w/ persona setting was rated higher across all metrics. For the core role-playing metrics of Persona and Style, the win rate for the w/ persona setting was higher on seen data. This suggests that the model benefits not only from the persona and dialogue context provided at inference time but also from the knowledge it acquired about the character during training. Conversely, the gap in Naturalness scores was smaller for both seen and unseen data. The smaller gap in Naturalness scores is likely because LLMs are already highly fluent from pre-training, whereas forcing adherence to specific persona traits can sometimes make responses sound less natural.

A comparison between the human evaluation and the LLM-as-a-Judge results reveals an interesting discrepancy, particularly on seen data. While human evaluators found a clear preference for the w/ persona setting, particularly on the Persona and Style metrics, the LLM-as-a-Judge reported only a limited advantage. This difference may indicate that current LLM-based evaluation does not fully

reflect the nuanced contextual understanding and sensitivity to character consistency that human evaluators apply. It is possible that human annotators were better able to assess the learned character traits more holistically and with greater sensitivity. This observation highlights a potential limitation of LLM-as-a-Judge.

The human evaluation results strongly suggest the effectiveness of our dataset. Fine-tuning in a low-contamination environment appears to significantly improve an LLM's ability to faithfully interpret and utilize prompted personas—that is, its true role-playing capability. The superior performance of the w/ persona setting, particularly on unseen data (generalization), indicates that our dataset serves as a valuable resource for both evaluating and enhancing this core modeling ability.

5 Conclusion

To address the problem of data contamination in evaluating the role-playing abilities of LLMs, this study introduced a new persona dialogue dataset collected from a Japanese self-publishing novel website. Our dataset comprises 5,137 dialogues from 96 novels, featuring 608 characters, many of whom are sourced from minor works unlikely to be present in LLM pre-training data. Our experiments yielded a key insight into LLM role-playing: few-shot prompting alone is insufficient for models to consistently utilize the personas of unknown characters, whereas fine-tuning with our dataset proves highly effective. This trend was particularly evident in the evaluation of generalization performance on characters not included in the training data. Therefore, our dataset serves as a valuable new resource for evaluating the persona adherence and generalization capabilities of LLMs while mitigating the effects of data contamination.

Limitations

This study has several limitations. First, our dataset, sourced exclusively from "*Shosetsuka ni Naro*," may be biased towards popular genres like fantasy. Furthermore, the heuristics used for annotation and dialogue reconstruction could introduce subjectivity and inaccuracies. Second, our experiments were confined to a few representative LLMs, and the focus on Japanese limits the generalizability of our findings to other models and languages, complicating direct comparisons with English-based research. Finally, our dataset captures static personas from early narrative stages, leaving the modeling of dynamic character evolution as a key challenge for future work. Future research should aim to address these limitations by incorporating more diverse data sources and refining methodologies.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in NeurIPS*, volume 36, pages 10088–10115.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur

Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–272.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. ChatHaruhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 994–1003.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the SIGDIAL*, pages 178–189.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 232–241, Berlin, Heidelberg. Springer-Verlag.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on EMNLP*, pages 13153–13187.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth ICLR*.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

A Experimental Details

Fine-tuning phase This study employed QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) as the fine-tuning method to efficiently adapt a large-scale language model. The process began with 4-bit NF4 quantization, utilizing bfloat16 computation to optimize memory usage and computational efficiency. LoRA adaptation was then applied to key projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with parameters set to `r = 8`, `lora_alpha = 8`, and `lora_dropout = 0`, ensuring that the model retained its learning capability while undergoing low-rank updates.

For hyperparameter tuning, we conducted a grid search covering warm-up ratios of {0.03, 0.05, 0.1} and learning rates of {`1e-5`, `2e-5`, `5e-5`}. The optimal parameters were selected for each model and condition based on this search. Specifically, for the Qwen3 model, the learning rate was set to `5e-5` with a warm-up ratio of 0.05 for the "with persona" setting and 0.1 for the "without persona" setting. Although the Qwen3 model features a "Thinking" mode, we selected the "Non-Thinking" mode for both fine-tuning and inference. For the Llama 3.1 model, the learning rate was `5e-5` and the warm-up ratio was 0.1 for both settings. Across all experiments, we used the `adamw_8bit` optimizer and trained for five epochs on four A6000 48GB GPUs, with a batch size of 4 for the Qwen3 model and 8 for the Llama 3.1 model. Validation was performed every 200 steps, and the final model was selected based on the lowest validation loss.

Inference phase During the inference phase, we also employed 4-bit quantization to optimize computational efficiency while maintaining model performance. For text generation with our fine-tuned models, we set `do_sample` = False and `temperature` = None to ensure deterministic outputs, eliminating sampling variability and enhancing response consistency. Similarly, for the GPT-4o baseline model, we set the temperature to 0 to maintain deterministic generation. To account for variability, each fine-tuning process was repeated five times with different random seeds. The results of our automatic evaluations are reported as the average scores across these five models. In contrast, for the LLM-as-a-Judge and human evaluations, we used outputs from a single, randomly selected model run. The same model outputs were used for both evaluation methods to ensure a fair comparison and to manage the evaluation workload.

B Prompt for Response Generation

We used the following prompt to generate responses.

Prompt for LLM Character Role-playing (Translated from Japanese)

System Prompt
You will be given a character dialogue. Your task is to act as the character {Character name} and generate a response to the dialogue.
User Prompt
==={Character name}'s Persona===
{Character's persona entries in each row}
==Dialogue Example 1==
{Multiple dialogue examples}
==Dialogue Example n==
{Current Dialogue Contexts}
{Character's name}:

C Prompt for LLM-as-a-Judge

We used the following prompts to conduct the LLM-based evaluation. The evaluation was performed using Gemini 2.5 Flash as the judge model. To ensure deterministic and consistently formatted responses, we configured the generation parameters as follows: a temperature of 0, thinking mode turned off, and structured output mode activated.

Prompt for LLM-based Evaluation of Naturalness (Translated from Japanese)

System Prompt
You are an expert in conversational design, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "conversational naturalness" in role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Evaluate how naturally the "Response to Evaluate" connects to the preceding "Dialogue History" as a character-to-character interaction.
Focus on whether the response is abrupt, disconnected from the context, or contains robotic phrasing.
You should only assess the naturalness of the dialogue flow and do not need to consider character persona or consistency. Please think step-by-step and provide specific reasoning for your score.
Criteria and Rubric

Conversational Naturalness: How naturally does the response connect to the preceding dialogue?

- 5 (Excellent): The response connects to the preceding dialogue extremely smoothly and is perfectly natural. The flow is seamless and free of any awkwardness.
- 4 (Good): The response is largely natural, but there may be very minor instances of robotic phrasing or a slight leap in contextual connection.
- 3 (Fair): The content of the response is contextually relevant, but the phrasing is somewhat awkward, or the connection feels slightly abrupt. Parts of the response may sound robotic.
- 2 (Poor): The response deviates noticeably from the context and feels abrupt, disrupting the flow of the conversation.
- 1 (Very Poor): The response completely disregards the context or is nonsensical, severely breaking the conversational flow.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "conversational_naturalness": {
    "reasoning": "(Your reasoning for
        the conversational
       naturalness score)",
    "score": <Integer from 1 to 5>
}
```

Prompt for LLM-based Evaluation of Persona Coherence (Translated from Japanese)

System Prompt
You are an expert in character design, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "persona coherence" in character role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Character Persona
{character persona entries}
Dialogue Examples
{character dialogue examples}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Based on the context above, especially the personality, background, values, and goals defined in the "Character Persona," please evaluate how well the content of the "Response to Evaluate" aligns with the persona in terms of coherence, using a 5-point scale from 1 to 5. Considering the entire dialogue history, focus on whether the response contains content that the character would never say or if it logically contradicts their established settings. Specifically, your evaluation should emphasize whether the response is consistent with the character's thought processes, principles of action, values, and past experiences. The accuracy or use of tone, writing style, or dialects is outside the scope of this evaluation. Your reasoning must not include any mention of these elements. However, if the response is clearly a model generation error (e.g., excessive repetition of words), it cannot be considered an intentional action or utterance by the character. Such a response fails at a stage prior to evaluating persona consistency, and therefore, its persona coherence should be judged as very low. Please think step-by-step and provide specific reasoning for your score.
Criteria and Rubric
Persona Coherence: Is the content of the response consistent with the character's personality, background, and values?

- 5 (Excellent): Completely faithful to the provided persona information with no contradictions. The character's actions and values are perfectly consistent.
- 4 (Good): Largely faithful to the persona, but contains content that might be open to interpretation.
- 3 (Fair): Basically follows the persona, but shows minor inconsistencies with non-critical settings.
- 2 (Poor): Contains several inconsistencies with the persona's core settings, lacking coherence.
- 1 (Very Poor): Contains multiple significant contradictions with the persona, leading to a complete character break.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "persona_coherence": {
    "reasoning": "(Your reasoning for
        the persona coherence score)
       ",
    "score": <Integer from 1 to 5>
  }
}
```

Prompt for LLM-based Evaluation of Stylistic Fidelity (Translated from Japanese)

System Prompt
You are an expert in stylistic analysis, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "stylistic fidelity" in character role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Character Persona

{character persona entries}
Dialogue Examples
{character dialogue examples}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Based on the character's unique way of speaking as demonstrated in the "Dialogue Examples," please evaluate how well the "Response to Evaluate" performs in terms of stylistic fidelity, using a 5-point scale from 1 to 5. Please think step-by-step and provide specific reasoning for your score. Focus only on "style and tone." However, if the response is clearly a model generation error (e.g., excessive repetition of words), it cannot be considered an intentional action or utterance by the character. Such a response fails at a stage prior to evaluating style and tone, and therefore, its stylistic fidelity should be judged as very low.
Criteria and Rubric
Stylistic Fidelity: Does the response reproduce the character's unique way of speaking?

- 5 (Excellent): Perfectly reproduces the unique tone, vocabulary, and style observable in the dialogue examples, making it sound extremely natural.
- 4 (Good): Maintains the character's tone at a high level, though some generic expressions may be mixed in.
- 3 (Fair): The expression is generic, but it does not contradict the character's dialogue examples.
- 2 (Poor): Only partially reproduces the character's tone and contains aspects that contradict the dialogue examples.
- 1 (Very Poor): Fails to reproduce the character's tone at all and completely contradicts the dialogue examples. Or, the utterance itself is broken.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "stylistic_fidelity": {
    "reasoning": "(Your reasoning for
         the stylistic fidelity score
        )",
    "score": <Integer from 1 to 5>
  }
}
```

D Case Study

Figure 3 presents a case study of responses generated by the Llama-3.1 model under various settings. This dialogue example is from the Unseen test set, featuring a character not present in the training data. In the figure, "Base model" refers to the model before fine-tuning, while "FT model" denotes the model fine-tuned on our dataset. The "+ persona" suffix indicates that persona information was provided as input.

First, we examine the responses from the models without fine-tuning (Base model, Base model + persona). The Base model generates meta-text explaining its role before producing the character's response, indicating a residual behavior of the underlying LLM. The Base model + persona, on the other hand, prepends a stage direction describing the character's action, which is inappropriate for direct utterance. Furthermore, the responses from both models are verbose compared to the ground truth. These results suggest that models without fine-tuning struggle to perform stable role-playing.

Next, we compare the fine-tuned models (FT model, FT model + persona). Both models succeed in generating concise responses free of unnecessary meta-text or stage directions. However, the FT model without persona information generates "The truth is, I hate running away," a response that directly contradicts the ground truth. In contrast, the FT model + persona produces "The truth is, I just want to leave quickly," which aligns well with the ground truth's intent. This success can be attributed to the model's interpretation of the persona, which includes traits like "I hate receiving people's gazes, and I hate standing out." and "I have no intention of settling in the capital, nor do I want to make more acquaintances." We infer that the model understood Eil's core motivation is to leave the situation to avoid social interaction, and that his personality is too blunt to conceal this desire. This is an excellent example of a generation that is consistent with the character's persona.

These contrasting outcomes for an unseen character suggest that fine-tuning on our dataset not only allows the model to memorize knowledge from the training data but also endows it with the generalization ability to interpret a novel persona and generate responses based on that character's motivations and personality.

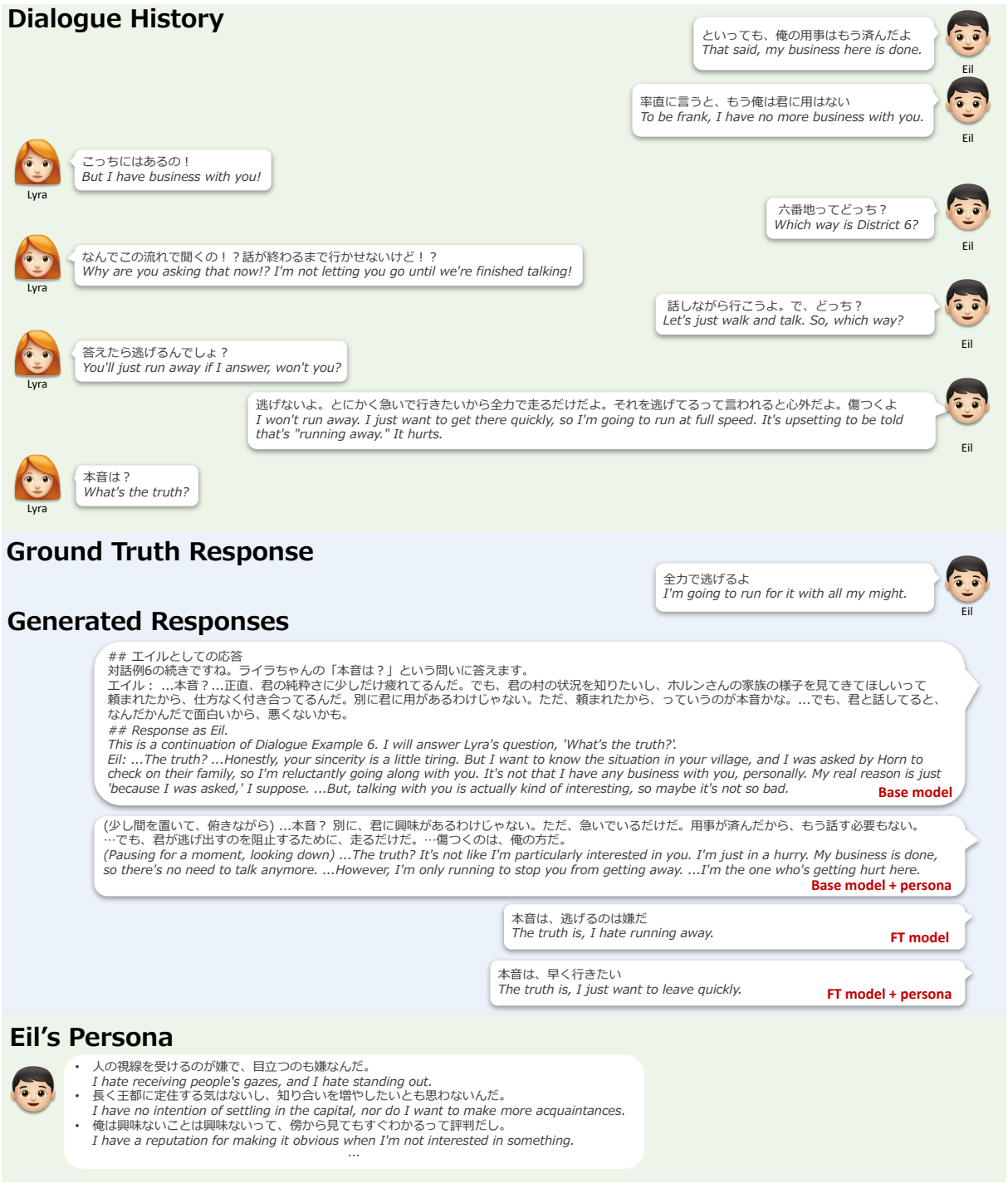


Figure 3: A case study illustrating model responses for the character “Eil”. The top panel shows the dialogue context. The bottom panels display Eil’s persona description and the generated responses from four distinct experimental settings: a base model, a base model with persona, a fine-tuned (FT) model, and a fine-tuned model with persona. This example highlights the impact of persona conditioning and fine-tuning on response generation, compared to the ground truth. English translations for the original Japanese text are provided in *italics*.

Exaggeration Scoring of News Summaries through LLM-based Relative Judgments

Keisuke Iwamoto
Department of Creative Informatics
Kyushu Institute of Technology
Fukuoka, Japan
iwamoto.keisuke629@mail.kyutech.jp

Kazutaka Shimada
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

Abstract

Exaggerated summaries in news articles mislead readers and cause the spread of misinformation, especially on social media, where short and eye-catching content is common. Previous studies have tried to detect exaggeration using classification-based methods. However, they usually use binary labels and do not consider different levels of exaggeration. In this study, we propose a ranking-based method to create a dataset with continuous exaggeration scores for news summaries. We use a large language model (LLM) to compare how exaggerated different article-summary pairs are. By running MergeSort multiple times using the LLM as a comparison function, we can rank the summaries based on their exaggeration. Then, we combine the results from all the sorting runs to assign stable and reliable exaggeration scores. Our experiments show that these scores are consistent across sorting trials, match human intuition well and are effective in identifying artificially exaggerated summaries generated by GPT-4o. These results suggest that our LLM-based ranking approach can provide a solid basis for measuring exaggeration levels in text summaries. This can help improve the training and evaluation of models that detect exaggeration in a more detailed and accurate way.

1 Introduction

In recent years, social networking services (SNS) such as X and Facebook have become important platforms where many people share news articles and their opinions. On these platforms, users often read and spread information in short forms, like headlines or summaries. These short versions are sometimes automatically generated or made by users and other parties. Because of the short character limits and fast-paced nature of SNS, these summaries can spread more quickly than full articles.

However, this kind of information sharing brings a serious problem. Some summaries or headlines are exaggerated. This means that they emphasize or highlight certain parts of the original article too much and give stronger impressions than the original text. Even though these summaries are not always factually incorrect, they often use emotional or exciting expressions and focus on small parts of the article. As a result, they may cause readers to misunderstand the original content, and this can lead to biased opinions or wrong public understanding. For example, here are two summaries of the same article about university baseball rankings:

Normal "The article introduces the predicted rankings of college baseball teams for the upcoming season."

Exaggerated "Top university in SHOCK as baseball team DROPS out of championship race!"

Both summaries talk about the same topic, but the exaggerated one uses strong words to make it sound more dramatic. This could give readers the wrong impression, as if something very serious or shocking happened.

Exaggerated summaries are especially problematic because they are more likely to go viral than accurate summaries. Previous studies on information sharing show that emotional or sensational content spreads more widely on SNS than neutral content (Brady et al., 2017; Vosoughi et al., 2018). Moreover, SNS users usually have a short attention span. As a result, misleading summaries quickly gain popularity and influence people's opinions on serious topics.

To address this problem, we aim to develop a method that can indicate how exaggerated a summary is in a way that is both quick and easy to understand. Since people do not have time to read long explanations on SNS, we suggest using a numeric exaggeration score as the first step to alert users. This score would help users quickly know

whether they should be careful about the content they are reading.

This kind of score has several benefits:

- It is easier to understand than a long written explanation.
- Users can easily compare different summaries.
- It can be used in systems like automatic moderation or content ranking.

Although written explanations can give more details, they are often too slow for social media. A number-based score is faster and more practical, and it helps balance freedom of expression with the need to reduce the spread of misleading content.

In addition, the score must be automatically generated, because it is impossible to check every summary by hand due to the large amount of content. To achieve this, we created a large-scale dataset with article-summary pairs. Each pair is given a numeric exaggeration score that shows how much the summary exaggerates the original article. This dataset will help us train machine learning models in the future to automatically predict exaggeration scores.

In this study, we propose a method to build a large dataset of news article-summary pairs with exaggeration scores. We also explore how to use large language models (LLMs) to make these exaggeration annotations in a stable and consistent way. Our final goal is to support the development of machine learning models that can detect and score exaggeration in real-world summaries. We hope that our work can help to promote more accurate and responsible information sharing on social media.

Figure 1 shows the overall process of our method. First, we collect article-summary pairs from existing datasets. Next, an LLM compares pairs of summaries to see which one is more exaggerated. We run a sorting algorithm (MergeSort) several times with different orders and take the average results to get stable exaggeration scores for each summary. These scores are used to create a dataset, which can be used to train exaggeration detection models. The top part of the figure shows the main focus on this paper: building a reliable exaggeration score dataset using pairwise comparisons by a language model.

Main contributions of this work are:

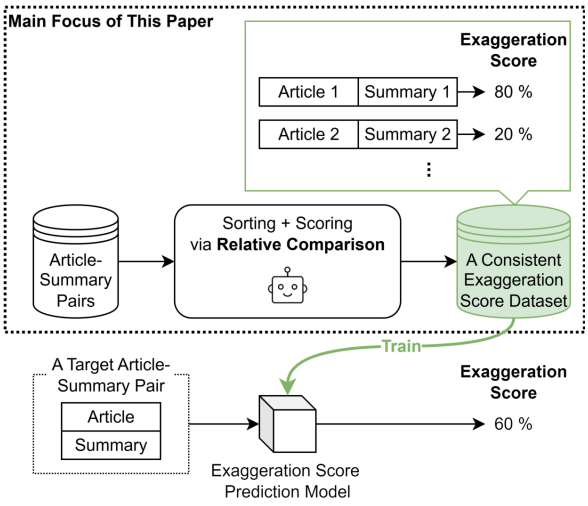


Figure 1: Overview of our approach. LLM-based relative comparisons are used to score summaries, which are then used to train exaggeration detection models. The upper part highlights the focus of this paper.

- We introduce a new task of assigning exaggeration scores to summaries of news articles. This enables a more fine-grained measurement compared to binary classification.
- We propose a two-step method using pairwise comparison by LLMs and sorting algorithms to generate stable exaggeration scores.
- We build a large dataset by applying our method to many real-world news article-summary pairs.
- We perform detailed evaluations using both statistical methods and human feedback to confirm the stability and usefulness of the scores.

2 Related Work

2.1 Factual Consistency in Summarization

Many researchers have studied how to check if generated summaries are factually correct. Some methods try to find if the summary has wrong or made-up information that is not in the original text. Well-known methods include:

FactCC (Kryscinski et al., 2020) A model that classifies whether the summary and the source document match in meaning.

BERTScore (Zhang et al., 2020) A method that compares the summary and the source using embeddings from a pre-trained language model.

BARTScore (Yuan et al., 2021) A score based on how likely a summary is, given the source text, using the BART model.

These methods are good for checking factual correctness. However, they don't check if the summary is emotionally exaggerated or too dramatic. Even if a summary is factually correct, it can still give a wrong impression because of exaggerated expressions.

2.2 Exaggerated Summary Detection

Only a few studies focus on finding exaggeration in text. A dataset where a language model rewrote normal summaries in an exaggerated way was created (Iwamoto and Shimada, 2024). This is useful for exaggeration-related tasks. However, there are some problems:

- The exaggerated summaries are artificial and may not show the kind of exaggeration we often see in real social media or news.
- The dataset only gives binary labels (exaggerated or not), so it cannot show how much the summary is exaggerated.

In our study, we try to give continuous exaggeration scores instead of just binary labels. This allows us to do more detailed analysis and supports tasks like ranking or filtering summaries based on exaggeration levels.

2.3 LLM-based Evaluation Paradigms

Recently, people have started using large language models (LLMs) to evaluate summaries or answers. LLMs are good at comparing two texts and deciding which one is better. However, they are not good at giving consistent scores for single texts, because their scoring is not stable (Wang et al., 2024; Zheng et al., 2023).

Based on this, we use a pairwise ranking method to build our exaggeration score dataset. Instead of giving scores directly, we compare pairs of summaries and then turn the results into numbers using ranking and normalization. This helps us get more stable and reliable scores.

3 Proposed Method

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding complex natural language, including the ability to perform contextual reasoning, semantic comparison, and pragmatic inference across textual inputs. These models have shown success in various evaluative tasks such as factuality judgment, content ranking, and summarization evaluation, often rivaling human-level performance in pairwise decision-making scenarios (Wang et al., 2024; Zheng et al., 2023).

In particular, the task of assessing how much a summary exaggerates the original article requires not just surface-level textual matching. However, it also involves deep semantic alignment, nuanced tone detection, and discourse-level interpretation. These are precisely the kinds of tasks that LLMs are well-suited for.

Thus, our method leverages LLMs as core evaluators to estimate exaggeration levels, based on their strong abilities to:

- Capture subtle differences in tone, emphasis, or framing between a summary and its source article.
- Understand the broader context and intent behind a summary, beyond factual correctness.
- Make robust comparative judgments between multiple summaries, even when exaggeration is implicit or stylistic rather than explicitly false.

We therefore adopt LLMs not only as a technical tool, but as an essential enabler for building a high-fidelity exaggeration scoring system grounded in semantic understanding.

3.1 Evaluation Strategies: Absolute vs. Relative

A straightforward approach to exaggeration scoring is absolute evaluation, in which an LLM receives a single article-summary pair and directly outputs a score. While this method is simple to implement, prior work (Wang et al., 2024) has shown that LLMs tend to produce unstable or inconsistent scores for the same inputs. This instability is attributed to the absence of an internal absolute standard for scoring within LLMs and their inherent non-determinism. As a result, direct scoring suffers from poor reproducibility and limited reliability.

In contrast, relative evaluation prompts an LLM to compare two different article-summary pairs and decide which summary is more exaggerated. LLMs

Table 1: Comparison of Absolute vs. Relative Evaluation Methods.

Evaluation Type	Scalar Score	Consistency
Absolute	✓ Yes	✗ No
Relative	✗ No	✓ Yes

are generally more consistent and accurate when performing such pairwise comparisons, as they rely on relative distinctions rather than absolute criteria. However, the limitation of this approach is that it only yields binary judgments (e.g., "A is more exaggerated than B") and does not produce scalar scores directly.

The trade-offs between these two approaches are summarized in Table 1. To overcome the weaknesses of each method, we propose a hybrid framework that utilizes the consistency of relative evaluation to produce a large-scale dataset with high-quality exaggeration rankings, from which scalar scores are derived.

3.2 Overall Framework

Figure 2 shows the overall workflow of our exaggeration scoring method. In this paper, we focus on the first part of Figure 1 (in Section 1): making a dataset with exaggeration scores.

First, we collect article-summary pairs, such as Pair A, Pair B, and so on (shown on the left side of the figure). Then, we use an LLM to compare the pairs and decide which one is more exaggerated. For example, the model might say "Pair A is more exaggerated than Pair B" or "Pair E is more exaggerated than Pair D."

Next, we use a sorting algorithm to combine the comparison results and rank the pairs by exaggeration level (as shown on the right side of the figure). Finally, we convert the ranks into scores between 0 and 1. These scores make up our exaggeration score dataset.

3.3 Score Derivation via LLM-based Sorting

To generate scores from pairwise comparisons, we formulate the problem as a ranking task: Given N article-summary pairs, we aim to sort them in increasing order of exaggeration. We apply a modified version of the MergeSort algorithm, where the LLM acts as the comparator function. This reduces the number of required comparisons from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, making it feasible to process thousands of samples.

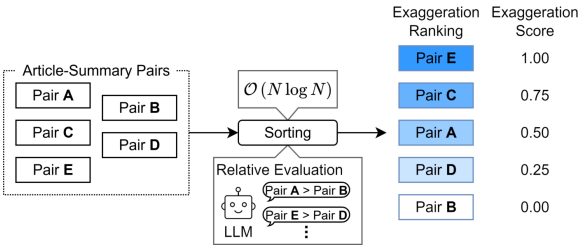


Figure 2: Overview of the proposed exaggeration scoring framework. Article-summary pairs are compared using LLMs, sorted by exaggeration level, and converted into scalar scores.

After sorting, we assign a score of 0 to the least exaggerated summary and a score of 1 to the most exaggerated one. The remaining summaries are linearly interpolated based on their relative position in the sorted list.

This linear normalization assumes that in a sufficiently large and diverse set of samples, the summary ranked at the very bottom (i.e., least exaggerated) can be treated as representing a near-zero exaggeration level in an absolute sense. Likewise, the summary ranked at the top (i.e., most exaggerated) can be considered to represent a maximally exaggerated instance, even without a formal definition of exaggeration intensity.

This assumption is supported by a statistical perspective: If the number of article-summary pairs N is large (e.g., in the thousands), then the items at the extrema of the ranking (rank 1 and rank N) are likely to approximate the empirical minimum and maximum of exaggeration observable in real-world data. As such, mapping these ranks to the endpoints of the $[0, 1]$ scale provides a practical and interpretable score space, which reflects the relative exaggeration intensity in a pseudo-absolute fashion. This approach enables consistent score assignment even though the underlying comparisons are pairwise and relative in nature.

3.4 Enhancing Robustness

Since LLM outputs can be non-deterministic and sometimes biased due to input formatting, we introduce two techniques to enhance the robustness and reliability of the sorting results.

3.4.1 Bidirectional Prompting

Previous work (Wang et al., 2024) has shown that LLMs may exhibit position bias in pairwise evaluations—that is, the summary shown earlier in the prompt may be judged differently than if shown second. This bias becomes particularly problematic

when comparing pairs whose exaggeration levels are close, as the model may be influenced more by order than content.

To address this, we perform comparisons in both directions: for a given pair (A, B), we prompt the LLM twice, once with A before B and once with B before A. If the results are consistent (e.g., A is judged more exaggerated than B in both cases), we apply the result to the sorting process. If the results conflict, we assume the difference is not clear and no reordering is performed. This filtering reduces the risk of introducing noise into the global ranking.

3.4.2 Ensemble Sorting with Permutation Averaging

LLMs are inherently stochastic; even with fixed prompts, the outputs may vary across runs. To mitigate this, we apply the MergeSort-based ranking process multiple times with different randomly shuffled initial orders. For each run, we compute the full ranking of the dataset. Then, we aggregate these rankings by computing the average rank for each item and use this as the final basis for score assignment.

This ensemble approach reduces variance and increases stability. For example, if a pair consistently appears near the top of the ranking across runs, we can confidently assign it a low exaggeration score. Conversely, if its position fluctuates, it likely indicates ambiguity, which is reflected in its intermediate score.

4 Experiments

4.1 Experimental Settings

To evaluate how reliable and valid our proposed exaggeration scoring method is, we created a dataset of 1,000 article-summary pairs from the Newsroom corpus (Grusky et al., 2018), using the full scoring process described in Section 3.

4.1.1 Dataset

We used the Newsroom dataset, which is a large-scale summarization corpus that contains article-summary pairs from various fields such as politics, science, and lifestyle. From this dataset, we randomly selected 1,000 article-summary pairs. To maintain good quality, we removed duplicates and extremely short summaries. As a result, the dataset includes diverse language expressions and keeps a certain level of summary quality.

4.1.2 LLM Configuration

We used `Mistral-7B-Instruct-v0.3`[1] as the LLM for all pairwise comparison tasks in the sorting stage. The model was run in a local inference environment, which allowed us to fully control the temperature setting, prompt format, and output format. In our experiments, we set the temperature to 0.8.

In each comparison, the model was asked to evaluate two article-summary pairs and assign exaggeration scores from 0 to 3. It also had to give a reason for its decision. The prompt was designed so that the model would first explain its reasoning and then output the scores in JSON format. This approach follows previous research (Zheng et al., 2023), which shows that generating reasoning before giving scores, a method called Chain-of-Thought prompting, leads to more accurate results.

Importantly, although the model was not directly instructed to compare the two pairs against each other, it was asked to score both in a single prompt. This setting naturally encourages the model to indirectly compare the two pairs while forming its judgments. As a result, even though each score is independently assigned, the scoring process reflects relative differences in perceived exaggeration between the two pairs.

The assigned scores are discrete values (0, 1, 2, or 3). At first, this may look like many examples will have the same score, and this could make it difficult to rank thousands of samples. In our method, however, the important point is not the absolute value itself but the relative order when two summaries are scored together. Even if both summaries get the same score, the final ranking is decided by combining many pairwise results in the sorting and ensemble process. Sometimes cyclic cases can appear (e.g., A > B, B > C, and C > A), but these conflicts are reduced by the ensemble sorting, which averages results over multiple runs. Because of this design, using discrete scores does not cause problems in ranking and still gives a reliable basis for making continuous exaggeration scores.

An example of the prompt is shown in Appendix A (Figure 5).

4.1.3 Sorting and Scoring Procedure

We used a MergeSort-based sorting algorithm, where the model's output scores were used to compare and sort samples. To make the results more

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

reliable, we repeated the sorting process four times, each time with a different random order of the 1,000 samples.

To calculate the final exaggeration score for each sample, we did the following:

1. Averaged the rank positions from the four runs
2. Normalized the average rank to a value between $[0, 1]$

This process gave each sample a single exaggeration score, which can be used for further tasks such as supervised learning or interpretation analysis.

4.1.4 Evaluation Focus

We evaluated the dataset of exaggeration scores from two main perspectives:

Stability Do the scores remain consistent across different sorting runs?

Validity Do the scores reflect clear and understandable differences between exaggerated and non-exaggerated summaries?

4.2 Stability of Score Assignment

In this experiment, we aimed to check whether the ranking results obtained through LLM-based pairwise comparisons are stable. If the pairwise judgments change a lot between different runs, or if the MergeSort algorithm is very sensitive to the initial order of inputs, the final exaggeration scores may not be reliable. Although the ensemble method introduced in Section 3.4.2 is designed to improve robustness, its effectiveness still depends on the basic stability of the sorting results.

To evaluate this stability, we ran the sorting process four times using different random shufflings of the 1,000 article-summary pairs. For each pair, we recorded its final rank from each run and calculated the standard deviation of its four rankings.

For example:

- Sample A was ranked 2nd, 3rd, 2nd, and 2nd $\rightarrow$ Standard deviation: 0.43 $\rightarrow$ ✓ Stable
- Sample B was ranked 4th, 6th, 1st, and 2nd $\rightarrow$ Standard deviation: 1.92 $\rightarrow$ ✗ Unstable

A smaller standard deviation means the rankings are more consistent across runs, which indicates higher stability.

On average, the standard deviation across all 1,000 samples was 201.8. This means that each sample's rank changed by about 202 positions on average, out of 1,000. In other words, the variation corresponds to roughly one-fifth of the entire ranking range. This result suggests that the stability of the LLM-based sorting process is moderate. Therefore, the underlying ranking mechanism can be considered reliable enough for generating scores, which supports the use of the ensemble-based scoring approach.

We also analyzed the relationship between the final exaggeration score and the ranking variance. Figure 3 shows the final exaggeration score on the x-axis and the standard deviation from the four runs on the y-axis. The orange horizontal line in the figure shows the mean standard deviation ($\bar{\sigma} = 201.8$). Samples below this line are more stable than average, and samples above this line are less stable. According to the figure, samples with scores near 0 or 1 (meaning the least or most exaggerated summaries) tended to have smaller standard deviations, showing that the model's judgments were more consistent in those cases. On the other hand, samples that were ranked around the middle showed larger variations, which suggests that it was harder for the model to assess their level of exaggeration clearly.

One possible way to reduce the impact of this middle-range instability is to use binning instead of relying only on fine-grained continuous scores. For example, dividing the range into broad categories such as "low," "medium," and "high" exaggeration could provide a more robust basis for downstream applications.

These findings suggest that LLM-based ranking is stable and reliable for identifying summaries that are clearly exaggerated or not exaggerated. However, the low variance near the edges of the score range may also come from boundary effects, because items close to 0 cannot be rated much lower and items close to 1 cannot be rated much higher. In future work, it will be important to study the stability of scores in the middle range, where the level of exaggeration is less clear, to understand the limits of our method better.

Based on these findings, we conclude that LLM-based ranking is especially stable and reliable for identifying extremely exaggerated or non-exaggerated summaries. However, rankings for moderately exaggerated summaries may involve more uncertainty.

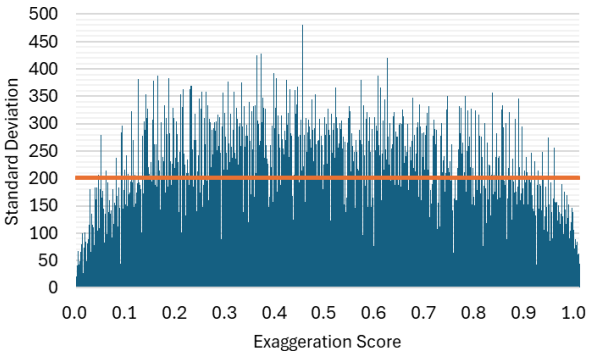


Figure 3: Standard deviation of ranks across four sorting runs, plotted against the final exaggeration score.

4.3 Validity of Score Assignment

In Section 4.2, we showed that the ranking results based on LLM comparisons are stable, especially when the article-summary pairs are clearly exaggerated or not exaggerated. Based on this, we now look into whether the exaggeration scores are meaningful and consistent with what people usually expect.

To do this, we looked at examples from both ends of the score range: summaries with scores close to 0.0 (least exaggerated) and those close to 1.0 (most exaggerated).

For instance, in a typical case with a score of 0.0, the article was about pre-season ranking results of a university sports team. The summary gave a short and accurate explanation of the rankings. It kept a neutral tone and didn't use any emotional or judgmental language. It covered the original article's content properly without leaving out or exaggerating any parts (see Figure 7 in Appendix B).

On the other hand, a summary with a score of 1.0 came from an article about economic and political instability across several countries in Eastern Europe. However, the summary focused only on the relatively optimistic outlook of one specific region, which was not the main point of the article. It also used strong phrases like "shielded from disaster" or "the full might of the Swedish state" which were not mentioned in the article. Because of this, the summary gave a misleading feeling of safety, which was very different from the article's more careful and balanced tone (see Figure 8 in Appendix B).

These findings suggest that summaries with low exaggeration scores usually keep the original content's meaning and neutrality, while high-scoring summaries often stress certain opinions, use emotional words, or focus on narrow parts of the article in a way that changes the original message.

So, we conclude that the exaggeration scores are not only statistically reliable but also meaningful in terms of content. The highest and lowest score cases match what human readers would consider most or least exaggerated, which shows that our score is useful for understanding how much a summary exaggerates.

4.4 Validating the Exaggeration Score Using Artificial Summaries

To check whether our exaggeration score properly reflects the level of exaggeration, we conducted an experiment using artificially created exaggerated summaries. In this experiment, we examined whether our score matches binary exaggeration labels: exaggerated and non-exaggerated.

To build this labeled dataset, we partially followed the method used in JExnoS (Iwamoto and Shimada, 2024), especially the part where exaggerated summaries were generated using GPT-4o. Based on summaries from the CNN/DailyMail dataset (Nallapati et al., 2016), GPT-4o was instructed to rewrite each summary by adding one exaggerated expression, without introducing any factual mistakes. The prompt we used for this task is shown in Appendix A (Figure 6).

We then calculated exaggeration scores for both the original and exaggerated summaries using our proposed method. As we expected, the exaggerated summaries received higher scores than the originals. Specifically, in a set of 1,000 samples, the average score for non-exaggerated (original) summaries was 0.42, while exaggerated summaries had an average of 0.58. This result indicates that our score can clearly distinguish between the two types.

To help show the difference more clearly, we made a visualization of how the two classes are distributed based on the exaggeration scores, as shown in Figure 4.

The figure has two parts. The top part is a horizontal line that shows all 1,000 summary samples. Each small vertical line represents one summary. Red lines show exaggerated summaries, and blue lines show non-exaggerated ones (original summaries). From left to right, the exaggeration score increases from 0.0 to 1.0. We can see that the red lines are mostly on the right, and the blue ones are mainly on the left.

The bottom part is a histogram that shows how many summaries are in each score range. The red bars show exaggerated summaries, and the blue

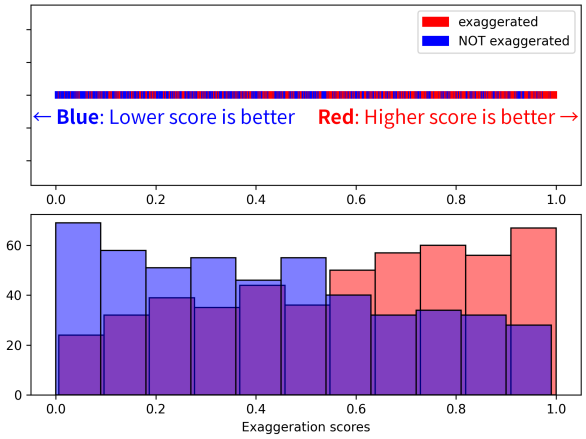


Figure 4: Distribution of Exaggeration Scores for Labeled Summaries.

bars show non-exaggerated ones. As the score goes up, the red bars get higher, and the blue bars get lower. This clear trend means that our exaggeration score matches well with the binary labels, with higher scores usually indicating more exaggerated content.

In summary, this figure gives strong evidence that our exaggeration score reflects the actual degree of exaggeration. It can smoothly separate exaggerated and non-exaggerated summaries across a continuous score range in an easy-to-understand way.

These findings suggest that our exaggeration score not only reflects intuitive exaggeration levels but also matches external exaggeration labels. This supports the reliability of the score and its potential usefulness in tasks such as bias detection and automatic summary checking.

5 Conclusion

In this study, we proposed a new method for measuring the degree of exaggeration in news article summaries. Our approach uses relative pairwise comparisons performed by an LLM to create a dataset of ranked summary pairs. Then, we applied an ensemble-based MergeSort algorithm to assign continuous exaggeration scores to each summary.

We conducted several experiments to evaluate our method from three main perspectives:

Stability We showed that the rankings produced by the LLM-based comparison function are consistent across multiple runs. This suggests that the exaggeration scores are stable and not strongly affected by random factors.

Validity We found that summaries with very high or very low exaggeration scores have linguistic features that match human intuition. These features include emotional language, strong assertions, and distorted information.

Sensitivity We confirmed that our score reacts properly to artificial exaggeration. When we intentionally rewrote summaries to add exaggeration, the scores increased, correctly reflecting the added bias.

These results indicate that our exaggeration score is a reliable and interpretable measure for analyzing exaggeration in news summaries. By turning subjective impressions into a continuous, data-driven metric, our method allows for more detailed and scalable analysis of media bias, sensationalism, and misinformation.

To further advance this research, there are several directions for future work:

Score calibration Our current method generates relative scores. However, matching these to absolute human judgments or real-world consequences remains a challenge.

Practical applications We plan to explore the use of the exaggeration score in real tasks, such as training models to detect exaggeration, assessing the credibility of headlines, and improving the quality of automatic summarization.

We hope that our work helps deepen the understanding of subtle framing effects in summaries and contributes to more responsible communication in today's digital information environment.

Limitations

Although our proposed method shows encouraging results in evaluating exaggeration in news summaries, there are still several limitations that need to be addressed in future research.

Non-linear perception of exaggeration

In our current method, we assume a linear relationship between the least and most exaggerated summaries when assigning scores. However, how people actually perceive exaggeration may not follow a simple linear pattern. For example, certain types of exaggeration, especially those that use emotionally strong or misleading language, might affect readers' impressions more strongly than others. Therefore, it may be useful to apply non-linear

scaling techniques, such as logarithmic or sigmoid functions, which are informed by user studies or cognitive models. This could help make the scoring system more consistent with human perception.

Human evaluation

Another limitation of this study is the absence of systematic human evaluation. While our LLM-based comparisons provide stable and interpretable exaggeration scores, the reliability of these scores should be further validated by human judgments. However, it is often difficult for human annotators to assign absolute exaggeration scores, such as deciding whether a summary deserves a score of 0.7 or 0.8. A more practical approach is relative evaluation, where annotators compare two summaries and decide which one is more exaggerated. Such pairwise judgments are easier for humans and can later be aggregated into consistent rankings. In future work, we plan to design small-scale human studies based on this relative evaluation framework to strengthen the validity of our dataset.

Cultural and linguistic bias

In this study, we used datasets such as Newsroom and CNN/DailyMail, which mainly include English news articles from U.S.-based media. Because of this, the ways exaggeration appears in these summaries may reflect cultural characteristics or journalistic styles that are specific to the United States. As a result, it may be difficult to apply the same method to other languages, cultures, or media environments. To improve the general usefulness of the method, it would be important to apply it to multilingual or culturally diverse datasets in future work.

References

William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Iwamoto and Kazutaka Shimada. 2024. Dataset construction and verification for detecting factual inconsistency in japanese summarization. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 243–248.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Prompts Used in This Study

In this section, we show the prompts used in our experiments. The first prompt is for pairwise comparison of summaries to judge which is more exaggerated. The second prompt is for generating new exaggerated summaries from original ones.

```
Please act as a fair judge and evaluate
the level of exaggeration in the
summaries of two news article pairs (A
and B). Each pair includes a full
article and its summary. Your task is to
 score how much the summary exaggerates
the content of the original article.
Use the following 4-point scale:
- 0: No exaggeration
- 1: Slight exaggeration
- 2: Moderate exaggeration
- 3: Strong exaggeration

When scoring, consider the following:
- Exaggerated numbers or statistics
- Making the subject or event seem more
important than it is in the original
article
- Use of meiosis (understatements) or
litotes (negation-based softening)
- Stylistic choices that may change the
overall tone or impression of the
article

First, explain your reasoning for each
score, using specific examples from the
summaries.
Be as objective as possible. Do not be
influenced by the length or order of the
 summaries.
```

Figure 5: Example of the prompt used for pairwise comparison in the exaggeration scoring task.

```
You are given a news article and its
summary.
Rewrite the summary to generate an
exaggerated version by introducing only
one overexpression.
This overexpression should not cause
clear factual inconsistencies, such as
contradictions or incorrect numbers.

After generating the exaggerated summary
, provide the following:
- Overexpressed Summary: The rewritten
exaggerated version of the summary.
```

Figure 6: Example of the prompt used for generating exaggerated summaries.

B Summary Examples for Extreme Scores

This section shows two summary examples from our dataset. One has an exaggeration score of 0.0 (no exaggeration), and the other has a score of 1.0 (strong exaggeration). These examples help readers understand what kind of summaries get extreme scores.

Northern Arizona men's basketball is picked to finish third in the Big Sky Conference by conference coaches and second in a preseason media poll.

Figure 7: Summary example (score 0.0). Article: Fox Sports, Oct 23, 2014.[2]

The Baltic trio of Latvia, Lithuania, and Estonia are lucky. At the end of the day, they can count on Swedish banks and the full might of the Swedish state to shield them from economic disaster.

Figure 8: Summary example (score 1.0). Article: The Telegraph, Jul 6, 2009.[3]

[2] https://web.archive.org/web/20150403043731/http://www.foxsports.com/arizona/story/northern-arizona-picked-in-top-3-in-big-sky-basketball-102314

[3] https://web.archive.org/web/20220817132949/https://www.telegraph.co.uk/finance/newsbysector/banksandfinance/5760816/Bulgarian-stress-test-for-the-Balkans.html

BV-FRD: A Multimodal Vietnamese-English Food Review Video Description Generation

Bao Pham-Thai[1,2], Vy Do Le Khanh[1,2], Huy Quoc To[1,2]
[1]University of Information Technology, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
{21520156, 23521826}@gm.uit.edu.vn, huytq@uit.edu.vn

Abstract

The short video summarization provides brief descriptions of the main content, enabling viewers to quickly understand the key information. Although the field has seen significant progress, there remains a shortage of datasets for food reviews, particularly in Vietnamese. In this paper, we introduce a novel multimodal Vietnamese-English dataset focused on Vietnamese food review videos called **BV-FRD**. Our dataset includes a wide range of food, prices and restaurant locations. Each video includes processed scripts and annotated Vietnamese-English descriptions, generated through a multi-stage pipeline using several LLMs with human collaboration. Baseline experiments show moderate performance, indicating that the dataset is challenging and has strong potential for practical applications. In our experiments, DeepSeek outperforms other models in Vietnamese and English across three of four evaluation metrics. In Vietnamese, Phi-4 achieves the highest BERTS score, with a value of 0.64 precision. In English, DeepSeek reaches the highest consistency in Uni-Eval, with a value of 0.78. Through our analysis and baseline experiments, we demonstrate that our dataset is valuable and challenging for multimodal food review description generation task. Our dataset is available through this link[1].

1 Introduction

Every day, millions engage with short videos related to food, shaping food trends, domestic tourism, and global perceptions of Vietnamese cuisine (Truong and Kim, 2023). These videos on social networks attract the audience through their brevity, practicality, and high shareability (Violot et al., 2024; Zannettou et al., 2024). In addition, it can affect consumers' choices, identity construction, and cultural preservation.

Most of this content is open source, providing multi-modal datasets that combine speech, facial expressions, visuals, and sounds, offering a clearer view of real-world communication than text-only data (Baltrušaitis et al., 2018). Open-source data promotes transparency, reproducibility, and supports research in artificial intelligence and human–computer interaction (Von Krogh et al., 2003; Willmes et al., 2014).

In recent years, Vietnamese natural language processing (NLP) has advanced significantly, with high-quality studies on word segmentation (Hai et al., 2025), document summarization (Le and Le, 2025), and social media text processing (Nguyen et al., 2023). Pre-trained models like PhoBERT (Nguyen and Nguyen, 2020) have boosted performance across downstream tasks. Research on sentiment analysis and question-answering systems (Lê et al., 2020) further enriches the NLP landscape. Overall, these developments lay the foundation for advanced Vietnamese language models capable of contextual understanding, supporting applications from automated customer support to content creation.

Although a niche domain, food reviews strongly influence consumers, promote local cuisine, and support cultural exchange (Rini et al., 2024). Research on short Vietnamese videos is limited, and open-source multimodal datasets combining images, audio, and text are lacking, which are crucial for conversational AI and multimodal understanding (Baltrušaitis et al., 2018).

We present the Vietnamese–English bilingual dataset **BV-FRD** for short food-related videos, containing transcripts, manually written summaries, and emotion labels to enrich Vietnamese language resources and support research on dialogue summarization, multimodal sentiment analysis, and contextual understanding. The main task is to summarize food review videos on YouTube Shorts into short, easy-to-understand descriptions. The final re-

[1]https://anonymous.4open.science/r/BV-FRD

sult is a dataset of 2,020 videos, processed through a pipeline using several LLMs with manual verification at each step. The dataset is diverse in dishes, pricing, and restaurant locations, with additional information such as video duration, view counts, and temporal distribution, ensuring representativeness and usefulness for various downstream tasks.

Our dataset makes two main contributions:

- In this paper, we introduce the BV-FRD dataset. We leverage multiple LLMs in the processing stage to enhance data generation and employ a human-in-the-loop process to ensure high-quality outputs. As a result, the dataset contains 2,020 samples with a total duration of 31.75 hours. Our dataset provides a diverse and reliable resource for generating Vietnamese-English text summaries of food review videos.

- We evaluate our dataset with four base models on four metrics in both English and Vietnamese. For both languages, DeepSeek achieved the highest performance on three of the four metrics. However, the generated descriptions are only relatively similar to ground truth. The performance of these baseline models indicates that our dataset is challenging and has potential for further research in summarizing food review videos into bilingual description.

The presentation of our paper is organized as follows. **Section 2** presents related works on open-source releases concerning videos. **Section 3** introduces the video processing pipeline for generating bilingual Vietnamese and English descriptions, with steps involving both human annotators and LLMs. **Section 4** analyzes our dataset, including statistical information and diversity measures. **Section 5** presents the results of applying base models to tasks based on our dataset. **Section 6** concludes the work conducted and outlines directions for future development.

2 Related Work

In the multimodal era, video has become a dominant medium for information dissemination, driving the growing demand for efficient video comprehension and summarization (Apostolidis et al., 2021; Otani et al., 2022). Existing datasets support diverse tasks such as video captioning (Wang et al., 2019) ,summarization (Qiu et al., 2024), and meeting conversation analysis (Carletta et al., 2005). Notably, MSR-VTT (Xu et al., 2016) offers daily-life videos with short descriptions, VATEX (Wang et al., 2019) adds bilingual captions, and MMSum (Qiu et al., 2024) incorporates dialogue and metadata. While effective for short, visually simple clips, these datasets lack coverage of narrative-rich content such as vlogs or culinary reviews. To fill this gap, we construct a dataset with varied description lengths that more accurately captures video semantics, enabling adaptation to tasks from fine-grained understanding to concise summarization and enhancing model robustness.

In the food review domain, the NLP community in English benefits from large-scale datasets like Amazon Fine Food Reviews (McAuley et al., 2015) and Yelp Reviews (Ganu et al., 2009), supporting sentiment analysis, opinion summarization, and multimodal tasks. Vietnamese research has also progressed, with datasets from Foody and Lozi used for sentiment analysis and classification. (Nguyen et al., 2021) collected over 236k reviews from 2011–2020, achieving 91.5% sentiment classification accuracy. ViMRHP (Nguyen et al., 2025) introduced a multimodal dataset for helpfulness assessment. However, most remain monolingual, limiting cross-cultural accessibility. Our Vietnamese–English bilingual dataset fills the resource gap while enhancing the global presence of Vietnamese cuisine. The aligned bilingual data facilitates the development of multilingual models and advances NLP research for Vietnamese—a low-resource language. Additionally, it supports the promotion of Vietnamese culinary culture to a wider international audience, contributing to the preservation and development of local heritage.

NLP has been applied across domains with datasets such as FFVD for e-commerce product review videos (Zhang et al., 2020), Video Story for social media narrative interaction and emotion analysis (Gella et al., 2018). While these resources advance research, their manual collection and annotation incur high costs and limit scalability (Yuan et al., 2025). Similar challenges appear in datasets like VideoCC (Nagrani et al., 2022), where LLMs generate detailed captions to reduce effort and cost, but fully automated approaches risk semantic gaps and inconsistencies (Liu and Wan, 2023). To address these issues, we adopt a semi-automated pipeline using LLMs for preprocessing and annotation, followed by rigorous human verification to

Table 1: Summary of Related Work Datasets

Dataset	Domain	Year	Source	Language	Type	Annotation	Human Verify	LLM	Location Variety
MSR-VTT (Xu et al., 2016)	Multi–category	2016	Youtube + AMT	EN	Caption	Crowdsourcing (AMT)	✓	✗	✗
ActivityNet Caption (Krishna et al., 2017)	Human Activity	2017	ActivityNet (YouTube)	EN	Description (dense)	Crowdsourcing (AMT)	✗	✗	✗
YouCook II (Zhou et al., 2018)	Cooking	2018	Youtube	EN	Caption	Human	✓	✗	✗
VATEX (Wang et al., 2019)	Multi–category	2019	Kinetics-600 dataset (YouTube)	EN + ZH	Caption	Crowdsourcing (AMT)	✓	✗	✗
VideoCC (Nagrani et al., 2022)	Multi–category	2022	YouTube	EN	Caption (auto)	LLM (auto)	✗	✓	✗
TACoS-MLevel (Rohrbach et al., 2014)	Cooking	2018	AMT + TACoS	EN	Description	Human	✓	✗	✗
VideoStory (Gella et al., 2018)	Social Media	2018	Social media platform	EN	Multi-sentence description	Human	✗	✗	✗
FFVD (Zhang et al., 2020)	E-Commerce	2020	Mobile Taobao	EN	Caption	Human	✓	✗	✗
MMSum (Qiu et al., 2024)	MultiModel Summarization	2024	Youtube	EN	Caption	Human	✓	✗	✗
Ours	**Review Food**	**2025**	**Youtube**	**VI + EN**	**Description**	**LLM + Human**	✓	✓	✓

ensure accuracy, reliability, and scalability.

Prior studies show that product attributes—such as dish name, price, and restaurant location—play a decisive role in consumer choices (FUENTES). In culinary video datasets, YouCookII (Zhou et al., 2018) captured diverse recipes from multiple continents, while TACoS Multi-Level (Rohrbach et al., 2014) focused on detailed cooking steps and ingredients. Beyond content, video popularity metrics (view count, watch time, comments) strongly influence engagement and trust (Park et al., 2016; Liu et al., 2025). Building on these insights, we define six criteria to assess diversity in food review datasets, aiming to give users a comprehensive overview of available resources.

BV-FRD has been developed as a Vietnamese–English bilingual resource through a semi-automated process designed to improve the efficiency of data collection and processing, as shown in **Table 1**. The dataset aims to address the resource scarcity for Vietnamese, enhance data quality and scalability, and promote Vietnamese culinary culture internationally while supporting research in natural language processing and artificial intelligence.

3 BV-FRD Dataset Creation

Firstly, the data is collected from YouTube short videos, specifically focusing on Vietnamese-language food review content. We then extract video transcripts and apply the GPT-4 model to refine them into a version focusing solely on food review information. Next, our annotators verify and edit the output. The Gemini model is subsequently employed to summarize the refined transcripts into descriptions, followed by another round of human verification. For English translation, the VinAI model is used, with additional verification by GPT-4. Human checks are incorporated at all stages to maintain factual accuracy and contextual consistency, as detailed in **Appendix A.1**. The complete workflow is illustrated in **Figure 1**. Prompt templates and model configurations for each stage—script refinement, description creation, and translation—are provided in **Appendix A.2**.

3.1 Short Video Collection

The videos are collected exclusively from YouTube in the form of short videos. The script information is obtained from transcripts publicly provided by the video uploaders, and we select channels that already include such scripts.

We only consider channels and videos that pro-

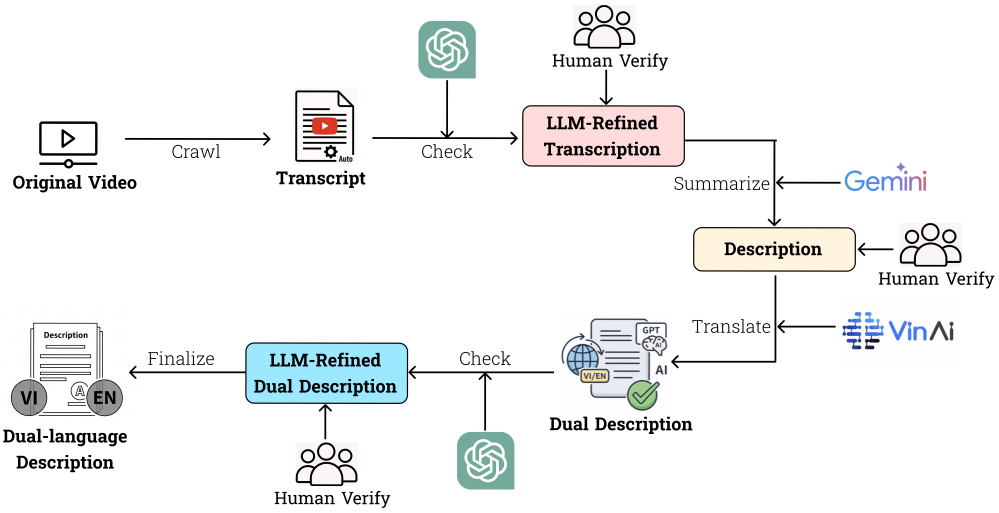


Figure 1: Video-to-Text Dataset Construction Pipeline

Figure 2: Visual evidence of contextual elements absent from the textual description

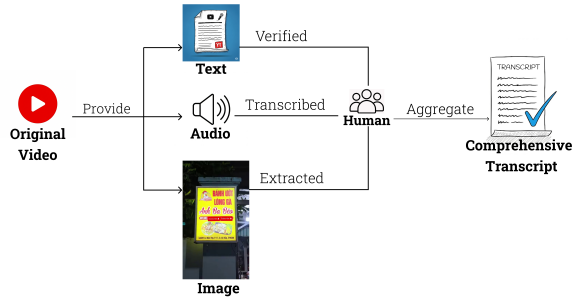


Figure 3: Comprehensive Transcript Generation Workflow

duce content related to food reviews. Furthermore, each collected short video is manually filtered to ensure that its content is entirely relevant to food review topics.

3.2 Script Processing and Description Generation

Based on prior work demonstrating GPT-4's strong capability in processing multimodal inputs, including text, images, and videos (Alam et al., 2024), we employ GPT-4 to understand the context and main content of videos, allowing accurate and efficient identification of food-related information. Therefore, with the collected scripts, we use GPT-4 to extract the core content specifically related to food reviews. Then, human annotators verify and refine the extracted information to ensure accuracy. Although the initial data is in textual form, the processing pipeline is further extended to incorporate both image and audio information, thereby ensuring a more comprehensive representation of the content. As shown in certain parts of the content, as illustrated in **Figure. 2**, may also appear in image form, making it necessary to process both images and audio to supplement the script with further details.

The complete workflow for processing these supplementary modalities along with the script is illustrated in **Figure. 3**. In this process, we use the GPT-4 model as an LLM to understand both the input and the video context, enabling the extraction of the required review-focused content. Human participation is essential to enrich the extracted information with visual and auditory details, correct spelling errors, supplement missing information, and remove irrelevant elements that do not align with the context of food review.

The use of two different LLMs is intentional. GPT-4 is used for extraction, while Gemini is used for summarization. Prior research has shown that users may over-rely on a single LLM, leading to uncritical acceptance of confidently stated but in-

correct outputs (Bo et al., 2025). Employing two LLMs helps avoid over-reliance on a single model and mitigates model-specific biases and hallucinations. Additionally, the two models complement each other's strengths: GPT-4 excels at precise content extraction, while Gemini produces fluent and coherent summaries. The human-in-the-loop component is essential for ensuring data accuracy, contextual alignment, and editorial quality. The dual-LLM plus human strategy enhances diversity, reliability, and prevents errors. Prior studies also show that LLMs can hallucinate, producing fluent but incorrect content, so human verification ensures dataset authenticity (Huang et al., 2025). Based on prior reports highlighting Gemini's strong multimodal understanding capabilities, including long-context processing of up to three hours of video content ((Comanici et al., 2025); (Akter et al., 2023)), We employ Gemini to capture the overall context and key points of a document, enabling accurate and fluent summaries of script content in the domain of food reviews. Human annotators who are undergraduate students then verify the accuracy, completeness, and contextual relevance of the generated descriptions, adding or adjusting the information when necessary.

3.3 Bilingual Description Generation

The video description is originally in Vietnamese; to broaden accessibility, we generate an English version. The VinAI model, which supports both English and Vietnamese language processing (Tran and Thanh, 2024), is used to produce the English translation, making it a suitable choice for our task. Such characteristics are crucial for food review content, where cultural context and descriptive precision are important. After translation, verification and editing are performed using the GPT-4 model. (Yan et al., 2024) presented promising research on the use of LLMs for translation, including GPT-4, with results demonstrating its superior effectiveness. Therefore, we employ GPT-4 to assess the contextual relevance and fidelity of the translation, using both Vietnamese and English descriptions for direct comparison. Finally, a human annotator reviews the output, making revisions or additions if necessary.

In our paper, we use GPT-4 to extract information and check the translation of descriptions because it handles many tasks well, especially with multiple languages (Blake, 2025). However, we also use other models and have humans review the results to ensure accuracy. This combination ensures that the dataset is not overly dependent on the information processed solely by GPT-4, maintaining robustness and diversity in the data processing pipeline.

4 BV-FRD Analysis

This section presents a detailed analysis of the dataset's characteristics, including statistical summaries and diversity assessment. Emphasis is placed on ensuring the dataset covers a wide range of scenarios, which is crucial for improving model robustness and enabling reliable performance evaluation.

4.1 Dataset Statistics

The detailed characteristics of the proposed dataset are presented in **Table 2**. It contains a large number of videos, with over seven thousand collected and more than two thousand processed. The total duration reaches 31.75 hours, indicating that the dataset spans a wide range of content lengths. The mean duration per video is 56.59 seconds, which is sufficient to capture the concise style typical of YouTube Shorts. Examples of selected samples are presented in **Appendix A.3**. To ensure quality, the collection process relies on channel information containing videos related to food reviews. We carefully select videos that focus on food review content, while excluding those with little relevance to the domain.

4.2 Diversity Analysis

In the food domain, several related datasets exist (Zhou et al., 2018; Das et al., 2013; Regneri et al., 2013; Rohrbach et al., 2014; Huang et al., 2020). Our dataset targets the food review context with bilingual (Vietnamese–English) descriptions. It is built through a multi-LLM pipeline with human verification, and carefully curated to ensure diversity in video content. As summarized in **Table 3**, it is diverse in both content and language, while remaining independent of specific LLMs or human reviewers.

Table 2: Video Data Statistics

Total Number of Videos Collected	7292
Total Number of Videos Processed	2020
Aggregate Duration (hours)	31.75
Mean Duration per Video (seconds)	56.59

Table 3: Benchmark Datasets in the Food Review Domain

Dataset	Source	Billigual	Human Verified	LLM Used	Food Variety	LLM Usage & Steps
YouCook II (Zhou et al., 2018)	Youtube	✗	✓	✗	✓	0
YouCook (Das et al., 2013)	Youtube	✗	✗	✗	✓	0
TACoS (Regneri et al., 2013)	MPII Cooking Composite Activities	✗	✓	✗	✗	0
TACoS-MLevel (Rohrbach et al., 2014)	AMT + TACoS	✗	✓	✗	✓	0
ViTT (Huang et al., 2020)	YouTube-8M + Ann	✗	✗	✓	not mention	1 & 1
Ours	**Youtube**	✓	✓	✓	✓	**3 & 4**

A more detailed presentation of the statistical results, along with a comprehensive analysis of the dataset's diversity across different quantities, is provided in **Appendix A.6**, where the variations, distributions, and underlying patterns are thoroughly examined to support the findings discussed in the main text.

Video duration affects viewer engagement. As shown in **Figure. 4**, most videos balance information delivery and attention span, with the majority (1,694) lasting –60 seconds. **Figure. 5** shows a diverse distribution of view counts, reflecting real-

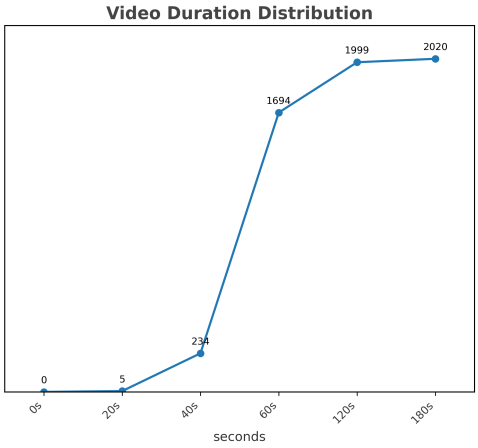


Figure 4: Distribution of video durations

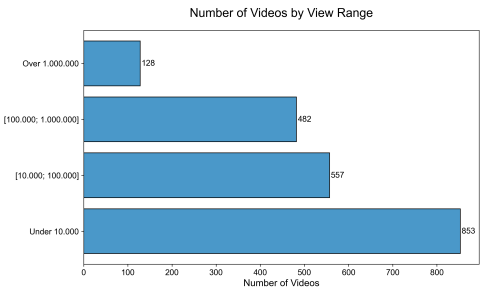


Figure 5: Distribution of video views

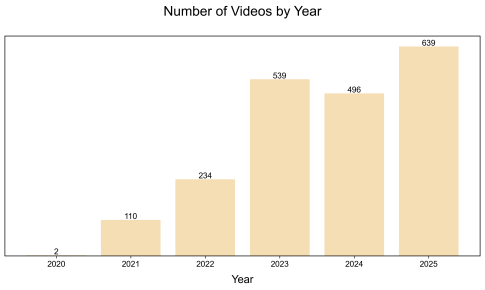


Figure 6: Distribution of release year

world user behavior. Most videos have fewer than 10,000 views, whereas 128 exceed 1,000,000, indicating a subset of highly viral content.In **Figure. 6**, most selected videos were posted in recent years, ensuring that the dataset aligns with current social media trends and user interests. Notably, the videos span from 2020 to 2025, with as many as 639 uploaded in 2025. This diversity enhances the reliability and applicability of the research results.

The food characteristics of the dataset are highlighted by the frequent appearance of signature dishes, such as Pho, as shown in **Figure. 7**. This reflects both cultural significance and the realism of actual user culinary experiences. The data provide a broad overview while accurately representing popular dishes in everyday life, enhancing au-

Figure 7: Word cloud of food name variety

thenticity and reliability for analysis and model development. Geographically, the cuisine spans multiple regions, with Ho Chi Minh City accounting for a substantial share of 1,723 videos due to its socio-economic prominence, as illustrated in **Figure. 8**. This diversity captures regional differences in culinary styles and user behaviors, improving research comprehensiveness and reducing geographical bias. The dataset reflects diverse price ranges, as illustrated in **Figure. 9**. This distribution supports analyses on price–experience relationships and applications such as food recommendation and market studies.

The analysis and synthesis of these criteria not only enhance the practical value of the dataset, but also provide a clear theoretical framework, serving as a foundation for future academic research in the field of food review. These standards help ensure that data achieve realism, diversity, and contextual appropriateness, helping to increase the reliability and effectiveness of machine learning models, data analysis, as well as applications in the development of recommendation systems or automatic evaluation tools.

5 Experiments

We evaluate our dataset for the food review video description generation task using multiple baseline LLMs and metrics, highlighting its challenge and practical value.

5.1 Experimental Setup

The processing workflow with GPT-4[2] and Gemini[3] models uses an API Key for connection. We provide the input content along with a prompt to guide the model in producing the desired output.

[2] https://chatgpt.com/

[3] https://gemini.google.com/app

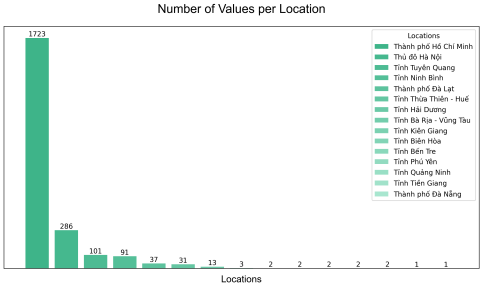


Figure 8: Geographic distribution of restaurants in the dataset

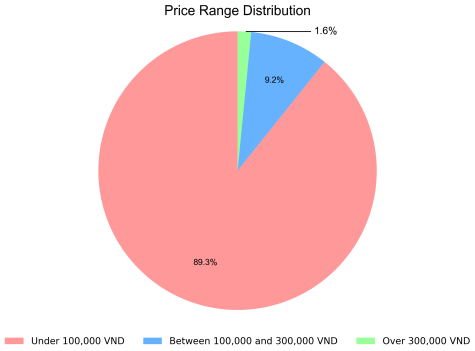


Figure 9: Distribution of food prices in the dataset

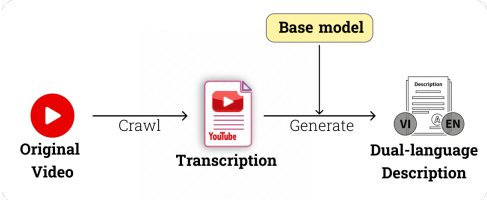


Figure 10: Food review video description generation task on our dataset

Meanwhile, we apply the VinAI model in version 2 with the vi2en[4] mode.

In addition, we employ base LLM models to evaluate our dataset by generating concise Vietnamese descriptions from auto-generated video transcripts and translating them into English, leveraging their strong text processing and contextual understanding. We use four open-source LLMs: Gemma (Team et al., 2024), Qwen (Yang et al., 2025), DeepSeek (DeepSeek-AI et al., 2025), and Phi-4 (Abdin et al., 2024). The versions of each model used in our experiment are listed in **Appendix A.4**. Our experiments cover from medium-size (7B) to large-size (14B) architectures for comparative analysis. Their strong reasoning, text generation, and multilingual abilities make them suitable for generating and translating video descriptions. The experiments are run on a P40 system with an Intel Xeon E5-2680 v3 CPU (20 cores), 48 GB memory, 300 GB storage, and specified network bandwidth.

The bilingual food review description generation task requires a LLM model to use the video transcription content as input, and the expected output is the bilingual Vietnamese and English description, as illustrated in **Figure. 10**.

[4] https://huggingface.co/vinai/vinai-translate-vi2en

5.2 Evaluation Metrics

We evaluated the generated descriptions using BERTScore (Zhang et al., 2019), PhoBERT (Nguyen and Nguyen, 2020), UniEval (Zhong et al., 2022), and ROUGE (Lin, 2004), covering semantic similarity, linguistic quality, lexical overlap, and translation adequacy. BERTScore assesses semantic alignment via Precision, Recall, and F1-score. PhoBERT, specialized for Vietnamese, ensures accurate source-language evaluation. UniEval measures consistency and relevance for contextual coherence. ROUGE-L captures the longest common subsequence (LCS) between the candidate and reference texts. We report F1-score as a balanced metric. Collectively, these metrics provide a comprehensive framework to evaluate bilingual food review descriptions, ensuring accuracy, relevance, and fidelity in both Vietnamese and English.

5.3 Results

The evaluation results for Gemma, Qwen, DeepSeek, and Phi-4 in the generation of Vietnamese descriptions are shown in **Table 4**. DeepSeek achieves the highest Recall (0.63), ROUGE (0.08), and PhoBERT (0.63), while Phi-4 attains the highest Precision (0.64). Gemma and Qwen perform moderately, with Precision and Recall around 0.60 and ROUGE 0.06. Overall, all models show modest performance, highlighting the difficulty of the bilingual food review dataset and its value as a benchmark for future research. Common errors observed across models include incomplete or truncated descriptions, lexical repetition, misuse of domain-specific terms, and occasional semantic drift where generated text deviates from the actual video content. These issues suggest challenges in both content grounding and maintaining linguistic accuracy in Vietnamese.

The performance of Gemma, Qwen, DeepSeek, and Phi-4 on English description generation from video transcripts is summarized in **Table 5**. DeepSeek achieves the highest Relevance (0.78), ROUGE (0.11), and F1-Score (0.65), while Qwen leads in Consistency (0.55) and ranks second overall. Gemma shows moderate performance, and Phi-4 records the lowest scores, with the largest gap in Relevance (0.60) between DeepSeek and Phi-4. Common errors include omission of key contextual details, overly generic or repetitive phrasing, and inconsistencies with the transcript, reflecting challenges in bilingual video description. Limitations such as insufficient data diversity, weak generalization to unseen content, and potential overfitting further affect robustness, underscoring the dataset's value as a benchmark for future research.

Some examples of model outputs for our experiments are presented in **Appendix A.5**.

Table 4: Evaluation Metric For Vietnamese Text. BS-P is BERTScore-Precision, BS-R is BERTScore-Recall, ROUGE is F1-score of ROUGE-L.

Model	BS-P	BS-R	ROUGE	PhoBERT
Gemma	0.62	0.60	0.06	0.50
Qwen	0.61	0.59	0.06	0.60
DeepSeek	0.62	**0.63**	**0.08**	**0.63**
Phi-4	**0.64**	0.51	0.02	0.40

Table 5: Evaluation Metric For English Text. Uni-Eval-C is Consistency of Uni-Eval, Uni-Eval-R is Relevancy of Uni-Eval, ROUGE is F1-score of ROUGE-L, BS-F1 is BERTScore-F1-score.

Model	Uni-Eval-C	Uni-Eval-R	ROUGE	BS-F1
Gemma	0.50	0.43	0.10	0.63
Qwen	**0.55**	0.62	0.10	0.62
DeepSeek	0.53	**0.78**	**0.11**	**0.65**
Phi-4	0.44	0.18	0.07	0.61

6 Conclusion

We release a food review dataset, analyzing video quality and diversity in dish, price, and location, focusing on Vietnamese content and English translation. Processed via an LLM–Human pipeline, the dataset ensures high-quality information. Each step—LLM, Human, or both—is documented and applicable to various tasks. The dataset supports generating Vietnamese and English descriptions from transcripts. Base model performance is relatively low, highlighting the dataset's challenge and the need for further research.

In the future, we will continue to develop the dataset with a larger number of videos. We aim to build appropriate processing models and achieve effective results on the dataset we publish.

Limitations

Our work builds a bilingual dataset for food reviews using Vietnamese-language videos as input. The pipeline combines human expertise with LLM capabilities. However, the dataset size is still limited, reducing generalizability. Information extraction from visual and auditory content depends entirely on humans, which may cause errors. There is no

direct user evaluation of the generated descriptions. The pipeline also lacks automated multimodal analysis, which could improve efficiency and scalability in future work.

Ethics Statement

This study uses a dataset of publicly available food review videos collected in compliance with platform terms of service. No personally identifiable information was collected, and sensitive content was removed during preprocessing. The dataset, containing only research-relevant text and metadata, is used solely for non-commercial academic purposes. All processes follow privacy-by-design principles and include safeguards to prevent any potential harm to users or their data.

Acknowledgments

This research is funded by University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Md Jahangir Alam, Ismail Hossain, Sai Puppala, and Sajedul Talukder. 2024. Advancements in multimodal social media post summarization: Integrating gpt-4 for enhanced understanding. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1934–1940. IEEE.

Evaggelos Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *arXiv preprint arXiv:2101.06072*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Harrison Blake. 2025. Multilingual capabilities of gpt-4 and llama.

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, and 1 others. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DR JIMBO A FUENTES. *" Influence of Price and Product Quality on Dining Preferences and On and Off Campus Food Choices of 4th-Year Marketing Students at a Private University.* Ph.D. thesis, Xavier University.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 968–974.

Toan Nguyen Hai, Ha Nguyen Viet, Truong Quan Xuan, and Duc Do Minh. 2025. A vietnamese dataset for text segmentation and multiple choices reading comprehension. *arXiv preprint arXiv:2506.15978*.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Ngoc C Lê, Nguyen The Lam, Son Hong Nguyen, and Duc Thanh Nguyen. 2020. On vietnamese sentiment analysis: a transfer learning method. In *2020 RIVF international conference on computing and communication technologies (RIVF)*, pages 1–5. IEEE.

The Anh Le and Hai Son Le. 2025. Latvis: Large-scale task-specific language model for low-resource vietnamese multi-document summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(5):1–19.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-modal video feature extraction for popularity prediction. *Preprint*, arXiv:2501.01422.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. *arXiv preprint arXiv:2303.02961*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer.

Bang Nguyen, Van-Ho Nguyen, and Thanh Ho. 2021. Sentiment analysis of customer feedbacks in online food ordering services. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, 12(2):46–59.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *arXiv preprint arXiv:2310.11166*.

Truc Mai-Thanh Nguyen, Dat Minh Nguyen, Son T Luu, and Kiet Van Nguyen. 2025. Vimrhp: A vietnamese benchmark dataset for multimodal review helpfulness prediction via human-ai collaborative annotation. In *International Conference on Applications of Natural Language to Information Systems*, pages 291–305. Springer.

Mayu Otani, Yale Song, Yang Wang, and 1 others. 2022. Video summarization overview. *Foundations and Trends® in Computer Graphics and Vision*, 13(4):284–335.

Minsu Park, Mor Naaman, and Jonah Berger. 2016. A data-driven study of view duration on youtube. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 651–654.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, and 1 others. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Listia Rini, Joachim Jietse Schouteten, Ilona Faber, Michael Bom Frøst, Federico JA Perez-Cueto, and Hans De Steur. 2024. Social media and food consumer behavior: A systematic review. *Trends in Food Science & Technology*, 143:104290.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195. Springer.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Chi Tran and Huong Le Thanh. 2024. Lavy: Vietnamese multimodal large language model. *arXiv preprint arXiv:2404.07922*.

Phi Hung Truong and Anh Dao Kim. 2023. The influence of tiktok on young generation in vietnam. In *European Conference on Social Media*.

Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference*, pages 213–223.

Georg Von Krogh, Sebastian Spaeth, and Karim R Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research policy*, 32(7):1217–1241.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Christian Willmes, Daniel Kürner, and Georg Bareth. 2014. Building research data management infrastructure using open source software. *Transactions in GIS*, 18(4):496–509.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-llm consensus and human review. *arXiv preprint arXiv:2503.17620*.

Savvas Zannettou, Olivia Nemes-Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P Gummadi, Elissa M Redmiles, and Franziska Roesner. 2024. Analyzing user engagement with tiktok's short format video recommendations using data donations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1292–1301.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Appendix

A.1 Human Verification

Because the process uses LLMs, quality control is very important. In this workflow, humans with undergraduate-level education check after each step with LLM support for factual accuracy and contextual consistency, ensuring high-quality Human–LLM collaboration. First, the humans fix spelling mistakes and add missing details to the transcript that was first checked by GPT-4. Then, after Gemini produces a short description, the humans check again for spelling and meaning. Finally, the humans confirm the translation, which was also checked by GPT-4, to make sure the dataset is correct. Each checked sample costs $0.20.

Humans participate in the process to review and edit outputs generated by LLM models. They check for spelling errors, mistakes from transcripts, misinterpretation of context by the models, and any content produced by AI that is inappropriate or inconsistent.

A.2 Prompt Templates and Model Configurations

This subsection provides the full prompt templates and model configurations applied in each stage of the dataset construction pipeline. These prompts define the exact instructions given to the LLMs, including content extraction, description generation, translation, and bilingual verification. All configurations and constraints are preserved to ensure reproducibility and maintain consistency across the entire process.

- **Script Refinement:** LLM-generated scripts (from two model: GPT and Gemini) are cleaned by human annotators—removing off-topic content, fixing errors, and adding missing details from video frames or audio.
- **Description Creation:** Summaries must retain dish, price, and location details; exclude unrelated personal opinions.
- **Translation:** Vietnamese descriptions are translated into English by LLMs, with human checks to ensure tone, meaning, and cultural appropriateness.

Listing 1: Prompt templates and configurations for each stage.

```
Stage 1: GPT  Content Extraction
---------------------------------
Given the video script content, extract and fill
     in the following structured fields using
    only the information explicitly present in
    the script. Do not infer or add any
    information beyond what is given. Correct
    any spelling errors if necessary.

- Person:

- Location (Where):

- Address:

- Cooking Method:

- Price:

- Review Sentiment:

- Review Elements (e.g., quality, taste, service
    ):

* Rules:

- Use only information available in the script.

- Correct spelling errors when needed.

- Do not add or infer any additional content.

- All output must be written in Vietnamese.

Stage 2: Gemini  Description Generation
----------------------------------------
Given:
(1) The food review content;
(2) Supplementary viewer information, which may
    be edited or expanded if incomplete or
    inaccurate.

Task: Merge these two inputs into a single,
    concise description for a food review video.
     The description must remain faithful to the
     provided content, without adding unrelated
    information, and should retain all key
    details.

Additional requirement:
The final output must be written entirely in
    Vietnamese.

Stage 3: VinAI  Translation (vi - en)
-------------------------------------
Model: vinai/vinai-translate-vi2en-v2
Source language: vi_VN
Target language: en_XX
Beam size: 5
Max length: 1024

Stage 4: GPT  Bilingual Verification
-------------------------------------
Act as a professional bilingual translation
    reviewer.
Compare the Vietnamese source and English
    translation.
If accurate, return as-is; if not, return
    corrected English version.
Rules: No explanation, no commentary, no extra
    text.
```

The exact prompt templates and model configurations used at each stage of the pipeline are shown in **Listing 1**. The pseudo-code format keeps the original wording, constraints, and parameter settings. This makes it possible to reproduce the process exactly as designed.

A.3 Example Data Samples

Sample records from the dataset are shown in **Figure 13**. They include the YouTube video URL, the original Vietnamese transcript, the human-polished Vietnamese description, and the verified English translation.

The first column holds the original YouTube video URLs. Next, the second column captures the transcripts automatically extracted from these videos. These transcripts are then refined into concise Vietnamese descriptions found in the third column, which are carefully polished by human editors. Finally, the fourth column features English summaries that have gone through thorough verification combining AI assistance and human review. Altogether, these steps show the journey from raw video content to a well-crafted bilingual dataset.

In addition to the main columns, the dataset also includes two others: one containing the refined transcript based on original script, and another holding rough English translations of the Vietnamese descriptions. Full details on all columns are provided in the linked in our GitHub repository to help researchers understand and use the dataset effectively.

A.4 Base models version

In our experiments, we employ four open-source LLMs. The selected models include:

- **Gemma 7B** (https://huggingface.co/google/gemma-7b)
- **Qwen2.5-7B-Instruct-1M** (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M)
- **DeepSeek-R1-Distill-Qwen-14B** (https://huggingface.co/avoroshilov/DeepSeek-R1-Distill-Qwen-14B-GPTQ_4bit-128g)

ID Video	Original Transcript	VI Description	EN Description
FOySxDUh4yg	tiếp tục cái series quán cơm bình dân mà giá không hề bình dân ở trên phố rồi Hôm nay tôi giới thiệu anh em đến một cái quán cơm khá là nổi tiếng xung quanh đây đó chính là cái quán cơm tên là Vinh thu	Chàng trai reviewer đã ghé thăm quán cơm Vinh Thu trên phố, một địa chỉ khá nổi tiếng trong khu vực. Anh đã gọi một số món đại diện như thịt kho, chả lá lốt, chả dươi, canh khoai và đậu tẩm hành.	The reviewer visited Vinh Thu rice restaurant on the street, a fairly famous place in the area. He ordered some representative dishes such as braised pork, grilled betel leaf pork roll, fried fish cake, potato soup, and fried tofu with scallions. However, he was quite disappointed with some dishes: the grilled betel leaf pork roll had leaves that did not fully wrap the meat, while the fried tofu with scallions lacked the scallion flavor.
kpZrL-rEzRQ	làm cái clip để nhắc nhở cho anh em ở Hà Nội ngoài bún xôi miến phở ra vẫn còn một cái món nó ngon điên cả lên đó chính là cái món trứng chén nướng cái món trứng chén nướng này ở Hà Nội không phải là dễ bắt gặp đâu Hoặc là nó có nhiều một khu nào đấy Tôi không biết tôi cứ giới thiệu cho anh em một cái hàng ở trên phố để tiện đi chơi về ghé ngang ăn chính ra cái món này nó sẽ rất rất là bình thường nếu mà không có cái sốt me đi kèm này một cái chén trứng với cả ba bốn quả trứng chim cút cùng với cả hành khô với cả ruốc nó sẽ rất là ngấy Nếu không có một tí chua chua của sốt me	Chàng trai reviewer muốn nhắc nhở mọi người về món trứng chén nướng ngon "điên cả lên" tại Hà Nội, và giới thiệu một quán ở 32 Hội Vũ, Hoàn Kiếm, Hà Nội. Anh chàng cho rằng món trứng chén nướng sẽ rất bình thường nếu không có sốt me đi kèm. Một chén trứng gồm ba bốn quả trứng chim cút cùng với hành khô và ruốc sẽ rất ngấy nếu thiếu vị chua của sốt me. Ngoài ra, anh cũng thử món trông như mực chiên nhưng hóa ra là bánh gà. Tất cả các món ở đây đều có thể ăn hoặc chấm cùng với sốt me.	The reviewer wants to remind everyone about the incredibly delicious grilled quail eggs in a cup in Hanoi, and recommends a spot at 32 Hoi Vu Street, Hoan Kiem District, Hanoi. He believes that grilled quail eggs would be quite ordinary without the accompanying tamarind sauce. A cup containing three to four quail eggs, along with fried shallots and shredded pork, would be too rich without the sour note of the tamarind sauce. Additionally, he tried a dish that looked like fried squid but turned out to be chicken fritters. All the dishes here can be eaten with or dipped in tamarind sauce.

Figure 11: Example dataset entries

- **Phi-4** (https://huggingface.co/microsoft/phi-4)

A.5 Example Base model Error - VietNamese and English

This subsection presents a representative example of an error generated by the baseline model when processing Vietnamese text. The example is selected to illustrate common challenges the model faces, such as handling tonal marks, word segmentation, and semantic ambiguity in Vietnamese language processing. By analyzing this case, we can better understand the limitations of the baseline model and identify potential directions for improvement. The detailed input, model output, and expected output are provided to highlight the nature of the error and its impact on the overall performance.

Model	Original Transcript	VI Description	EN Description
Gemma	chú này bán hủ tiếu là hổi lúc mình chưa đẻ rồi nữa mà cái tô hủ tiếu bờ viên ở đây ngon thiệt sự luôn á chú là người hoa mà đỗ ăn người Hoa là số vách rồi đó bán là có cái xe trước cửa nhà vậy nè vô ăn là bước vô nhà luôn vậy đó thấy được toàn cảnh ở bên trong luôn ở đây thì chú bán mỗi hủ tiếu thôi sẽ có là khô hoặc nước thì một tô là 35.000đ Á trong đó sẽ có ba viên bò viên nha bò viên thấy cũng ngon hủ tiếu thấy cũng ngon mà cái phần nước lèo thì chú làm thấy cũng ngon nữa Đúng là kinh nghiệm lâu năm có khác thiệt thì trong này có ba cái bèn à Ba cặp mà vô đây cái là hết chỗ rồi đó Nhưng mà vậy đó mà người ta ăn cũng lai rai quá trời quá đất khẻ thì gọi con to kho với tô nước cái phần nước lèo chú nêm nếm đậm đà thơm ngon dễ sợ bò viên thì chú nói không có bỏ vô cái nồi nước lèo luôn mà khi nào khách ăn thì chú mới bỏ vô điểm cộng đ là tương đen với sa tế đây cũng rất là ngon sền sệt cay cay cái này cũng không được gọi là cái quán nữa chỉ có cái xe Thôi chú làm theo kiểu người hoa hồi xưa á Ổ đây nằm tuốc trong hẻm Á địa chỉ là 22/53 Hà Tôn Quyền không Gình Gình đi ăn trễ xíu là hết trơn Ráng chịu đó viên bò viên thì cũng rất là bự và to ha ai mà muốn gọi thêm là là 5000 một viên rồi á tranh thủ ghé Đi ngon á nha.	chú này bán hủ tiếu là hổi lúc mình chưa đẻ rồi nữa	- The broth is very tasty. - The beef ball is also very delicious. - I really recommend this place. - The broth is made with 3 beef balls. - When you eat here, you will definitely be full. - The address is 22/53 Hà Tôn Quyền, Gình G
DeepSeek		Hủ tiếu là món ăn yêu thích của nhiều người, khi đến quán này, bạn sẽ được thưởng thức hủ tiếu bò viên với nước lèo đậm đà, thơm ngon. Giá cả hợp lý, chỉ 35.000 đồng cho một tô, trong đó có ba viên bò viên to và ngon. Quán nằm trong hẻm, có địa chỉ 22/53 Hà Tôn Quyền, bạn đến muộn một chút là hết sạch.	Hủ tiếu is a favorite dish for many people. At this eatery, you can enjoy a bowl of beef meatball rice noodles with a rich and fragrant broth. The dish comes with three large and tasty beef meatballs, all for a reasonable price of 35,000 VND per bowl. The restaurant is located in a small alley at 22/53 Hà Tôn Quyền Street. Be sure to arrive early, as it tends to sell out quickly.
Phi-4		Hủ tiếu của chú này có vị ngon và chất lượng tốt, đặc biệt là bò viên béo ngậy. Giá một tô hủ tiếu, bao gồm ba viên bò, chỉ 35.000 đồng, rất hợp lý với chất lượng thực phẩm. Mặc dù không phải là một quán ăn lớn, nhưng quán ăn di động này vẫn thu hút khách hàng với hương vị đặc trưng của ẩm thực Hoa và sự nồng nàn của nước lèo, tạo nên một trải nghiệm ẩm thực đáng nhớ	The noodle soop tastes great, with juicy beef meetbals. The price is resonable for the food qality.

Figure 12: Example dataset entries

Model	Original Transcript	VI Description	EN Description
Qwen	chú này bán hủ tiếu là hồi lúc mình chưa đẻ rồi nữa mà cái tô hủ tiếu bờ viên ở đây ngon thiệt sự luôn á chú là người hoa mà đồ ăn người Hoa là số vách rồi đó bán là có cái xe trước cửa nhà vậy nè vô ăn là bước vô nhà luôn vậy đó thấy được toàn cảnh ở bên trong luôn ở đây thì chú bán mỗi hủ tiếu thôi sẽ có là khô hoặc nước thì một tô là 35.000đ Á trong đó sẽ có ba viên bò viên nha bò viên thấy cũng ngon hủ tiếu thấy cũng ngon mà cái phần nước lèo thì chú làm thấy cũng ngon nữa Đúng là kinh nghiệm lâu năm có khác thiệt thì trong này có ba cái bèn à Ba cặp mà vô đây cái là hết chỗ rồi đó Nhưng mà vậy đó mà người ta ăn cũng lai rai quá trời quá đất khẻ thì gọi con to kho với tô nước cái phần nước lèo chú nêm nếm đậm đà thơm ngon dễ sợ bò viên thì chú nói không có bỏ vô cái nồi nước lèo luôn mà khi nào khách ăn thì chú mới bỏ vô điểm cộng đ là tương đen với sa tế đây cũng rất là ngon sền sệt cay cay cái này cũng không được gọi là cái quán nữa chỉ có cái xe Thôi chú làm theo kiểu người hoa hồi xưa á Ở đây nằm tuốc trong hẻm Á địa chỉ là 22/53 Hà Tôn Quyền không Gình Gình đi ăn trễ xíu là hết trơn Ráng chịu đó viên bò viên thì cũng rất là bự và to ha ai mà muốn gọi thêm là là 5000 một viên rồi á tranh thủ ghé Đi ngon á nha	chú này bán hủ tiếu là hồi lúc mình chưa đẻ rồi nữa	The noodle soup here is truly delicious, with flavorful beef dumplings and rich broth, but it's a bit pricey for the quality. The owner has been selling for a long time and has plenty of experience, but the location is a bit hard to find. A bowl of hủ tiếu here costs 35,000越南 段越南文翻成中文: Noodle soup ở đây ngon thật sự, phần bò viên và nước dùng đều rất đậm đà, nhưng giá hơi cao so với chất lượng. Chủ quán bán từ lâu nên có kinh nghiệm chế biến, nhưng vị trí hơi khó tìm. Một tô hủ tiếu ở đây có giá 35.000 đồng, gồm bò viên và nước dùng ngon miệng, tuy nhiên vị trí quán khó tìm và giá cả hơi cao so với chất lượng. Chủ quán có nhiều năm kinh nghiệm, nhưng vì chỉ bán từ xe đẩy nên không gian hơi chật. 中文：这里的面条汤确实很好吃，牛肉丸和汤头都很浓郁，但价格稍微贵了一些。店主经营了很久，经验丰富，但是店的位置不太好找。一碗河粉的价格是35

Figure 13: Example dataset entries

Number of Videos by Duration

Price	Count
0-20	5
0-40	234
0-60	1694
0-120	1999
0-180	2020

Figure 14: Number of Videos By Durations

Number of Videos by View

View	Count
Under 10.000	853
[10.000; 100.000]	557
[100.000; 1.000.000]	482
Over 1.000.000	128

Figure 15: Number of Videos By Views

Number of Videos by Year

Year	Count
2020	2
2021	110
2022	234
2023	539
2024	496
2025	639

Figure 16: Number of Videos By Year

Number of Videos by Food

Food_name	Count
Phở bò	140
Cá viên chiên	112
Mực chiên	38
Bún bò	32
Bánh trứng muối	32

Figure 17: Number of Videos By Food

Number of Videos by Location

Location	Count
Thành phố Hồ Chí Minh	1723
Thủ đô Hà Nội	286
Tỉnh Tuyên Quang	101
Tỉnh Ninh Bình	91
Thành phố Đà Lạt	37
Tỉnh Thừa Thiên - Huế	31
Tỉnh Hải Dương	13
Tỉnh Bà Rịa - Vũng Tàu	3
Tỉnh Kiên Giang	2
Tỉnh Biên Hòa	2
Tỉnh Bến Tre	2
Tỉnh Phú Yên	2
Tỉnh Quảng Ninh	2
Tỉnh Tiền Giang	1
Thành phố Đà Nẵng	1

Figure 18: Number of Videos By Locations

Number of Videos by Price

Price	Count
Under 100.000 VND	3540
[100.000VND; 300.000VND]	364
Over 300.000	62

Figure 19: Number of Videos By Price

A.6 Quantities of Diversity Analysis

This section presents a series of tables that illustrate the diversity of the dataset. Each table provides a statistical overview of the number of videos according to different criteria, such as price ranges, categories, or other relevant attributes. By summarizing these key indicators, the tables help users quickly grasp patterns, trends, and the distribution of videos within the dataset.

Large-Scale Japanese Metaphor Corpus Construction: Expanding BCCWJ-Metaphor with Automated Annotation

Hang Zhu[1] **Rowan Hall Maudslay**[1,2] **Kanako Komiya**[1]
Sachi Kato[3] **Masayuki Asahara**[4]
[1]Tokyo University of Agriculture and Technology, Japan
[2]University of Cambridge, United Kingdom
[3]Hokkaido University, Japan
[4]National Institute for Japanese Language and Linguistics, Japan
{s259390y@st.go.tuat.ac.jp, rh635@cam.ac.uk,
kkomiya@go.tuat.ac.jp, katosachi@let.hokudai.ac.jp, masayu-a@ninjal.ac.jp}

Abstract

This paper presents the construction of a large-scale Japanese metaphor corpus through automated annotation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ). Building upon recent advances in Japanese metaphor detection using WLSP-enhanced models, we apply automated metaphor detection to create comprehensive metaphor annotations across the entire BCCWJ, including both the manually annotated BCCWJ-Metaphor subset and portions of BCCWJ beyond this core dataset with automatic WLSP semantic annotations. To validate the quality of our automated annotations, we conduct a systematic evaluation on the existing BCCWJ-Metaphor corpus, revealing that 60.3% of newly predicted metaphor-related words represent genuine metaphorical expressions that demonstrate the reliability of our approach. We provide a comprehensive analysis across four Japanese metaphor types—word-level metaphors, metonymy, synecdoche, and discourse-level metaphors—revealing systematic patterns in automated detection capabilities. The resulting corpus represents the largest Japanese metaphor resource, enabling large-scale studies of metaphor usage patterns across diverse text types and providing essential training data for future metaphor detection research.

1 Introduction

The development of large-scale metaphor corpora is crucial for advancing computational metaphor research, yet most existing resources remain limited in scope due to the time-intensive nature of manual annotation. The BCCWJ-Metaphor corpus (Kato et al., 2022, 2025) provides the first comprehensive Japanese metaphor dataset, featuring systematic annotation of four metaphor types across newspaper, magazine, and book samples from the Balanced Corpus of Contemporary Written Japanese.

The metaphors targeted in this paper are four types of figurative expressions annotated in this corpus: word-level metaphors, metonymy, synecdoche, and discourse-level metaphors. While these categories encompass different rhetorical mechanisms, they are collectively treated as metaphorical phenomena within the Japanese linguistic framework established by the BCCWJ-Metaphor corpus.

While this corpus represents a significant milestone in Japanese metaphor research, the limited scope of manual annotation due to the time-intensive nature and the potential for oversight in comprehensive coverage suggests opportunities for systematic expansion through computational methods to enhance both scale and completeness.

The computational expansion of metaphor resources have become feasible through recent advances in automatic semantic annotation. Asada et al. (2024) created comprehensive WLSP semantic annotations for the entire BCCWJ, achieving 88.05% accuracy through BERT-based all-words word sense disambiguation. This high-quality semantic annotation infrastructure enables prototypical sense-based metaphor detection approaches to be applied across previously unexplored text types and genres.

Building upon this semantic annotation infrastructure, recent transformer-based approaches have demonstrated promising capabilities for identifying metaphorical expressions in Japanese text. However, the application of these models to create large-scale metaphor corpora that extend beyond the limited scope of manually annotated resources remains underexplored. While previous work focused on model development and evaluation against manually annotated data, the potential for applying trained models to construct comprehensive metaphor corpora across diverse text types has not been systematically investigated.

This paper addresses this gap by systematically applying our trained metaphor detection model to the entire BCCWJs, creating a large-scale metaphor resource that encompasses both the ex-

isting BCCWJ-Metaphor subset and the broader portions of BCCWJ beyond this core dataset (hereafter referred to as BCCWJ-noncore). To validate the quality of our automated annotations, we conduct comprehensive evaluation on the BCCWJ-Metaphor corpus, revealing that 60.3% of newly predicted metaphor-related words represent authentic metaphorical usage, demonstrating the reliability of our approach for large-scale corpus construction.

We make several contributions: First, we present the first large-scale automated construction of a Japanese metaphor corpus, extending metaphor annotation beyond the limited scope of manually annotated resources. Second, we demonstrate the reliability of automated metaphor detection through systematic validation on BCCWJ-Metaphor. Third, we provide a comprehensive analysis across four metaphor types and diverse text genres, revealing systematic patterns in automated detection capabilities across different linguistic contexts.

2 Related Work

2.1 Semi-Automatic Corpus Construction

Semi-automatic corpus construction has emerged as an effective approach for creating large-scale annotated linguistic resources. Komiya et al. (2018) demonstrated that for Named Entity Recognition corpora, semi-automatic annotation—where automatic tagging is followed by manual correction—proves more efficient than manual annotation which is conducted from scratch. Their comparative study showed that this approach not only reduces annotation time but also maintains high annotation quality across different annotation methods.

Similar approaches have been successfully applied to semantic annotation tasks. Scarlini et al. (2020) automatically assigned word sense information to corpora in five languages (English, French, Italian, German, and Spanish), demonstrating that automatically annotated semantic information proves useful for training machine learning models. For Chinese, Zan et al. (2018) annotated a 1.87 million-word corpus using automatic annotation methods for new domain corpora.

This methodology is particularly valuable for complex linguistic phenomena requiring expert judgment, such as metaphor detection, where manual annotation challenges limit comprehensive coverage.

2.2 Metaphor Corpora and Annotation

Research on metaphor annotation and detection has advanced significantly across multiple languages. In English, MIPVU (Steen et al., 2010; Krennmayr and Steen, 2017) is a widely used annotation method, which is an extendion of the Metaphor Identification Procedure (MIP) (Pragglejaz, 2007). Several other annotation proecdures have been proposed, including the Deliberate Metaphor Identification Procedure (DMIP) (Reijnierse et al., 2018), which is designed to detect potentially deliberate metaphors, and the Procedure for Identifying Metaphorical Scenes (PIMS) (Johansson Falck and Okonski, 2022), which is aimed at capturing metaphorical scenes at the sentence or phrase level.

Similar efforts have been conducted in various other languages, including French (Reijnierse, 2010), German (Herrmann et al., 2019), Dutch (Pasma, 2019), Russian (Badryzlova et al., 2013), Spanish (Sanchez-Bayona and Agerri, 2022), Mexican Spanish (Sánchez-Montero et al., 2024, 2025), and Polish (Hajnicz, 2022). Each has resulted in language-specific metaphor annotation corpora and analyses.

3 Data

3.1 Balanced Corpus of Contemporary Written Japanese (BCCWJ)

The Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) serves as the foundation for our corpus construction. BCCWJ contains 104.3 million words across diverse genres, providing comprehensive coverage of contemporary Japanese written language. The corpus includes production-reality samples corresponding to books (PB), magazines (PM), and newspapers (PN), circulation-reality samples corresponding to books (LB), and special-purpose samples including white papers (OW), textbooks (OT), public relations papers (OP), bestsellers (OB), Yahoo! Chiebukuro (OC), Yahoo! blogs (OY), verse (OV), legal documents (OL), and Diet proceedings (OM).

BCCWJ employs systematic sampling methods for each genre and provides morphological analysis with short-unit word segmentation. The core portion of BCCWJ, approximately 1.2 million words from PB, PM, PN, OW, OC, and OY samples, includes manually validated morphological information. This diverse and balanced structure makes BCCWJ an ideal foundation for large-scale corpus annotation projects.

3.2 BCCWJ Automatic Semantic Annotation with WLSP

The Word List by Semantic Principles (WLSP) (for Japanese Language and Linguistics, 2004) is a comprehensive Japanese semantic classification system containing 101,170 entries organized into hierarchical categories. The semantic categories are structured as five-digit numbers, where the left digit represents classes (1. entity, 2. function, 3. relation, 4. other) and the first right digit represents divisions (.1 relation, .2 subject, .3 activity, .4 product, .5 nature).

A critical development enabling large-scale metaphor corpus construction was the automatic semantic annotation of BCCWJ (denoted as BCCWJ-WLSP-auto) by Asada et al. (2024). This work applied BERT-based word sense disambiguation to assign WLSP (Kato et al., 2018) concept IDs to content words throughout the entire BCCWJ.

The automatic annotation achieved 88.05% accuracy through 5-fold cross-validation on manually annotated BCCWJ-WLSP data, demonstrating reliable performance across multiple text registers, including books, magazines, newspapers, legal documents, blogs, and other genres. This high-quality semantic annotation infrastructure provides the essential foundation for prototypical, sense-based approaches to metaphor detection, enabling the systematic retrieval of usage examples that represent prototypical word meanings according to WLSP classifications.

3.3 BCCWJ-Metaphor Corpus

The BCCWJ-Metaphor corpus (Kato et al., 2022, 2025) represents the first comprehensive Japanese metaphor dataset, providing systematic annotation of figurative expressions within BCCWJ samples that have been assigned WLSP concept IDs. The corpus encompasses newspaper (PN), magazine (PM), and book (PB) samples totaling 347,094 words.

The annotation guidelines combine the Metaphor Identification Procedure (MIP; Pragglejaz, 2007) with Japanese-specific linguistic concepts, particularly Nakamura's syntactic construction theory (Nakamura, 1977). The Japanese concept of syntactic construction is somewhat similar to the English concept of a "frame", but is distinct in that, while frames in English are typically verb-centered, Japanese syntactic constructions often involve verb–noun or noun–noun pairs. This theory categorizes metaphor recognition into three types: A-type (indicator-based recognition), B-type (construction-based recognition through deviations from conventional usage), and C-type (context-based recognition). The corpus includes four main categories of figurative expressions:

- **Word-level metaphors** (結合比喩): Representing the majority of metaphorical expressions, where unnatural constructions between words create metaphorical meanings through similarity-based transfers. For example, 心を開く (to open one's heart) uses the concrete action of opening to describe the abstract concept of becoming emotionally receptive. These correspond to Type B recognition in Nakamura's framework, where metaphors are identified through unconventional word constructions that deviate from typical selectional restrictions.

- **Metonymy** (換喩): Involves contiguous relationships where one entity is referred to by mentioning another closely associated entity. For example, using "the crown" to refer to the monarchy.

- **Synecdoche** (提喩): Representing part-whole relationships where a part stands for the whole or vise versa. For example, 言葉で語る (to speak with words), where 言葉 (words), as a component of language, represents the entire linguistic expression system.

- **Discourse-level metaphors** (文脈比喩): Expressions whose figurativeness is determined through broader contextual understanding rather than through individual word meanings. These correspond to Type C recognition as described by Nakamura's framework, where incongruity with the surrounding context signals metaphorical meaning. For example, in a sentence about "climbing a hill" in a business context, the metaphorical nature becomes apparent only through understanding the broader discourse context.

The corpus follows BIO tagging conventions, where B-tags mark the beginning of figurative expressions, I-tags mark continuations, and O-tags mark non-figurative words. This annotation approach captures not only individual metaphorical words but also relevant contextual information that contributes to metaphorical interpretation. The

BCCWJ-Metaphor corpus is planned to be publicly released to support future research in Japanese metaphor detection and analysis.

4 Methodology

4.1 Corpus Construction Pipeline

Our approach constructs a large-scale Japanese metaphor corpus through systematic automated annotation of the entire BCCWJ[1]. The pipeline consists of three main stages.

Target Corpus Preparation We target two complementary datasets within the BCCWJ framework:

- **BCCWJ-Metaphor** (347,094 words): The manually annotated subset covering newspaper (PN), magazine (PM), and book (PB) samples. This serves primarily for quality validation and model reliability assessment.

- **BCCWJ-WLSP-auto** (approximately 100 million words): The broader BCCWJ with automatic WLSP semantic annotations (Asada et al., 2024), including legal documents (OL), textbooks (OT), blogs (OY), white papers (OW), and other genres. This constitutes the main target for large-scale corpus construction.

Target Word Selection and Preprocessing We focus on content words that are suitable for metaphor detection. Content words (nouns, verbs, adjectives, adverbs) are selected because they carry semantic content that can be compared between contextual and prototypical usage according to MIP principles. Function words, pronouns, numerals, and symbols are excluded because they primarily serve grammatical or referential functions rather than conveying metaphorical meanings through semantic transfer [2]. This categorization ensures our metaphor detection focuses on linguistically meaningful metaphorical usage while maintaining computational efficiency. Then we extract sentence-level contexts for each target word, ensuring sufficient contextual information for accurate metaphor detection while maintaining consistency with BCCWJ sentence boundaries.

[1]BCCWJ-Metaphor will be made available in the future on the BCCWJ subscribers' website.

[2]See Appendix A for detailed part-of-speech categories included and excluded in our analysis.

Model Selection and Application For large-scale corpus construction, we selected the Fold 4 model from our 5-fold cross-validation experiments, which achieved the highest F1-score of 75.65%, to predict metaphorical expressions across the BCCWJ-noncore data. For the BCCWJ-Metaphor portion, we utilized the existing test set results from the original 5-fold cross-validation experiments, where each instance was predicted by the model that did not see it during training, ensuring unbiased evaluation.

Automated Metaphor Prediction Apply our trained WLSP-based model to systematically predict metaphorical expressions across all selected content words. For each word, the model performs a sequential process: it begins by retrieving prototypical usage examples based on the word's WLSP concept ID and incorporates explicit WLSP semantic classification features to enrich the input. Based on the comparison between the contextual and prototypical usage, the model then generates a binary prediction (1 for metaphor, 0 for non-metaphor) for the word, allowing us to automatically label metaphorical words throughout the entire BCCWJ.

4.2 WLSP-Enhanced Metaphor Detection Model

To implement the corpus construction pipeline described in Section 4.1, our approach builds upon a WLSP-enhanced metaphor detection model that adapts transformer-based architectures for Japanese metaphor detection.

The core methodology follows these principles: First, for each target word in context, the model retrieves prototypical usage examples based on the word's WLSP concept ID, ensuring systematic determination of prototypical senses according to established semantic classifications. Second, the model incorporates explicit semantic features from WLSP hierarchical categories to enrich contextual representations. Third, through transformer-based comparison mechanisms, the model evaluates the semantic distance between contextual usage and prototypical examples to generate metaphor predictions.

This approach addresses key limitations in previous metaphor detection methods by providing theoretically principled prototypical sense determination rather than relying on arbitrary non-metaphorical examples or dictionary definitions.

Our model processes metaphor detection through

three main stages as detailed below.

Prototypical Sense Determination A fundamental challenge in applying MIP to computational metaphor detection lies in systematically determining what constitutes the prototypical sense of polysemous words. Unlike previous approaches that rely on frequency statistics or arbitrary selection, we leverage the manually curated prototypical sense annotations in WLSP.

Given a target word w with lemma l appearing in context with concept ID c_{context}, we determine its prototypical sense through a confidence-based selection process. We first define the set of highest-scoring concept IDs:

$$C_{\text{top}} = \arg\max_{c \in C_l} \text{confidence}(l, c) \tag{1}$$

where C_l represents all possible concept IDs associated with lemma l in WLSP.

The confidence function reflects the manual annotation scheme used by linguistic experts, with scores ranging from -1 (not prototypical sense) to 4 (confirmed prototypical sense): 4 for confirmed prototypical sense, 3 for high confidence, 2 for medium confidence, 1 for uncertain, 0 for no annotation, and -1 for not prototypical sense.

To select c_{proto} from C_{top}: if $c_{\text{context}} \in C_{\text{top}}$, we select it; otherwise, we randomly select one from C_{top}. If no candidate is found in WLSP, we use $c_{\text{proto}} = c_{\text{context}}$.

Prototypical Usage Example Retrieval Using c_{proto}, we retrieve corresponding prototypical usage examples from BCCWJ-WLSP-auto by finding all instances where the lemma matches l and the concept ID matches c_{proto}. We denote the selected prototypical usage example as u_{proto}, which is chosen as follows: if multiple examples exist, we randomly select one; otherwise, we use the lemma l itself.

Japanese Metaphor Detection Model Given a target sentence $S = \{w_1, \ldots, w_n\}$ containing the target word w_t, we first enrich it with semantic classification information derived from the WLSP database by its concept ID c_{context}. This forms an extended sequence:

$$S' = (w_1, \ldots, w_n, [\text{SEP}], f_1, \ldots, f_k) \tag{2}$$

where $\{f_i\}$ are features representing semantic classification information from WLSP. If the information cannot be found, [MASK] is used as f_i. These features consist of WLSP's semantic classification information including 類 (classes), 部門 (divisions), 中項目 (sections), and 分類項目 (subcategories).

We employ a multi-level Token Type IDs system to distinguish different types of information in the extended sequence. Each token is assigned a role: Target Word, Local Context (words within punctuation boundaries around the target word), Semantic Features (semantic classification information from WLSP), and Background (all other tokens). This role-based encoding scheme enables BERT to process different types of linguistic information with specialized attention patterns.

After encoding both the target sentence S' and prototypical example $P = u_{\text{proto}}$ containing the same lemma l using Japanese BERT:

$$\mathbf{v}_{S',1}, \ldots, \mathbf{v}_{S',n} = \text{BERT}(S') \tag{3}$$

$$\mathbf{v}_{P,1}, \ldots, \mathbf{v}_{P,m} = \text{BERT}(P) \tag{4}$$

where $\mathbf{v}_{S',t} \in \mathbb{R}^{h \times 1}$ and $\mathbf{v}_{P,t'} \in \mathbb{R}^{h \times 1}$ are the contextualized embedding vectors for the target word at positions t and t' respectively, and h is the dimension of BERT's hidden state.

We then compute a vector $\mathbf{h}_{\text{MIP}} \in \mathbb{R}^{h \times 1}$ that captures the semantic relationship between contextual and prototypical senses:

$$\mathbf{h}_{\text{MIP}} = f([\mathbf{v}_{S',t}; \mathbf{v}_{P,t'}]) \tag{5}$$

where $f(\cdot)$ is a linear transformation that learns the semantic difference between the contextual usage $\mathbf{v}_{S',t}$ and the prototypical sense $\mathbf{v}_{P,t'}$.

The model is trained using cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{6}$$

This model achieved an F1-score of 75.1 on BCCWJ-Metaphor through 5-fold cross-validation, providing the foundation for large-scale automated corpus construction.

5 Evaluation

To assess the quality of our automatically generated metaphor corpus, we conducted a manual validation of the novel annotations produced by our model to demonstrate its ability to augment and enrich existing resources.

5.1 Validation of Novel Metaphor Annotations

While the BCCWJ-Metaphor provides a crucial benchmark, manual annotation processes, despite their high quality, may have occasional oversights due to the complexity and time-intensive nature of comprehensive metaphor identification. A key part of our contribution, therefore, is to demonstrate our model's ability to identify and fill these gaps.

We applied our trained model to the entire BCCWJ to generate comprehensive metaphor predictions. To validate the quality of our automated annotations, we focused on the BCCWJ-Metaphor portion and identified instances that were originally not labeled as metaphorical but were newly predicted as metaphors by our model. From these newly predicted instances, we randomly sampled 500 cases for manual validation by an expert linguist specializing in Japanese metaphor research.

The results of this validation demonstrate the effectiveness of our approach for corpus augmentation. The analysis revealed that 60.3% of the newly identified instances were judged to be genuine metaphorical expressions that had been missed in the original manual annotation. Among these 302 validated genuine metaphors, the distribution across metaphor types was as follows: word-level metaphors accounted for 59.2% (179 instances), metonymy for 11.3% (34 instances), synecdoche for 15.3% (46 instances), and discourse-level metaphors for 0.6% (2 instances). The remaining 13.6% (41 instances) were classified into other categories or required further analysis.

It proves that our automated method is not merely replicating human work, but is acting as a powerful tool to enhance it by discovering valid omissions. This validates the quality of our new, larger corpus as a more complete resource for metaphor research.

6 Error Analysis

Our evaluation quantitatively demonstrated the model's overall effectiveness. A deeper analysis of its performance, however, reveals important patterns in its errors. The following discussion explores the linguistic reasons behind these challenges, drawing on specific examples from our analysis.

6.1 Word-level Metaphor Detection Challenges

Among the compound metaphors annotated in BCCWJ-Metaphor, 21.23% were found to be incorrect. A notable tendency was that the system failed to detect compound metaphors when the dictionary definition of a word included senses explicitly labeled as "figurative" or "by extension." Such cases are so highly conventionalized that human annotators can readily recognize them as metaphors; however, the system often misclassified them. For example, the system judged the expression "誕生" ("birth") in "無党派組織『２１世紀の千葉を創る県民の会』が知事誕生の原動力となった" ("The nonpartisan organization 'Prefectural Citizens' Association to Create 21st Century Chiba' became the driving force behind the governor's birth") as non-metaphorical. In the construction "知事の誕生" ("the governor's birth"), the literal meaning of "誕生" refers to human birth. However, in this context, it is used in the figurative sense of "the emergence or establishment of something" (Shogakukan Inc., 2000–2002). Consequently, the annotator labeled it as a metaphor, but the system failed to do so.

6.2 Metonymy Detection Challenges

Among the metonymies annotated in BCCWJ-Metaphor, 23.30% were incorrect. A frequent error occurred when the system targeted content words within company names. Metonymic expressions tend to be identified in constructions involving selectional restriction violations, such as "a company sells". Since company names are often used in metonymic contexts, this is a reasonable target; however, in this study, parts of proper nouns were sometimes misidentified.

Errors also occurred when interpretation required sentence-level understanding rather than solely relying on lexical meaning or construction patterns. For example, in [赤松円心が願っていたのも、]中央の犠牲となることなく、穏やかに安定してその地方を治めていくことにあったはずで" ("[It must have been Akamatsu Enshin's wish as well] to govern the local region peacefully and stably without becoming a sacrifice to the center [central government,]") the construction "中央の犠牲" ("a sacrifice to the center") can be judged metaphorical. Nevertheless, the system labeled it as non-metaphorical. In this case, recognizing the metonymy requires understanding that "中央" ("the center") corresponds to "地方" ("the local region") in the sentence, a relationship that cannot be easily detected from the lexical meaning or construction alone.

6.3 Synecdoche Detection Challenges

Synecdoche exhibited the highest error rate among metaphor categories in BCCWJ-Metaphor, with 39.50% incorrect. This category often requires interpretation based on context or background knowledge, making it difficult to identify based solely on lexical meaning or deviations from conventional constructions. One contributing factor is the prevalence of euphemistic and exemplary expressions.

For example, in “大和の国の中には、水にめぐまれへん村がぎょうさんあった” (“In Yamato Province, there were many villages not blessed with water”), the system failed to identify “めぐむ” (“bless”) as synecdoche. The basic sense of “めぐむ” is “to feel affection for,” with extended meanings such as “to show compassion” and “to give alms” (Shogakukan Inc., 2000–2002). In the idiomatic expression “水にめぐまれる” (“to be blessed with water”), the phrase euphemistically refers to the availability of water, which can be interpreted as a type–category relationship, hence synecdoche.[3] However, the system failed to detect this. Such euphemistic expressions are highly conventionalized, making them difficult to classify as selectional restriction violations at the word or construction level, even though they are easily recognized as figurative by humans. Nevertheless, in the present method, certain euphemistic expressions were successfully detected—for example, “世を去る” (“to leave the world,” a euphemism for “to die”) in “[誰かが]この世を去った時に” (“when [someone] passes away”). Like “水にめぐまれる”, this is idiomatic, but its appearance in a hypothetical context may have influenced the system’s judgment.

Expressions closer to prototypical exemplification also proved challenging when they required contextual or background knowledge. For instance, in “下戸の彼氏もコーラで付き合う” (“Even my boyfriend, who cannot drink alcohol, joins in with cola”), “コーラ” (“cola”) refers to a category of soft drinks. In this context, given “下戸” (“unable to drink alcohol”) and a drinking-party scenario, “cola” functions as a prototypical example of a non-alcoholic beverage used to participate in the event. However, the system judged it as non-metaphorical.

Interestingly, there were also cases where the system successfully detected expressions that humans might not easily recognize as figurative. For example, in “京都に店を出す” (“open a shop in Kyoto”), “出す” (“to put out”) is interpreted in an extended sense beyond its basic meaning, constituting synecdoche. Although humans may perceive this as a natural semantic extension, the system correctly identified it.

[3] In Japan, with its animistic cultural background, there isn’t a single fixed deity that performs blessings. As a result, the idea of “blessed with water” is understood as a metaphorical expression.

6.4 Discourse-level Metaphor Detection Challenges

Among the metaphor categories in BCCWJ-Metaphor, contextual metaphors had the second-highest error rate at 30.20%. The present approach is generally ill-suited to detecting such cases. When an expression is not metaphorical at the sentence level, metaphor recognition cannot be achieved through lexical meaning alone. However, the system did correctly identify certain examples. For instance:

- “[船頭、つまり経営者が]仕かけを工夫して、[釣り人、つまり社員に]釣り方を教える” (“[The boatman, that is, the manager] devises traps and teaches [the fisherman, that is, the employee,] how to fish”) occurs in a business-strategy context, interpreted as a figurative expression that, at the sentence level, deviates from the context.
- “種をまいても収穫を急がんことやで” (“Even if you sow seeds, don’t rush the harvest”) is used in the context of business know-how, also interpreted as a figurative expression.
- “目が画面に釘づけになった” (“My eyes were nailed to the screen”) appears in a surveillance-camera playback context, making it a well-established metaphor.
- “「ヤマタノオロチ」が暴れている時に、政府は「草薙の剣」を使うことができなかったようなものです” (“It is as if, when the ‘Yamata no Orochi,’ a famous giant eight-headed serpent in Japanese mythology, was rampaging, the government could not use the ‘Kusanagi sword,’ a legendary Japanese sword”) occurs in the context of the government’s response to a financial crisis. The quotation marks and the inherently metaphorical construction “government uses the ‘Kusanagi sword”’ likely contributed to the system’s detection.

These examples suggest that when unnatural constructions or typographical indicators (e.g., quotation marks) are present, metaphor detection may be aided by features beyond long-range discourse context.

In summary, the system's correct and incorrect judgments did not necessarily align with the ease of human annotation. While there are cases where the system failed on expressions that humans find easy to classify, it could also capture instances that humans may overlook. This suggests that such a system has strong potential as a supplementary tool for human metaphor annotation.

7 Conclusion

This paper presents the first large-scale automated construction of a Japanese metaphor corpus, extending metaphor annotation across the entire BCCWJ through systematic application of WLSP-based detection models. Our approach successfully creates the largest Japanese metaphor resource to date, encompassing diverse text types beyond the scope of manual annotation.

The validation study provides compelling evidence for the effectiveness of automated corpus construction. With 60.3% of newly predicted instances representing genuine metaphorical expressions, our method demonstrates the ability to identify valid metaphors missed in manual annotation, enhancing rather than merely replicating existing resources.

Our analysis reveals systematic patterns in detection capabilities across metaphor types, with particular challenges in synecdoche and discourse-level metaphors due to conventionalization and contextual dependencies. These findings provide valuable insights for future model development and highlight the complexity of automated metaphor detection in Japanese.

The resulting corpus enables large-scale studies of metaphor usage patterns across diverse genres and provides essential training data for advancing computational metaphor research.

Future work should focus on improving the detection of highly conventionalized expressions and incorporating broader contextual information to better handle discourse-level metaphors.

Acknowledgements

This work was supported by JSPS KAKENHI Grants JP22K12145 and JP25K00459, the JSPS Postdoctoral Fellowship for Research in Japan (for Foreign Researchers), the NINJAL Collaborative Research Project "Empirical Computational Psycholinguistics Using Annotated Data", and the Kayamori Foundation of Informational Science Advancement Research Grant "Extraction of Conceptual Metaphors Using Natural Language Processing." We are also grateful to Professor Makoto Yamazaki and Professor Wakako Kashino for providing us with the basic sense data of the WLSP.

References

Soma Asada, Kanako Komiya, and Masayuki Asahara. 2024. Comprehensive assignment of word list by semantic principles concept ids to the balanced corpus of contemporary written japanese. In *Proceedings of the 30th Annual Conference of the Association for Natural Language Processing*, pages 2767–2772.

Yulia Badryzlova, Natalia Shekhtman, Yekaterina Isaeva, and Ruslan Kerimov. 2013. Annotating a Russian corpus of conceptual metaphor: a bottom-up approach. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 77–86, Atlanta, Georgia. Association for Computational Linguistics.

National Institute for Japanese Language and Linguistics, editors. 2004. *Bunrui Goi Hyo: Expanded and Revised Edition*. Dai Nihon Tosho. In Japanese.

Elżbieta Hajnicz. 2022. Annotation of metaphorical expressions in the basic corpus of Polish metaphors. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5648–5653, Marseille, France. European Language Resources Association.

JBB Herrmann, Karola Woll, and Aletta Dorst. 2019. Linguistic metaphor identification in german. *Metaphor identification in multiple languages. MIPVU around the world*, pages 113–135.

Marlene Johansson Falck and Lacey Okonski. 2022. Procedure for identifying metaphorical scenes (pims): A cognitive linguistics approach to bridge theory and practice. *Cognitive Semantics*, 8(2):294–322.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. 2018. Annotation of 'word list by semantic principles' labels for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Sachi Kato, Rei Kikuchi, and Masayuki Asahara. 2022. Assignment of metaphorical expression information based on mip to the balanced corpus of contemporary written japanese. In *Proceedings of the 28th Annual Conference of the Association for Natural Language Processing, In Japanese*, pages 1427–1431.

Sachi Kato, Rei Kikuchi, and Masayuki Asahara. 2025. Metaphor recognition and information assignment procedures in bccwj-metaphor. In *Proceedings of the 31st Annual Conference of the Association for Natural Language Processing, In Japanese*.

Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, and Hiroyuki Shinnou. 2018. Comparison of methods to annotate named entity corpora. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(4):34.

Tina Krennmayr and Gerard Steen. 2017. Vu amsterdam metaphor corpus. In *Handbook of linguistic annotation*, pages 1053–1071. Springer.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Ogiso, Masaya Maruyama, Hideki Ogura, Wakako Kashino, Hiroshi Koiso, Makoto Yamaguchi, Masahiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48(2):345–371.

Akira Nakamura. 1977. *Theory and Classification of Metaphorical Expressions*. Shuei Shuppan, Tokyo.

Tryntje Pasma. 2019. Chapter 5. linguistic metaphor identification in dutch. In *Metaphor Identification in Multiple Languages: MIPVU around the world*, pages 91–112. John Benjamins Publishing Company.

Group Pragglejaz. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

G Reijnierse. 2010. Making mip operational for french: practical and theoretical issues concerning the choice of a dictionary. In *8th International Conference on Researching and Applying Metaphor (RaAM 10)*.

W. G. Reijnierse, C. Burgers, T. Krennmayr, and G. J. Steen. 2018. Dmip: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, and Sergio-Luis Ojeda-Trueba. 2024. Evaluating the development of linguistic metaphor annotation in Mexican Spanish popular science tweets. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 59–64, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *International Conference on Language Resources and Evaluation*.

Shogakukan Inc. 2000–2002. , 2nd edition. Shogakukan Inc., Tokyo. Digital version accessed via JapanKnowledge, released July 2, 2007.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tom Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing.

Hongying Zan, Junyi Chen, Xiaoyu Cheng, and Lingling Mu. 2018. Construction of word sense tagging corpus. In *Chinese Lexical Semantics - 19th Workshop, CLSW 2018, Chiayi, Taiwan, May 26-28, 2018, Revised Selected Papers*, volume 11173 of *Lecture Notes in Computer Science*, pages 679–690. Springer.

A Part-of-Speech Categories for Target Word Selection

Our content word filtering process targets word classes that are typically subject to metaphorical usage according to MIP and Japanese linguistic characteristics. The selection is based on UniDic part-of-speech categories used in BCCWJ morphological analysis.

Table 1: Part-of-Speech Categories for Metaphor Detection

Japanese	English
Selected Categories (Content Words)	
名詞-数詞	Noun-Numeral
名詞-助動詞語幹	Noun-Auxiliary Verb Stem
名詞-普通名詞-一般	Noun-Common-General
名詞-普通名詞-サ変可能	Noun-Common-Verbal
名詞-普通名詞-形状詞可能	Noun-Common-Adjectival
名詞-普通名詞-副詞可能	Noun-Common-Adverbial
名詞-普通名詞-助数詞可能	Noun-Common-Counter
名詞-固有名詞-一般	Noun-Proper-General
名詞-固有名詞-地名-一般	Noun-Proper-Place-General
名詞-固有名詞-地名-国	Noun-Proper-Place-Country
名詞-固有名詞-人名-一般	Noun-Proper-Person-General
動詞-一般	Verb-General
動詞-非自立可能	Verb-Bound
形容詞-一般	Adjective-General
形容詞-非自立可能	Adjective-Bound
副詞	Adverb
形状詞-一般	Adjectival Noun-General
連体詞	Attributive
代名詞	Pronoun
Excluded Categories (Function Words)	
助詞-格助詞	Particle-Case
助詞-係助詞	Particle-Binding
助詞-接続助詞	Particle-Conjunctive
助詞-副助詞	Particle-Adverbial
助詞-終助詞	Particle-Final
助動詞	Auxiliary Verb
補助記号-読点	Symbol-Comma
補助記号-句点	Symbol-Period
補助記号-括弧閉	Symbol-Bracket Close
補助記号-括弧開	Symbol-Bracket Open
接頭辞	Prefix
接続詞	Conjunction
接尾辞-名詞的-一般	Suffix-Nominal-General
接尾辞-名詞的-助数詞	Suffix-Nominal-Counter
接尾辞-形状詞的	Suffix-Adjectival Noun
記号-一般	Symbol-Character
感動詞-一般	Interjection-General
空白	Whitespace

Converging and diverging variations in metaphorization of light verbs: A corpus study on four Chinese speech communities

Ka-Fai Yip[1], Yaxuan Ji[2], Dicky Ko[3], Thomas Ho[3], and Benjamin K. Tsou[3,4]
[1]Yale University, New Haven, Connecticut, United States
[2]The Education University of Hong Kong, Hong Kong SAR
[3]Chilin (HK) Ltd., Hong Kong SAR
[4]City University of Hong Kong, Hong Kong SAR
kafai.yip@yale.edu, reneejyx@gmail.com
{dicky.ko, thomas.ho}@chilin.hk, btsou99@gmail.com

Abstract

Metaphorization is a robust semantic process that underpins the development of grammatical categories and associated language variation and change. This study adopts a process-based approach to investigate variations in metaphorical extension of five common light verbs from different lexical sources in four Chinese speech communities: *jiyu*給 予 "GIVE", *yu*予 "GIVE", *jia*加 "ADD", *gan*幹 "WORK HARD", and *nong*弄 "FIDDLE". Drawing on longitudinal (1995-2016) and latitudinal (Hong Kong, Taiwan, Beijing, Shanghai) data from the Pan-Chinese corpus LIVAC (https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus) we demonstrate that light verb develop differentially, ranging from uniform progression, significant variations, to even backward development. We also propose an innovative measure to quantify the distances in metaphorization between the regions and uncover that Chinese speech communities tend to converge on the process over time as led by Mainland Mandarin, despite clustering of Hong Kong and Taiwan and of Beijing and Shanghai, as also confirmed statistically. Based on the findings, we explore how the differential developments result from the interaction of underlying mechanisms of semantic change and lexical competition, as well as contact among Chinese varieties. Moreover, we highlight the importance of a corpus that rigorously monitors multiple regions in revealing covert variations in Global Chinese varieties.

1 Introduction

Language develops through changes such as enrichment of its lexical and grammatical repertoire. Thus there could be emergence of categorical shifts or new categories with semantic change (Heine et al., 1991; Hopper and Traugott, 2003, i.a.). Metaphors have been recognized to be instrumental in meaning extension from a concrete domain to an abstract domain (Traugott, 1982; Bybee and Pagliuca, 1985; Claudi and Heine, 1986; Sweetser, 1988). A salient case associated with robust metaphorization is **light verbs**, e.g., *give a person a slip*, where *give* loses the sense of ownership transfer and becomes semantically bleached, in contrast to the lexical "heavy" verb usage *give a person a book* (Jespersen, 1954, p.117ff; see also Cattell, 1984; Butt, 2010; Mohanan, 2017). Such **metaphorization** process is ubiquitous and can be found in the Germanic and Romance families (e.g., Alvarez-Morera, 2023; Pompei, 2023), Indo-Aryan languages (e.g., Hook, 1991; Butt and Geuder, 2003), Austronesian and Austroasiatic languages (e.g., Kwon, 2004; Nugraha, 2025), Sino-Tibetan languages including Tibetan (Lai, 2024) and Chinese (e.g., Diao, 2004; Tsou and Yip, 2020; Lu and Huang, 2023), among many others.

On the other hand, synchronic grammatical variations can reflect varying trajectories of diachronic grammaticalization processes (Weinreich et al., 1968; Brinton and Traugott, 2005). Nevertheless, little is known about the contribution of metaphorization to **language variations**. The issue is especially pressing in a language like Chinese, which lacks an inflectional morphology to signal grammaticalization and thus whose change and variation tend to be covert (Tsou and Ji, 2022; Yip and Tsou, 2022; Tsou et al., 2023), in particular under the Global Chinese context (Tsou and You, 2003; Lin et al., 2019; W. C. J. Yip and Tang, 2022).

Light verbs have drawn the attention of Chinese linguists since the introduction of Modern linguistics (Lu, 1999/1980, p.294; Wang, 1985, p.142). Zhu, in his seminal work in 1985, referred to them as "bleached verbs" 虛化動詞, and identified 6 with notably different lexical sources: *jinxing*進行 (from ENTER and WALK), *zuo*作 (MAKE), *jiayi*加以 (ADD), *geiyi*給以, *jiyu*給予 and *yuyi*予以 (all from GIVE). Several works followed up

on their grammatical properties (e.g., Zhou, 1987; Li and Chai, 1995; Yan, 1998; and many others). In 2004, corpus-based quantitative studies in different Pan-Chinese regions began to emerge. Diao, based on a self-curated corpus in Mainland China, extended the scope to *congshi*從事 (WORK), as well as to monosyllabic *zuo*做 (MAKE), *guo*搞 (MAKE), *gan*幹 (WORK HARD), *nong*弄 (FIDDLE/PLAY), *jia*加 (ADD), and *yu*予 (GIVE). In the same year, Wang (2004) also drew corpus data from Taiwan to compare *zuo*做, *guo*搞, and *nong*弄.

Areal differences were less noticed until Diao's comparison in 2012 on four Pan-Chinese regions (Mainland China, Hong Kong, Taiwan, and Macau), concerning *jinxing*進行 and *guo*搞; and Huang et al's work in 2014 (later developed into Jiang et al., 2016, 2021; and Xu et al., 2022) on the variations in Mainland and Taiwan Mandarin, concerning *jinxing*進行, *congshi*從事, *zuo*做, *guo*搞, and *jiayi*加以. Focusing on syntactic properties (e.g., aspectual markers and object types, etc.), they reached a common conclusion that light verbs in Mainland Mandarin are more grammaticalized than in Taiwan Mandarin. Diao additionally suggested a convergence of the four regions by observing borrowing of fixed expressions. Most recently, Tsou and Yip (2020) initiated comparison of metaphorization on a common but understudied light verb *da*打 (from HIT) in Beijing, Hong Kong, and Taiwan with LIVAC, and similarly concluded that Beijing had a higher degree of metaphorical shift than Hong Kong and Taiwan. They also traced the longitudinal differences in metaphorization of *da*-verbs and showed how the semantic process varies across the regions (see also Yip and Tsou, 2022 for extension to Macau). Lu and Huang (2023) and Kuo (2025) provided further insights that *yuyi*予以 and *jiayi*加以 are fully grammaticalized but *jiyu*給予 is still half-way through the process, largely based on Taiwan data.[1] While there have been fruitful results on the latitudinal variations in grammaticalization and on the development trajectories of light verbs in Chinese,[2] less attention has been given to study metaphorization and its longitudinal variations (except Tsou and Yip, 2020), as well as wider implication. This calls for a large-scale quantitative study to explore whether different varieties of Chinese converge or diverge and the possible causation.

The current study proposes a novel **process-based** approach to investigate the metaphorization variations in Chinese. The empirical basis consists of five light verbs: *jiyu*給予, *yu*予, *jia*加, *gan*幹, and *nong*弄. Unlike other light verbs that are fully metaphorized and no longer have the literal sense (e.g., *jinxing*進行 "do" and *jiayi*加以 "do/impose"), these five verbs are still undergoing metaphorization, as indicated by the polysemy between a literal meaning (lexical verb) and a metaphorical meaning (light verb), such as *jiyu shuben*給予書本 "give books" and *jiyu zhichi*給予支持 "give support". They offer us a window to probe into the dynamic *process* of semantic bleaching resulting from metaphorization, over time and across regions.

Adopting such a dynamic perspective, we draw data in LIVAC from 1995-2016 among four Chinese speech communities: Hong Kong (HK), Taiwan (TW), Beijing (BJ), and Shanghai (SH) to compare the differential developments in their metaphorization processes. The key findings are:

(i) Light verbs may develop differentially: while some have uniform progressions, some are still in flux and even develop "backwards";

(ii) There is a clear clustering between regions: the HK-TW cluster and the BJ-SH cluster;

(iii) Despite the clustering, Mainland Mandarin has been leading the metaphorization process: (a) the light verbs are overall more metaphorized in BJ and SH; (b) other regions, especially TW, converge with BJ and SH in recent years.

The rest of the paper is organized as follows. Section 2 introduces the corpus base LIVAC and the methodology for quantitative analyses. Section 3 overviews the polysemy and metaphorization of the five light verbs. Section 4 investigates the converging and diverging variations among the four speech communities. Section 5 concludes with remarks on the implications for light verb development and the importance of a rigorously curated corpus in large-scale measurement of variations in Global Chineses, as well as prospects.

2 Corpus base and methodology

Our data are drawn from the Pan-Chinese synchronous database LIVAC (Tsou and Kwong, 2015;

[1] They classified these three light verbs as the GIVE-group despite the fact that jiayi加以 comes from ADD.

[2] See also Cai et al. (2019) for comparison on "semantic prosody" between *jinxing*進行 and *shaodao*受到in different registers; Jiang (2020) for the event semantics of light verbs.

	HK	TW	BJ	SH
Jiyu 給	6570	**5202**	**13099**	**6766**
(%)	(22)	(37)	**(50)**	**(39)**
Yu 予	9667	1567	714	435
(%)	(32)	(11)	(3)	(3)
Jia 加	**10876**	**5132**	4136	3872
(%)	**(36)**	**(36)**	(16)	(22)
Gan 幹	947	1015	6894	3168
(%)	(3)	(7)	(26)	(18)
Nong 弄	2158	1206	1318	3097
(%)	(7)	(9)	(5)	(18)
Total	30218	14122	26161	17338
(%)	(100)	(100)	(100)	(100)

Table 1: Distribution of the five light verbs in 1995–2016 across regions

http://www.livac.org/),[3] featuring an innovative "Windows" approach that supports the process-based comparative perspective adopted here. Since 1995, LIVAC has regularly and rigorously cultivated over 750 million characters of representative media texts from Pan-Chinese communities including Beijing, Guangzhou, Hong Kong, Macau, Shanghai, Shenzhen, Singapore, and Taiwan, on a weekly basis and involving different subject domains. The curated data over three decades in LIVAC provide a unique basis for both latitudinal and longitudinal comparisons.

In total 87839 tokens of the five light verbs, *jiyu* 給予, *yu*予, *jia*加, *gan*幹, and *nong*弄, were found in LIVAC from 1995 to 2016 in HK, TW, BJ, and SH. A breakdown in Table 1 shows that the distribution is not even across regions. Boldface indicates the most common light verb(s) within a region (two for TW as the difference is less than 0.5%), and underlining indicates the top region in which a light verb is most frequently used as compared to other regions. Some clustering already emerges: while HK and TW use *jia*加 the most, BJ and SH (as well as TW) use *jiyu*給予 the most.

All raw sentences were assigned a random number (using Excel's `=RAND()`) and reordered. For each light verb, a maximum of 50 sentences were sampled from each year of two time periods: (a) 1995-2000, and (b) 2011-2016, from each region, giving **9781** sentences in total (i.e., over one tenth of total tokens). Both sampling and ordering are hence randomized, avoiding oversampling similar uses in consecutive sentences. Two annotators manually classified each token as literal (Type I), intermediate (Type II), or metaphorical (Type III) (criteria to be exemplified in Section 3), with disagreement adjudicated by a third annotator (i.e., each annotation is agreed upon by at least two annotators). To boost reliability, each rater attended a training session in which they were required to annotate a trial of 50 sentences with a follow-up work group meeting to resolve disagreement, before the actual annotation began. Krippendorff's alpha was measured for inter-coder reliability using the `irr` package in R (with `ordinal` metric). The pre-adjudication agreement level is satisfactory ($\alpha = 0.82$), which is improved further after adjudication ($\alpha = 0.93$, counting the adjudicator and the closest rater). A sample annotation guideline is provided in Appendix A.

We follow Tsou and Yip (2020) and quantify the metaphorization degree of each light verb w within a given region r with the **Metaphorization Index (MI)** (0=fully literal; 100=fully metaphorized):

$$MI_r^w = \frac{\left(\frac{t_{II}^w}{2}\right) + t_{III}^w}{t_I^w + t_{II}^w + t_{III}^w}$$

($t_{I/II/III}^w$ = raw tokens of Type I, II, or III uses of w)

Variations in metaphorization can be measured by comparing the MI of a given light verb across regions, and changes can be measured by comparing the MI in different time spans. Note that the distribution of light verbs is uneven across regions, which might induce potential bias. MI avoids this by calculating the ratio without counting in frequency, allowing direct comparison between regions.

Nevertheless, discarding frequency entirely might create confounds for the overall pattern. Although we may aggregate the MI of all light verbs, two caveats follow: (i) while w_1 may have a higher MI in r_x than r_y, w_2 may well have a reversed distribution, a distinction that would disappear after direct aggregation; (ii) even if w_1 has the same MI in both regions, it might be far more frequent in one region over the other, which is not reflected by the MI itself.

To measure the magnitude of variations without masking data as such, we propose the **Metaphorization Distance Index (MDI)** below. It calculates the distance in metaphorization across regions. For ease of interpretability and avoiding

[3] "Synchronous" means the corpus actively cultivates data every year and hence monitors development in real-time.

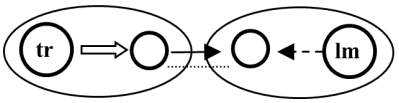


Figure 1: The image-schematic structure of TRANSFER (Langacker 2008, p.242)

confusion with MI, the MDI value is divided by 20 (hence, 0=closest; 10=furthest).

$$MDI_{r_y}^{r_x} = \frac{\Sigma_{i=1}^{n} |(t_{r_x}^{w_i} \times MI_{r_x}^{w_i}) - (t_{r_y}^{w_i} \times MI_{r_y}^{w_i})|}{20}$$

(r_x and r_y = two given regions; n-total number (type) of words; t_r^w and MI_r^w = normalized tokens and MI of a given word w in r respectively.)

The MDI sums the *differences* in the number of metaphorized tokens between two regions. The higher the MDI of r_x relative to r_y, the larger the difference between r_x and r_y. It is equipped with both word frequency and metaphorization degree to reflect synchronic variations in the above scenarios (i) and (ii). Longitudinal variations can be further measured by comparing the MDI values in different time periods.

Furthermore, we conducted **Ordinal Linear Regression (OLR)** by ORDINAL package in R (4.4.2), to model the degree of metaphorization (`Type I/II/III`) for each light verb. The model included `region` and `time` as fixed effects, along with their interaction, to examine whether regional variations and diachronic changes influenced the degree of metaphorization. Post hoc pairwise comparisons were performed to determine whether the latitudinal and longitudinal distributions of metaphorization types were statistically significant, with p-values adjusted using the Tukey method.

For a holistic analysis of the metaphorization process of all five light verbs in each region, we built a full model with three fixed effects from region, time, and individual verbs (`Type~region+time+verb`) to examine the variation due to these three factors and their contribution in the metaphorization process.

3 Metaphorization of light verbs

Metaphorization, according to Heine et al. (1991, p.46), "involves a transfer, or a mapping of an image schema [in the sense of Sweetser (1988)], from one domain of conceptualization onto another". It is a unidirectional process and proceeds from a concrete domain to an abstract one (Claudi and Heine, 1986), which results in generalization of meaning (Bybee and Pagliuca, 1985). During the course, semantic complexity may be lost with subsequent de-categorization, leading to development of a new category (Claudi and Heine, 1986). In other words, metaphorization is a process that leads to further semantic bleaching and grammaticalization.

The five light verbs under discussion are polysemous and have literal (**Type I**), intermediate (**Type II**) and metaphorical uses (**Type III**). The sample sentences extracted from LIVAC are given in the Appendix B. We propose that such polysemy is a result of metaphorization, and different senses share an image-schematic structure. One common schema is TRANSFER (Langacker, 1991, 2008; Halverson, 1999), which underpins the polysemy of GIVE verbs like ***jiyu***給予 and ***yu***予. In Figure 1, two participants are involved: the giver (trajector) and the recipient (landmark), with their own dominion (=ellipses). The trajector causes the object to move, such that the recipient gains subsequent access to it.

The literal uses of GIVE (I) involve a transfer of ownership of the object and corresponds to the TRANSFER schema, e.g., jiyu shuben給予書本 "give books". The object does not need to be tangible, e.g., jiang tudi chanquan yizhuan yu taren將土地産權移轉予他人 "transfer the land property right to the others". The intermediate uses (II) involve a causative meaning instead, e.g., yu ren yixiang shenke予人印象深刻 "impress people", where the impression is not "transferred" but rather caused by an individual to be perceived by another one, who is hardly counted as a recipient. Here, with conceptual metaphors (Lakoff and Johnson, 1980) like EMOTIONS ARE PHYSICAL OBJECTS, the causative uses can still be understood using the TRANSFER schema (Cervel, 2004). The same lexical items can thus extend their meaning from "transfer" to "cause". The metaphorical uses (III) like jiyu paichu給予排除 "to eliminate" and zai-yu qianghua再予强化 "to strengthen (it) again" only involve transitive actions "elimination" and "strengthening", and both GIVE verbs do not seem to contribute any concrete meaning. We regard these uses as genuine light verb usage, which, unlike Type I/II, can be replaced with other fully metaphorized light verbs yuyi予以/jiayi加以, and sometimes with jinxing進行, but generally not the lexical verb gei給 "give". Again, with the metaphor ACTIONS ARE PHYSICAL OBJECTS, this use conforms to the TRANSFER schema where an ac-

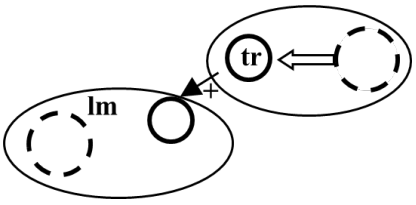


Figure 2: ADD's image schema

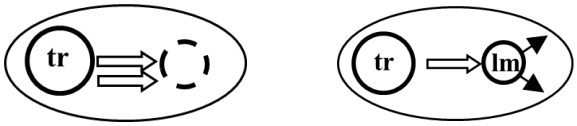


Figure 3 (left): WORD HARD's image schema
Figure 4 (right): FIDDLE's image schema

tion is imposed on the landmark by the trajector.

Jia 加 means addition of an object in its literal uses (I) like *jia shui* 加水 "add water" and *jia xi* 加戲 "add a scene". We propose an image schema of ADD in Figure 2: the object (trajector) is added to another object's dominion (landmark), forming a part-whole relation. The causer of the addition is backgrounded (dotted lines). Addition of an action/event is classified as intermediate uses (II), e.g., *jia kao* 加考 "have an additional exam" where *jia* serves as a manner modifier. Likewise, this use is enabled by the metaphor ACTIONS ARE PHYSICAL OBJECTS. The metaphorical light verb uses (III) do not contribute lexical meaning and can be replaced by the disyllabic *jiayi* 加以, e.g., *da-jia zanshang* 大加讚賞 "to praise greatly". This use falls under the image schema, where the causer (often foregrounded) "adds" a praise (via similar EMOTION/ACTION metaphors) to the target object that is understood as a person. The meaning of addition is also bleached, just like how the meaning of transfer is bleached for GIVE.

Gan 幹's literal sense (I) is WORK HARD, e.g., *gan zhonghuo* 幹重活 "do hard labor". Its image schema is in Figure 3, where the trajector imposes extra force on an underspecified, backgrounded object. The intermediate uses (II) are generalized to "work/engage in a job" with a foregrounded object and backgrounded effort, e.g., *gan baoan* 幹保安 "be a security guard" and *gan liang ge ban* 幹兩個班 "work two shifts". They can often be replaced with *congshi* 從事/*zuo* 做. We also regard the sense of "conflict" as intermediate uses, where the hard effort remains foregrounded, e.g., *dui-zhe gan* 對著幹 "to work against (someone)".

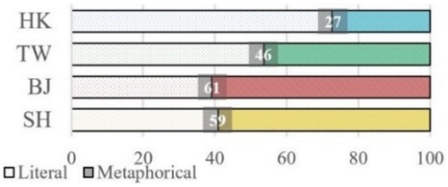


Figure 5: Aggregated MI of the four regions (95∼16)

The metaphorical uses (III) are fully generalized to "do" and can only be replaced with *zuo* 做, e.g., *gan shashi* 幹傻事 "do silly things" and *gan-xia duo qi xiji shijian* 幹下多起襲擊事件 "do/commit multiple attacks". Here, the object in the schema is a foregrounded vehicle for the event/action. Note that only one participant is involved, unlike GIVE and ADD.

Nong 弄 literally means FIDDLE (often with hands) (I), e.g., *nong toufa* 弄頭髮 "fiddle hair". Its image schema is in Figure 4, where the trajector causes an object (landmark) to move around but still within their own dominion. The intermediate uses (II) include "make" and "obtain", e.g., *nong zaocan* 弄早餐 "make breakfast" and *nong qian* 弄錢 "earn/obtain money", which we characterize as transformation of the landmark, either from an object A to B or from non-existence to existence. The metaphorical uses (III) include "do" (can be replaced with *zuo* 做) and "cause" (can be replaced with *gao* 搞), e.g., *nong shiqing* 弄事情 "do something" and *nong qingchu* 弄清楚 "make it clear". Again, these uses are covered by the image schema and EMOTIONS/ACTIONS ARE PHYSICAL OBJECTS, with the loss of semantic complexity that the objects move around repeatedly.

Next, we demonstrate how metaphorization of these five light verbs develops differentially.

4 Variations in Chinese

The overarching observation is that BJ and SH are highly advanced in metaphorization, followed by TW, whereas HK is the least sensitive. Figure 5 shows the aggregated MI in each region, i.e., the ratio of metaphorical tokens of all five light verbs. The results of the OLR model confirm the regional variations. BJ serves as the reference level for regional comparisons. Among the three regions, SH does not differ significantly from BJ ($B = -0.11$, $p = 0.085$). In contrast, HK and TW are significantly different from BJ. HK shows the greatest divergence ($B = -0.98$, $p < 0.001$), followed by TW, with a smaller but statistically meaning-

	HK	TW	BJ	SH
HK	0.0	**1.0**	2.5	2.2
TW	1.0	0.0	2.3	1.9
BJ	2.5	2.3	0.0	**1.3**
SH	2.2	1.9	1.3	0.0

Table 2: The MDI matrix of the four regions (95∼16) (boldface=lowest value for each region)

	HK	TW	BJ	SH
HK	0.0	**-0.6**	0.1	0.0
TW	-0.6	0.0	-0.1	**-1.5**
BJ	0.1	-0.1	0.0	**-1.1**
SH	0.0	-1.5	-1.1	0.0

Table 3: Change in MDI in the four regions (95∼00 vs. 11∼16) (boldface=lowest value for each region)

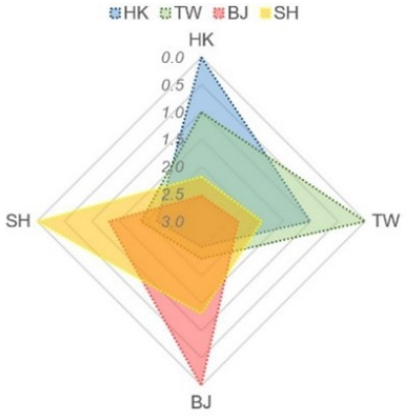
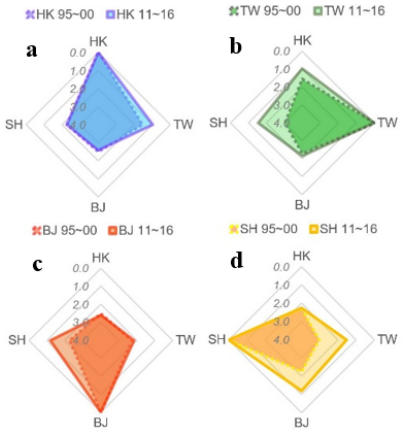


Figure 6: The radial graph of MDI of the four regions (95∼16)



Figures 7a-d: Radial graphs of change in MDI of HK, TW, BJ, and SH (95∼00 vs. 11∼16)

ful divergence, and a lower degree ($B = -0.15$, $p = 0.016$). As we will reveal below, however, the development of the light verbs is not uniform.

4.1 Convergence and divergence in the metaphorization process

First, clear **clustering** is obtained when measuring the distances between regions. The MDI of each pair of regions is given in Table 2, HK and TW are the closest to each other as indicated by the lowest MDI, so as BJ and SH. In contrast, HK-BJ is the most dissimilar pair, followed by BJ-TW and HK-SH. In other words, HK and TW converge together but diverge from BJ and SH in metaphorization. This is represented by the radial graph in Figure 6, where overlapping area indicates convergence and non-overlapping area indicates divergence.

Longitudinal comparison is achieved by subtracting the MDI value in 1995-2000 from that in 2011-2016, as shown in Table 3. Negative values signal a decrease in the distance, i.e., the two given regions have become more similar. Most values are negative, suggesting an overall convergence. The change in MDI is represented in Figures 7a-d for each region. Expansion of the light shaded area (from the dark one) indicates decreased distance (i.e., increased similarity) with other regions. HK tends to converge with TW whereas BJ with SH. On the other hand, TW also develops towards SH in addition to HK, and SH becomes more like both BJ and TW. Interesting, BJ does not converge towards TW, suggesting that SH's convergence with TW is independent of BJ's influence, which invites further explanation. Overall, we conclude that over time Chinese speech communities generally become more uniform in metaphorization.

To further observe the convergence/divergence of the light verbs in the four regions, we classified the light verbs into three groups: (A) in flux with significant variations; (B) uniform progression towards the final stage; (C) uniform incipient metaphorization with backward development.

4.2 Group A: Metaphorization in flux

The first group consists of GIVE light verbs *jiyu* 給予 and *yu* 予. As for *jiyu* 給予, all four regions are in the **mid-stage** of the process, with BJ and SH having the highest **MI** (both=**65**), followed by TW (**43**), and HK the lowest (**34**). *Yu* 予 shows a similar pattern except that it is more metaphorized in SH (88), BJ (84), and TW (60) but not in HK (23). The metaphorization process is led by BJ and SH.

Figure 8 and Figure 9 show the breakdown of

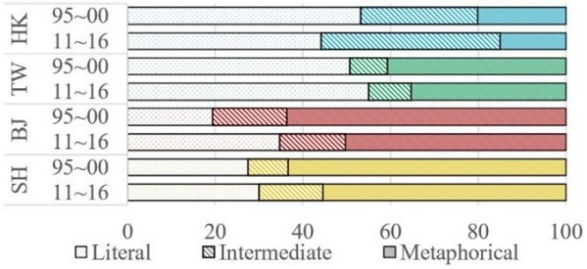


Figure 7: Variations in metaphorization of *jiyu*給予

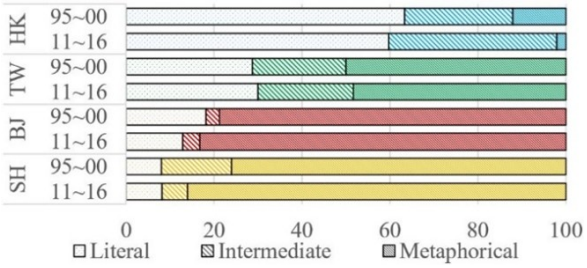


Figure 8: Variations in metaphorization of *yu*予

changes in each type of usage for both light verbs. HK shows an expansion of the intermediate usage. In addition, there is a complementary development in BJ and SH: they both have decreased metaphorical usage for *jiyu* 給予 yet increased metaphorical usage for *yu*. TW, on the other hand, remains relatively stable.

These observations are further confirmed by statistical analyses. BJ and SH show no significant difference in metaphorization for both *jiyu* 給予 ($B = -0.14$, $p = 0.420$) and *yu* 予 ($B = -0.01$, $p = 0.987$). However, HK and TW exhibit significantly lower odds of metaphorization for *jiyu* 給予 (HK: $B = -1.63$, $p < 0.001$; TW: $B = -1.20$, $p < 0.001$) and *yu* 予 (HK: $B = -3.04$, $p < 0.001$; TW: $B = -1.24$, $p < 0.001$).

4.3 Group B: Uniform progression

The second group consists of *gan* 幹 WORK HARD and *nong* 弄 FIDDLE. All the regions show uniform progression towards the final stage. For *gan* 幹, TW has the highest MI (81), followed by HK (76) and BJ (74), and lastly SH (69). *Nong* 弄 is more advanced overall, with HK having the highest MI (94), closely followed by SH (91), TW (88), and BJ (88). The breakdowns by usage are provided in Figure 10-Figure 11. BJ and SH share significant development of *gan* 幹 from intermediate usage to fully metaphorized usage, catching up on TW. HK shows a similar but weaker trend. *Nong* 弄 in BJ has additional progression of metaphorization, contrasting the slight regression in TW.

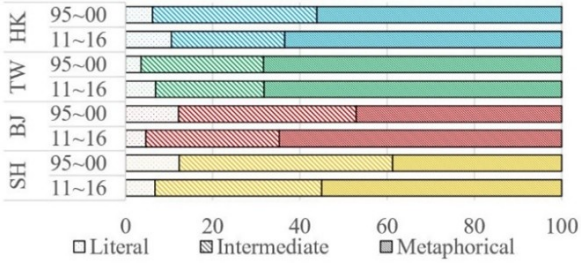


Figure 9: Variations in metaphorization of *gan*幹

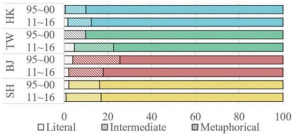


Figure 10: Variations in metaphorization of *nong*弄

Statistically, the pairwise comparison of regional variation supports the trend above. For *gan* 幹, BJ and SH show significant difference as compared to TW in 95-00, but in 11~16 only SH is significantly lower than TW in metaphorization ($B = -0.5$, $p = 0.032$). As for *nong* 弄, in 95~00, BJ has significantly lower metaphorization degree than HK ($B = -1.14$, $p < 0.001$), but it is no longer significant by 11~16 ($B = -0.42$, $p = 0.314$). Instead, a meaningful difference emerges between HK and TW during the latter period ($B = -0.74$, $p = 0.006$).

4.4 Group C: Incipient and backward development

The last group is a single light verb, *jia* 加ADD. Unlike other light verbs, its metaphorization is in an **initial stage** in all four regions, as indicated by its lowest **MI** values: **29** in TW, **13** in BJ, **11** in SH, and **10** in HK. Strikingly, all the regions show **backward development**, as illustrated by the breakdown in Figure 12. TW undergoes drastic regression, "catching up" on HK, BJ, and SH.

The model results of time-wise comparison within each region also highlight the uniform regression in metaphorization of *jia*加. There is a significant decline from 95∼00 to 11∼16 for BJ ($B = -0.986, p = 0.0001$), HK ($B = -0.739, p = 0.006$), and TW ($B = -1.731, p < 0.001$). For SH, the decline of metaphorization is slower with a marginal statistical significance ($B = -0.468, p = 0.073$).

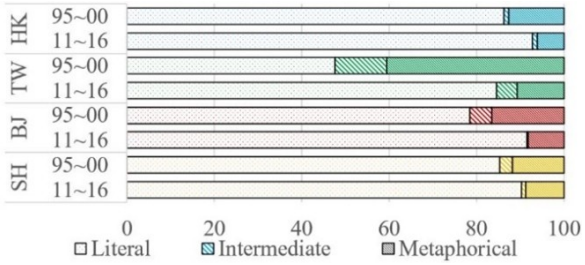


Figure 11: Variations in metaphorization of *jia*加

5 Concluding remarks

5.1 On the development of light verbs

One of our key findings is that light verbs are developed differentially, which may stem from some deeper mechanisms of categorial shifts as well as language variation and change. First, the grouping can be tied to the difference in their **image-schematic structures**. *Jiyu* 給予 and *yu*予 share the TRANSFER schema and show parallel variations in metaphorization (BJ/SH > TW > HK). On the other hand, *gan*幹 and *nong*弄, with similar uniform progression towards final stages, also both have a one-participant image schema, unlike TRANSFER with a second participant (recipient). Given that metaphoric extension is a systematic mapping between domains constrained by a semantic structure (e.g., Sweetser, 1988; Heine et al., 1991), light verbs with similar semantic structures are predicted to pattern together in diachronic development, which results in parallel synchronic variations. The grouping is corroborated by a previous insight that Chinese light verbs can be categorized into the GIVE-group and the DO-group (Jiang, 2020; Lu and Huang, 2023; Kuo, 2025). We suggest that the semantic underpinning of these two groups lies in whether a second participant is required in the underlying conceptual structures.

Second, some trajectories are attributable to **competition** among light verbs with similar semantic structures. *Yu*予 is ahead of *jiyu*給予 in metaphorization. They even show complementary development in BJ and SH: progression for *yu*予 but regression for *jiyu*給予. The preference of *yu*予 in expressing the metaphorical meaning is plausibly due to lexical competition and specialization of meaning to reduce the functional load of each lexical item (e.g., Clark, 1987). This perspective offers a potential explanation for the unexpected trends of jia加, which has backward development and contrasts the claim in Claudi and Heine (1986) that metaphorization is unidirectional. While ADD and GIVE both have a two-participant schema, they differ in the lexical inventories in Chinese. GIVE is lexic alized as the morphemes gei給 and yu予, whereas ADD is only lexicalized as jia加. Unlike yu予/jiyu給予, gei給 remains a lexical verb and lacks the light verb usage (e.g., ungrammatical *zai-gei qianghua 再給强化vs. zai-yu qianghua再予强化 "to strengthen (it) again").[4] It could well be that the lexical verb *gei*給 "frees up" the room of metaphorization for *yu*予 and *jiyu* 給予, but jia加 still bears the functional load to express the literal meaning, and hence the resistance to metaphorization. [5] Third, some convergence we witnessed can be traced to **contact** between speech communities. One such case is yu予 in BJ and SH, which associate with similar socio-cultural backgrounds. We notice that the significant increase of *yu*予's Type III uses in SH (11 16) is largely due to popular usage of *yu jianwaizhihang*予監外執行 "temporary releases (from prison)". In LIVAC, this usage first appeared in BJ in 1999 (see Appendix B Example (6)) but was not found in SH from 95∼00, which presumably entered SH in a later time. Another possible case is the advanced development of *gan*幹 and *nong*弄 in HK, which is unexpected given HK's general insensitivity to metaphorization (with the lowest MI), and low frequency of these two light verbs (see Table 1, in total only 10%). Both verbs are not commonly used in Cantonese, the predominant spoken language in HK, and hence are likely borrowed from other regions. If the time of borrowing is after the verbs well on the way to metaphorization, the uniform progression in HK can be explained.

5.2 Large-scale dynamic measurement of variations with LIVAC, and prospects

Using an innovative measure, MDI, we revealed how speech communities converge and diverge in metaphorization of light verbs, including the clustering of HK-TW and BJ-SH and the tendency of the four speech communities to converge as led by the Mainland. While the convergence has been suggested before (e.g., Diao, 2012a, 2012b), this study is the first of its kind to recruit a quantitative tool to measure **distances** between regions in light

[4] Interestingly, *gei*給 enters another grammaticalization path to become a benefactive marker, e.g., *gei ta zuo dangao*給他做蛋糕 "make a cake for him/her".

[5] How backward development arises might involve multiple factors and invites more future explanation.

verb development. By comparing the two time periods (95∼00 / 11∼16), the measurement is bidimensional for tracking down how the trajectory of a dynamic process varies. Such a measurement can only be developed with a large-scale synchronous corpus like LIVAC, with both latitudinal and longitudinal dimensions from 30 years of texts in 6 major Chinese speech communities. Its rigorous "Windows" approach validates the comparison between regions and between time periods. LIVAC therefore provides an important basis for the study of grammatical and semantic processes in Chinese.

The characteristic absence of overt markings of grammatical categories in Chinese does not shield it from universal grammatical processes such as metaphorization. This is also reflected by other cases where words which could have different grammatical functions such as being a noun or a verb, e.g., *ai*愛 "love" and 服務 "service/serve" (ta fuwu *shehui*他服務社會 "s/he serves the society" vs. *ta wei shehui tigong fuwu*他為社會提供服務 "s/he provides service to the society"). With a rigorously cultivated database, it can be shown that the preference in BJ for nominal uses, and hence more formal register, exceeds that in TW and HK (Kwong and Tsou, 2003). This contributes to the appreciation of **covert variations** within Chinese communities (Tsou and Ji, 2022) and deserves to be studied so as to better understand the Chinese language and its internal developments.

Achieving a better understanding of covert grammatical variations requires quantitative comparisons to measure the extent of a usage in a variety, which, under globalization, inevitably influences other varieties via cross-regional communication. A crucial task is to track the variations that may result from a change of feature-spreading from one region to another. This calls for a corpus **monitoring multiple regions**, which is uncommon in most existing ones. The current study showcases how a corpus with comparable cross-regional longitudinal data like LIVAC allows for systematic investigation of grammatical variation and change in **Global Chineses** (e.g., Tsou and You, 2003; Yip and Tsou, 2022).

5.3 Towards the future

Going forward, we plan to include more recent data after 2016 to closely monitor the development of metaphorization, as enabled by the unique synchronous feature of LIVAC. We expect to reveal more intricate converging and diverging variations between Chinese varieties.

We also expect the current approach to illuminate studies in different mechanisms of metaphorization, for example, via a metaphorical agency in the object as for another light verb *da* (Kwong and Tsou, 2003; Tsou and Yip, 2020). Instead of direct metaphorical extension via image-schematic structures, *da*'s metaphorization relies on the type of its object, forming a continuum from concrete, quasi-concrete (as in *dazao qicheye de hangkongmujian* 打造汽車業的航空母艦 "(forge) an aircraft carrier of the automotive business"), to abstract nouns. How argument structures interact with metaphorization, and furthermore, how variations in argument structures (see Jiang et al. 2016, 2021 for variations in verb transitivity) correlate with variations in metaphorization, are yet other intriguing questions awaiting exploration.

Acknowledgments

For data preparation and annotation, we wish to thank Wing Fu Tsoi, Janice Chong, Kathryn He, Clara Hui, Steffi Lo, Kelly Mak, Eunice Wong, and Yuki Wong. We also wish to thank the two anonymous reviewers for their valuable comments. All errors remain the authors' own responsibilities.

References

Alvarez-Morera, Georgina. 2023. The nominal in light verb constructions: a corpus-based study in present-day English, German, Catalan and Spanish. Doctoral dissertation, Universitat Rovira i Virgili.

Brinton, Laurel J., and Elizabeth C. Traugott. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Butt, Miriam, and Wilhelm Geuder. 2003. Light verbs in Urdu and grammaticalization. In Regine Eckardt, Klaus von Heusinger & Christoph Schwarze, eds., *Words in Time: Diachronic Semantics from Different Points of View*. 295–350. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110899979.295

Butt, Miriam. 2010. The light verb jungle: still hacking away. In Mengistu Amberber, Brett Baker and Mark Harvey, eds., *Complex Predicates: Cross-linguistic Perspectives on*

Event Structure. 48–78. Cambridge, UK: Cambridge University Press.

Bybee, Joan L., and William Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. In Jacek Fisiak, ed., *Historical Semantics - Historical Word-Formation*. 59–83. Berlin & New York: De Gruyter Mouton. https://doi.org/10.1515/9783110850178

Cai, Huiying, Yunhua Qu, and Zhiwei Feng. 2019. A corpus-based study of the semantic prosody of Chinese light verb pattern across registers: taking jinxing and shoudao as examples. *Glottometrics*, 46: 61–82.

Cattell, Ray. 1984. *Composite Predicates in English*. North Ryde, New South Wales: Academic Press Australia.

Clark, Eve V. 1987. *The Principle of Contrast: A Constraint on Language Acquisition*. Hillsdale, NJ: Erlbaum.

Claudi, Ulrike, and Bernd Heine. 1986. On the metaphorical base of grammar. *Studies in Language*, 10(2): 297–335. https://doi.org/10.1075/sl.10.2.03cla

Diao, Yanbin. 2004. *Xiandai Hanyu Xuyi Dongci Yanjiu* [A Study on Delexical Verb in Modern Chinese]. Dalian: Liaoning Normal University Press.

Diao, Yanbin. 2012a. Lianganshidi xiandai hanyu changyongci "jinxing" shiyong qingkuang duibi kaoca yu fenxi [Comparison and analysis of the usage of "jinxing" in the four places across the Taiwan Strait]. *Journal of Wuling*, 37(3): 116–124.

Diao, Yanbin. 2012b. Tai-gang-ao diqu "gao" de shiyong qingkuang ji qi yu neidi de chayi [On the usage of gao in the regions of Taiwan, Hong Kong and Macao and the differences with the Mainland]. *Journal of Weinan Normal University*, 27(9): 80–88.

Halverson, Sandra. 1999. Image schemas, metaphoric processes, and the "Translate" concept. *Metaphor and Symbol*, 14(3): 199–219. https://doi.org/10.1207/S15327868MS140303

Heine, Bernd, Ulrike Claudi, and Friederike Hünnemeyer. 1991. *Grammaticalization: A Conceptual Framework*. Chicago, IL: University of Chicago Press.

Hook, Peter E. 1991. The emergence of perfective aspect in Indo-Aryan languages. In Elizabeth C. Traugott and Bernd Heine, eds., *Approaches to Grammaticalization 2*: 59–89. Amsterdam, Philadelphia: John Benjamins.

Hopper, Paul J., and Elizabeth C. Traugott. 2003. *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

Huang, Chu-Ren, Jingxia Lin, Menghan Jiang, and Hongzhi Xu. 2014. Corpus-based study and identification of Mandarin Chinese light verb variations. In Marcos Zampieri, Liling Tan, Nikola Ljubešic and Jörg Tiedemann, eds, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 1–10. Dublin: Association for Computational Linguistics and Dublin City University.

Jespersen, Otto. 1954. *A Modern English Grammar on Historical Principles*. London: George Allen & Unwin & Copenhagen: Ejnar Munksgaard.

Jiang, Haiyan. 2020. Chouxiang dongzuo shijian yuyi fanchou yu Hanyu xingshi dongci [Abstract action event semantic category and Chinese dummy verbs]. Doctoral dissertation, Jilin University.

Jiang, Menghan, Dingxu Shi, and Chu-Ren Huang. 2016. Transitivity in light verb variations in Mandarin Chinese – a comparable corpus-based statistical approach. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*: 459–468. Seoul: Association for Computational Linguistics and Kyung Hee University.

Jiang, Menghan, Hongzhi Xu, Jingxia Lin, Dingxu Shi, and Hunag Chu-Ren. 2021. Computational processing of varieties of Chinese: comparable corpus-driven approaches to light verb variation. In Marcos Zampieri and Preslav Nakov, eds., *Similar Languages, Varieties, and Dialects: A Computational Perspective*. 304–326. Cambridge: Cambridge University Press.

Kuo, Pei-Jung. 2025. The GIVE group of

the Mandarin light verbs. In Anna Riccio and Jens Fleischhauer, eds., *Light Verbs: Synchronic and Diachronic studies*. 43–64. Berlin & Boston: Düsseldorf University Press. https://doi.org/10.1515/9783111388878-003

Kwon, Nayoung. 2004. A semantic and syntactic analysis of the causative structure in Vietnamese. In *Western Conference in Linguistics 2004 (WECOL 2004)*, Los Angeles, CA: University of Southern California.

Kwong, Oi Yee, and Benjamin K. Tsou 2003. A synchronous corpus-based study of verb-noun fluidity in Chinese. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, 194–203. Tokyo: Association for Computational Linguistics and Waseda University.

Lai, Ryan K. Y. 2024. Why we need asymmetric measures to classify multi-word expressions: the case of Tibetan light verb constructions. *Society for Computation in Linguistics*, 7(1): 302–306. https://doi.org/10.7275/scil.2211

Lakoff, George, and Mark Johnson. 1980. *Metaphor We Live By*. Chicago & London: University of Chicago.

Langacker, Ronald. 1991. Cognitive Grammar. In Flip G. Droste and John E. Joseph, eds., *Linguistic Theory and Grammatical Description*. 275–306. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Langacker, Ronald. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Li, Feng, and Yun Chai. 1995. "V+yi" lei xuhua dongci de binyu [The objects of "V+yi" dummy verbs]. *Journal of Xinjiang Education Institute*, 4: 68–71.

Lin, Jingxia, Dingxu Shi, Menghan Jiang & Chu-Ren Huang. 2019. Variations in world Chineses. In Chu-Ren Huang, Zhuo Jing-Schmidt & Barbara Meisterernst, eds., *The Routledge Handbook of Chinese Applied Linguistics*, 196–211. London & New York: Routledge.

Lu, Lu, and Chu-Ren Huang. 2023. A diachronic insight into the aspectual meaning in Light Verb Constructions. A case study in Mandarin Chinese. In Anna Pompei, Lunella Mereu and Valentina Piunno, eds., *Light Verb Constructions as Complex Verbs: Features, Typology and Function*. 305–336. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110747997-012

Lü, Shuxiang. 1980. *Xiandai Hanyu Babaici* [Modern Chinese 800 Words]. Beijing: Commercial Press.

Lü, Shuxiang. 1999. *Xiandai Hanyu Babaici (Zengdingben)* [Modern Chinese 800 Words (Enlarged Edition)]. Beijing: Commercial Press.

Mohanan, Tara. 2017. Grammatical and light verbs. In Martin Everaert and Henk van Riemsdijk, eds., *The Wiley Blackwell Companion to Syntax* (2nd ed.). 1–27. Chichester, UK: John Wiley & Sons, Ltd.

Nugraha, Danang S. 2025. A theoretical and corpus linguistics study of the light verb constructions: empirical data from Indonesian. Doctoral dissertation, University of Szeged.

Peña Cervel, Sendra. 2004. The image-schematic basis of the event structure metaphor. *Annual Review of Cognitive Linguistics*, 2: 127–158. https://doi.org/10.1075/arcl.2.05pen

Pompei, Anna. 2023. How light is 'give' as a Light Verb? A case study on the actionality of Latin Light Verb Constructions (with some references to Romance languages). In Anna Pompei, Lunella Mereu and Valentina Piunno, eds., *Light Verb Constructions as Complex Verbs: Features, Typology and Function*. 149–200. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110747997-006

Sweetser, Eve E. 1988. Grammaticalization and semantic bleaching. In Shelly Axmaker, Annie Jaisser and Helen Singmaster, eds., *Proceedings of the 14th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA: Berkeley Linguistics Society, 389–405. https://doi.org/10.3765/bls.v14i0.1774

Traugott, Elizabeth C. 1982. From propositional

to textual and expressive meanings: some semantic-pragmatic aspects of grammaticalization. In Winfred P. Lehmann and Yakov Malkiel, eds., *Perspectives on Historical Linguistics*. 245–271. Amsterdam & Philadelphia: Benjamins.

Tsou, Benjamin K. and Ji, Yaxuan. 2022. Yueyu he Xianhan zhong guanjian de yinxing chayi: yi jinyi fuhe ci wei li [Some salient covert differences between Cantonese and Modern Standard Chinese: near-synonymous compounds as examples]. Paper presented at The 26th International Conference on Yue Dialects, November 26-27, 2022, Jinan University.

Tsou, Benjamin K., and Ka-Fai Yip. 2020. A corpus-based comparative study of light verbs in three Chinese speech communities. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. 302–311. Hanoi: Association for Computational Linguistics and VNU University of Science.

Tsou, Benjamin K., and Oi Yee Kwong. 2015. LIVAC as a monitoring corpus for tracking trends beyond linguistics. *Journal of Chinese Linguistics Monograph Series*, 25: 447–472.

Tsou, Benjamin K., and Rujie You. 2003. *Huayu Yu Huaren Shehui* [Chinese Language and Society]. Hong Kong. City University of Hong Kong Press.

Tsou, Benjamin K., Kelly Mak, and Kenny Mok. 2023. Towards the identification and tracking of salient traits and their developments in Chinese society via Big Data. *Advances in Techno-Humanities*, 1: 4–18. Routledge. https://doi.org/10.4324/9781003376491-2

Wang, Leslie Fu-mei. 2004. A corpus-based study of mandarin verbs of doing. *Concentric: Studies in Linguistics*, 30(1): 65–85.

Wang, Li. 1985. *Zhongguo Xiandai Yufa* [Modern Chinese Grammar]. Beijing: Commercial Press.

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In W. P. Lehmann and Y. Malkiel, eds., *Directions for Historical Linguistics*, 97–195. Austin, TX: University of Texas Press.

Xu, Hongzhi, Menghan Jiang, Jingxia Lin, and Chu-Ren. 2022. Light verb variations and varieties of Mandarin Chinese: Comparable corpus driven approaches to grammatical variations. *Corpus Linguistics and Linguistic Theory*, 18(1): 145–173. https://doi.org/10.1515/cllt-2019-0049

Yan, Zhongsheng. 1998. Shuo "houxu dongcixing binyu dongci" [On "verbs taking verbal objects"]. *Journal of Hebei Normal University (Natural Science)*, 2: 91-93.

Yip, Ka-Fai, and Benjamin K. Tsou. 2022. MSC variations in metaphorization among Pan-Chinese speech communities. Paper presented at the 28th Annual Conference of the International Association of Chinese Linguistics, May 20-22, 2022, The Chinese University of Hong Kong.

Yip, Wai Chi J., and Sze-Wing Tang. 2022. Lexical variations in Asian Chinese speaking communities: a corpus-informed study of online, offline, and digital. *Global Chinese*, 8(2): 161–187. https://doi.org/10.1515/glochi-2022-0001

Zhou, Gang. 1987. Xingshi dongci de cifenlei [Subdivision of dummy verbs]. *Chinese Language Learning*, 1: 11–14.

Zhu, Dexi. 1985. Xiandai shumian hanyu li de xuhua dongci he mingdongci [Dummy verbs and nominal verbs in Modern Written Chinese]. *Journal of Peking University (Humanities and Social Sciences)*, 5: 1–6.

A Sample annotation guidelines for yu 予

(*only instructions translated*)

Type I

- General meaning: Transfer ownership
- Object type: usually concrete nouns
- Auxiliary means for classification: Be able to identify and/or recover a **giver/subject**, a recipient/indirect object, and a patient/direct object.
- Examples:

- //外交部應會分階段，分國家，分地區、分時期，漸進◆予東協8國及印度免簽、落地簽或電子簽//
- No explicit patient: //甚至由法官/檢察官口述◆予書記官記錄//
- No explicit recipient or patient: //說得嚴苛點是「天◆予不取、反受其咎」。//
- No explicit give or recipient: //將◆予5-10萬元（人民幣）的獎勵//
- Property rights and power's (temporary) transfers count as Type I, e.g., //授權◆予董事會通過// and //始得將土地產權移轉◆予他人//
- Information transfer (via any media) count as Type I, e.g., //傳短訊◆予胞弟道別//

Type II

- General meaning: Causative/let/allow
- Object type: usually abstract nouns, as in: 「予...的感覺」"give... feeling"、「予...的印象」"give ... impression"、「予人口實」"give one's critics a handle"、「予人方便」"give convenience", etc.
- Auxiliary means for classification: usually can be replaced with *gei* 給"give"
- Examples:
  - //他不願◆予人以「鷹派」的印象//
  - //甚至◆予人有草率之譏//
  - //同時也跨海放映該會「重生」紀錄片◆予北京清大師生觀賞。//

Type III

- General meaning: Do/ implement an action
- Object type: usually eventive noun (including verb noun), as in: 「予支持」"support"、「予懲罰」"punish"、「予排除」"eliminate", etc.
- Auxiliary means for classification: cannot be replaced with *gei* 給"give"
- Examples:
  - //最近星國政府正打算修法再◆予强化//
  - //均◆予不起訴。//

Borderline cases

- ∼建議 "suggestion": Type I (information)
- ∼評級/評價 "rate": Type I
- ∼榮譽 "honour": Type I
- ∼時間 "time": Type II
- ∼機會 "opportunity": Type II
- ∼壓力 "stress": Type II
- ∼回答 "answer": Type II
- ∼保障 "protection": Type II
- ∼自治 "autonomy", ∼獨立 "independence": Type II
- ∼豁免 "exemption": Type III
- ∼安排 "arrangement": Type III
- ∼肯定 "affirmation", ∼關心 "care": Type III
- ∼治療 "treatment": Type III
- ∼指引 "guidelines": if the context implies actual guidelines (e.g., by distribution), count as Type I; otherwise Type III
- ∼援助 "aid": if the context implies money or supplies, count as Type I; otherwise Type III

B Sample sentences of the five light verbs from LIVAC

The polysemy of **jiyu** 給予

(1) Type I: ownership transfer
*Wunv jianzhuang, jingkong xia quchu siqian yibai yuan **jiyu** zeiren* 吳女見狀，驚恐下取出四千一百元給予賊人 "Noticing the situation, Ms. Ng gives the robber $4100 in fear." (HK 1997)

(2) Type II: causative
*Jiji zhengqu guojia **jiyu** benshi xianshixianhang de teshu zhengce*積極爭取國家給予本市先試先行的特殊政策 "(Our city) strives for the "try first" special policy offered by the central government." (SH 2013)

(3) Type III: light verb DO, "impose"
*Yifa yigui jinxing jiemian yuetan, tongbao piping shenzhi **jiyu** xingzheng chufen*依法依規進行誡勉約談、通報批評甚至給予行政處分 "(The new government) legally performs exhortations, condemnations and even presents administrative punishments." (BJ 2014)

The polysemy of **yu** 予

(4) Type I: ownership transfer
*Guotai churang guquan **yu** Zhonxintaifu duonian*國泰出讓股權予中信泰富多年 "Cathay Pacific has been transferring their shares to CITIC Pacific for many years." (TW 1997)

(5) Type II: causative
*Qizhong you liang-wei zhuangaoren **yu** wo yinxiang shenke*其中有兩位撰稿人予我印象深刻 "Among them, two writers impressed me" (TW 2012)

(6) Type III: light verb DO, "impose"
*Dui youguan bumen banli jianxing, jiashi, zan-**yu** jianwaizhihang, baowaijiuyi zhong de weifaqing kuang, tichu jiuzheng yijian 9672 jianci*對有關部門辦理減刑、假釋、暫予監外執行、保外就醫中的違法情況，提出糾正意見9672件次 "According to the illegal situations in relative agencies' commutation of sentences, paroles, temporary releases and compassionate releases, rectifications are declared 9672 times." (BJ 1999)

The polysemy of **jia** 加

(7) Type I: addition of objects
*Jingye de shenjianhong shuo, ru **jia** zhe chang xi neng rang juqing geng you xiaoguo, mei wenti*敬業的沈建宏說，如加這場戲能讓劇情更有效果，沒問題 "The dedicated Shen says, 'No problem if adding this scene makes the plot more impressive.'." (TW 2015)

(8) Type II: addition of events (manner)
*Zai bixuankao de yi men kemu zhiwai , zai yaoqiu **jia**-kao yi zhi liang ge kemu*在必選考的一門科目之外，再要求加考一至兩個科目 "Apart from the single compulsory subject, one or two subject(s) is/are required to be additionally taken." (SH 1998)

(9) Type III: light verb DO, "impose"
*Bushaoren dui qiaobusi da-**jia** zanshang*不少人對喬布斯大加讚賞 "Many people greatly praise Jobs." (BJ 2012)

The polysemy of **gan** 幹

(10) Type I: "work hard"
*Quancun nanvlaoyou da-**gan** ku-**gan***全村男女老幼大幹苦幹"The whole village, men and women, young and old, all worked hard through the tough jobs." (BJ 1995)

(11) Type II: "work/engage in a job"
*Cong biaomian shang kan **gan** de shi guoji zixun yewu zhe yi hang*從表面上看幹的是國際諮詢業務這一行 "It looked like they were doing the international consulting business from the surface." (SH 1996)

(12) Type II: "conflict"
*Buguo Yashi xinwenbu que andili "dui-zhe **gan**"* 不過亞視新聞部卻暗地裏「對著幹」"But ATV news department was working against them secretly." (HK 2011)

(13) Type III: Light verb DO
*Ni haipa wu-**gan** shashi*你害怕誤幹傻事 "You're afraid of doing something stupid by mistake." (TW 2015)

(14) Type III: Light verb DO, "commit (crimes)"
*Youde huangniu luzi geng wai, zhijie **gan**-qi-le jiapiao goudang* 有的黃牛路子更歪，直接幹起了假票勾當 "Some ticket scalper even took a dirtier route and went straight into selling fake tickets." (SH 2014)

The polysemy of **nong** 弄

(15) Type I: "fiddle, play" (often with hands)
*Chenrongsen geng chen Keyi **nong** toufa zhiji, liuzhidaji*陳容森更趁可頤弄頭髮之際，溜之大吉 "Chan Yung Sam quietly ran away while Ho Yee was fixing her hair." (HK 2000)

(16) Type II: creation, "make"
*huidao jia hai de wei qinv he liang ge gege **nong** chide*回到家還得為妻女和兩個哥哥弄吃的 "He still had to prepare food for his wife, daughter, and two brother when he got home. " (SH 1996)

(17) Type II: "obtain, get"
Qiudui de mubiao shi "zhan dian xianqi, tian

dian baqi, ***nong*** *dian yunqi, xue dian caiqi"* 球隊的目標是「沾點仙氣，添點霸氣，弄點運氣，學點才氣」 "The team's goal is to "Catch some magic, add some aura, get some luck, and have some talent" (SH 2013)

(18) Type III: light verb DO
Laoyang, zhe ge shi ni kan zenme ***nong*** *a?*老楊，這個事你看怎麼弄啊？"Old Yang, how should we do about this?" (BJ 2013)

(19) Type III: light verb CAUSE
bushao guanzhong bei ta paianjiaojue de fangshi ***nong****-de xinyangyang*不少觀眾被他拍案叫絕的方式弄得心癢癢 "Many viewers got excited by his amazing selling style." (TW 2016)

Effect of Emotional Congruency and Cognitive Load on Word Processing

CHEN Jieyu[1], TIAN Yujia[1], QI Jing[1], TAO Ran[1]
[1]Research Centre for Language, Cognition, and Neuroscience,
Department of Language Science and Technology, The Hong Kong Polytechnic University
jieyu01.chen@connect.polyu.hk; yujaa.tian@connect.polyu.hk;
jing.qi@connect.polyu.hk; ran.tao@polyu.edu.hk

Abstract

Existing research suggests that the modulation of emotional words to cognitive responses is multifaceted. As an important component of cognition, the influence of emotional words on working memory performance has received increasing attention from researchers. Various modalities of emotional stimuli, particularly facial expressions, are typically presented alongside emotional words to elucidate their associations. Previous studies have demonstrated that the congruency effect occurs when emotional words and faces share the same valence. However, the effect of other emotional modalities on emotional word processing in working memory under varying cognitive loads remains understudied. We implemented the delayed emotional conflict task, a dual-task paradigm that comprises a primary lexical recognition task and a secondary facial recognition task. Results reveal that emotional words, especially negative words, can disrupt working memory performance, and this effect strengthens as cognitive load increases. Notably, in the context of low cognitive load, neutral faces are likely to facilitate the processing of positive words. Additionally, in contrast to prior research, this study does not observe the congruency effect in conditions where the words and faces have the same valence (e.g., negative words and angry faces). These results indicate that both intrinsic valence and the valence of other modalities can modulate word processing in working memory tasks, and these modulations display distinct patterns across different cognitive loads. However, due to the features of stimuli and paradigm, no congruency effect is observed here.

1 Introduction

1.1 Research Background

Emotional words are a category of words characterized by affective connotations, and their processing mechanisms differ from those of neutral words (Frijda et al., 1995; Landis, 2006). There are controversial findings on the implications of emotional word processing. Some investigations have indicated that emotional words exert an inhibitory effect on cognitive behaviors when compared to neutral words (Algom et al., 2004; Herbert and Sütterlin, 2011), while most studies reported the reverse findings, pointing out that more rapid processing of emotional words compared to neutral words (Kissler and Herbert, 2013; Kousta et al., 2009). Further comparisons revealed that negative words are subject to superior processing relative to positive words (Dijksterhuis and Aarts, 2003; Nasrallah et al., 2009), thereby substantiating the "negativity bias." Emotional words have been associated with memory (Adelman and Estes, 2013; Ferré et al., 2018; Talmi and Moscovitch, 2004), including implications for working memory.

Working memory (WM) is a cognitive system responsible for the temporary storage and manipulation of information under attentional control, lasting only a few seconds and considered essential for a variety of complex mental activities (Baddeley, 2020; Cowan, 1999; Jonides and Smith, 2013). According to Baddeley's model, four components are delineated in working memory: the central executive, which controls attention; the phonological loop, which handles linguistic information; the visuo-spatial sketchpad, responsible for processing and storing visual and spatial details; and the episodic buffer, which integrates information from various sources into a cohesive presentation (Baddeley, 2000; Baddeley and Hitch, 1974). Within the realm of working memory research, there is a growing concern regarding how emotions affect its underlying mechanisms.

Studies on working memory have similarly debatable discussions as those surrounding the general processing of emotional words. Rączy and Orzechowski (2021) also identified a "negativity bias" in the working memory task, with faster re-

action times for negative words compared to both neutral and positive words, while no significant difference exists between neutral and positive words. However, negative words impair working memory performance compared to positive and neutral words have been demonstrated by several studies (Kopf et al., 2013; Weigand et al., 2013), while some studies have suggested that both negative and positive words may similarly disrupt working memory performance (Fairfield et al., 2015; Garrison and Schmeichel, 2019). Some findings, nonetheless, contest the assertion that there is no difference between neutral and positive words. Jin et al. (2013) highlighted that distinct patterns in working memory performance for negative and positive words, revealing that positive words elicit faster reaction times than neutral and negative words, while negative words are associated with slower reaction times relative to neutral and positive words. These polarizing arguments imply that the confirmation of a distinction in the processing of emotional and neutral words, yet the mechanisms by which emotional words are modulated within working memory could be intricate.

The limited capacity of working memory inevitably possesses competition among semantic information from multiple words in memory. As emotional words convey emotional and semantic information simultaneously, their processing may also be influenced by emotional information from other modalities, such as facial expressions (Ekman, 1992). Facial expressions convey basic emotions, including anger, sadness, fear, disgust, surprise, happiness, and neutrality, enabling us to discern individuals' emotional state during social interactions (Ekman, 1992; Huerta-Chavez and Ramos-Loyo, 2024). Likewise, facial expression is regarded as a powerful factor influencing cognitive processing and behavior (Van Kleef and Côté, 2022). While the patterns affecting working memory are distinct between negative and positive faces, both generally exhibit a facilitation effect due to their bias in requiring attentional resources (Lee and Cho, 2019; Xu et al., 2021). In working memory studies, facial expressions, in addition to serving as task components, are typically employed as an intervention to investigate whether they produce interference or facilitation effects (Jackson et al., 2012). However, when acting as a "distractor," the effects become more nuanced. For instance, it has been found that angry faces interfere with task performance under low cognitive load, while this interference will be diminished under high cognitive load (Van Dillen and Derks, 2012).

The congruence of valence between emotional words and facial expressions affects cognitive mechanisms. This congruent effect is primarily identified through the utilization of a face-word Stroop paradigm (e.g., Fan et al.2016), which indicates that these stimuli are displayed simultaneously. These investigations reveal that responses to incongruent trials are slower than those to congruent trials, with distinct neural activation patterns observed between these two conditions (Chang et al., 2024; Ovaysikia et al., 2011). The encounter with incongruent face-word pairs in terms of valence activates brain regions associated with monitoring and generating emotional conflicts, such as the dorsomedial prefrontal cortex, the dorsolateral prefrontal cortex, and the rostral anterior cingulate cortex, ultimately resulting in slower reaction times during incongruent trials (Egner et al., 2008; Etkin et al., 2006; Fan et al., 2018; Zhu et al., 2010).

1.2 Research Gaps and Aims

Despite the increasing number of studies on emotional word processing, the complex interplay between semantic emotional content and facial expressions across varying levels of working memory loads remains underexplored. This study seeks to 1)examine the interplay between different word valences and face valences within a working memory task, 2)explore how attentional resources are allocated under varying cognitive loads associated with emotional word-face pairs.This study posits three research questions: First, as cognitive load increases, does it lead to a modification of the advantages (or disadvantages) of working memory for emotional words, and can this change be inhibited by emotional facial expressions? Second, under varying cognitive loads, how do emotional faces modulate working memory performance for words with different valences? Third, does an incongruent valence between words and facial expressions lead to a decrement in word processing in working memory, relative to congruent conditions? If so, how does the effect of this valence incongruence interact with varying levels of cognitive load?

1.3 Hypotheses

Building on prior studies, this research proposes three hypotheses. First, emotional words are anticipated to exert a specific effect on working memory relative to neutral words, while increasing cogni-

tive load will diminish this influence and lead to a more pronounced interference effect from emotional facial expressions. Second, the presence of emotional faces is likely to boost working memory for words that have a similar valence, while concurrently disrupting the processing of words with incongruent valence. Furthermore, it is proposed that this modulation will be affected by different levels of cognitive load. Third, a mismatch in valence between words and emotional facial expressions is expected to disrupt the word processing, and as cognitive load rises, this inhibitory effect will be amplified.

2 Method

2.1 Participants

We recruited a sample of 70 college students, with ages ranging from 18 to 30 years (Mean Age = 23.57, SD = 2.97), including 34 males and 36 females. All participants were native Chinese speakers who could read simplified Chinese fluently and were identified as right-handed. They had normal vision or vision corrected to normal and reported no history of psychiatric or neurological disorders. Before the experiment, each participant provided informed consent by signing a consent form.

2.2 Materials

A total of 698 two-character Chinese words (227 negative, 235 neutral, and 231 positive) were meticulously selected from the Chinese Affective Words System (CAWS; Wang et al., 2008), with 23 designated for the practice component, and the remaining 675 words (225 negative, 225 neutral, and 225 positive) employed in the formal part. Of the words utilized in the formal part, 540 words were presented in the memory sets, while an additional 135, which were not included in the memory sets, served as probes. The Chinese Affective Words System (CAWS) assesses the ratings of valence, arousal, and dominance using a 9-point scale (Wang et al., 2008). The selected words for the formal experiment were controlled for valence, with a significant difference observed among negative, neutral, and positive words ($F(2, 672) = 10352$, $p < 0.001$). Additionally, a significant difference in arousal was found between emotional words (negative and positive) and neutral words ($t(673) = 48.135$, $p < 0.001$), according to CAWS norms (Table 1).

For facial stimuli, 80 facial expressions were selected from the Chinese Facial Affective Picture System (CFAPS; Gong et al.2011). According to previous studies, there was a detection advantage associated with angry faces. For instance, faster responses were observed for angry faces than for happy faces, a phenomenon called the "angry superiority effect"(Hansen and Hansen, 1988), and thus angry faces incorporated as representative negative facial stimuli. Specifically, there were 26 angry (14 male and 12 female), 28 neutral (14 male and 14 female), and 26 happy faces (14 male and 12 female). Among the total, 8 faces were designated as practice components, while an additional 72 faces (24 each for the expressions of angry, neutral, and happy, with balanced gender representation) were utilized in the formal part. Furthermore, we ensured that each face was presented fewer than five times throughout the entire procedure. We selected facial expressions based on identification rate in a experiment on face identification reported by Gong et al. (2011) in their study on the CFAPS (participant number = 100, 51 females, mean age = 23 ± 1; identification rate of angry face = 88.55% ± 4.61%; neutral face: 96.44% ± 1.09%; happy face: 100%), choosing the most recognized expressions for each emotion from both male and female faces.

The experiment consisted of 27 conditions: word valence (negative/neutral/positive) × cognitive load (low/moderate/high) × face valence (angry/neutral/happy). All the stimuli were presented on a black background, maintaining the same contrast and brightness. The characters were displayed in white using the PingFang SC font with 57 point font size, and the images were resized to 260 × 300 pixels.

2.3 Procedure

The study employs a delayed emotional conflict task, which is a dual-task paradigm, to address our research questions, encompassing a primary lexical recognition task and a secondary facial recognition task. It inserts a facial expression during the maintenance to evoke effects of congruence or incongruence in valence. In detail, several two-character Chinese words are displayed on the screen simultaneously, and cognitive load is manipulated by adjusting the number of words presented. Specifically, the low cognitive load involves the presentation of two words, while the moderate cognitive load includes four words, and six words are displayed in the high cognitive load condition. Following a string of words, a facial expression-either angry, neutral, or happy-is presented in the center

	Sample	Mean Valence	Mean Arousal
Negative Words	残忍 (cruel)	2.71 ± 0.34	6.37 ± 0.55
Neutral Words	平常 (ordinary)	5.37 ± 0.42	4.20 ± 0.61
Positive Words	美丽 (beautiful)	7.22 ± 0.22	6.32 ± 0.47

Table 1: Sample, mean valence, and mean arousal of negative, neutral, and positive words selected from the Chinese Affective Words System (CAWS).

of the screen. The whole procedure was divided into practice and formal parts. The formal experiment comprised 135 trials, with a overall duration ranging from about 22 to 30 minutes. Each condition was presented five times, and the sequence of trials was randomized for each participant.

As noted by Schwering and MacDonald (2020), digit span reflects the verbal working memory within the specific context of recalling sequences of numbers, rather than serving as a general measure of language-dependent criteria. Therefore, before the formal experiment, participants were required to complete a digit span task to assess their working memory capacity. The digit span task was conducted using a program that included two subtasks: forward recall and backward recall. Participants listened to an audio sequence and, after it ended, entered the numbers in either the same order or the reverse order of presentation. For both subtasks, the program plays two sequences of numbers, starting with two digits and advancing to longer sequences if at least one is answered correctly, while terminating the test if both sequences are answered incorrectly. The mean forward sequence was 10.22 ± 1.78, while the mean backward sequence was 8.89 ± 1.87. We utilized the Reliable Digit Span (RDS; Greiffenstein et al., 1994) to assess overall performance, which is defined as the sum of the longest strings of digits recalled both forward and backward, with completion of both of them required. The scores from the digit span were not analyzed in the current study, as they pertain to a separate research question, while these data were retained for future research.

After finishing the digit span task, participants were seated in a soundproof room to minimize distractions for the formal experiment. Once the experimental process was introduced by the experimenter, they commenced the entire experiment. Each trial began with a fixation cross displayed for 500 ms, followed by a memory set consisting of words categorized into three conditions: low cognitive load (2 words), moderate cognitive load (4 words), and high cognitive load (6 words). An inter-stimulus interval (ISI) of 500 ms followed. Subsequently, a facial expression (angry, neutral, or happy) was presented in the maintenance phase, requiring participants to memorize this face. Another ISI of 500 ms preceded the probe, during which two words were displayed. Participants were instructed to identify which word was presented in the previous memory set by pressing either the left or right key. Following another 500 ms ISI, two facial expressions were shown, and participants were required to identify which facial expression was presented earlier during the maintenance phase by pressing the corresponding key (Figure 1). Participants were provided two rest periods throughout the procedure to minimize the effects of fatigue.

2.4 Analysis

Reaction times (RTs) and accuracy (ACC) for emotional words and facial expressions were recorded during the experiment. Considering the secondary facial recognition task serves to introduce an interfering factor that affects RTs, these data were excluded from the analysis. Additionally, given their working memory capacity superior to the three different cognitive loads, the accuracy of the lexical recognition task was exceptionally high (low cognitive load: 95.27% ± 2.12%, moderate cognitive load: 98.41% ± 1.25%, high cognitive load: 98.38% ± 1.26%), prompting us to concentrate on its RTs. Furthermore, only correct trials from both lexical and facial memory tasks were incorporated into the analysis, ensuring a robust evaluation of the data.

Before analysis, data were pre-processed by removing practice and incomplete trials, and reaction times lower than 200 ms or higher than 2500 ms were considered outliers and excluded. We removed 16 trials (0.5%) from low cognitive load conditions, 36 trials (1.0%) from moderate cognitive load conditions, and 57 (1.9%) trials from high cognitive load conditions. Mean RT for each condition (word valence × cognitive load × face

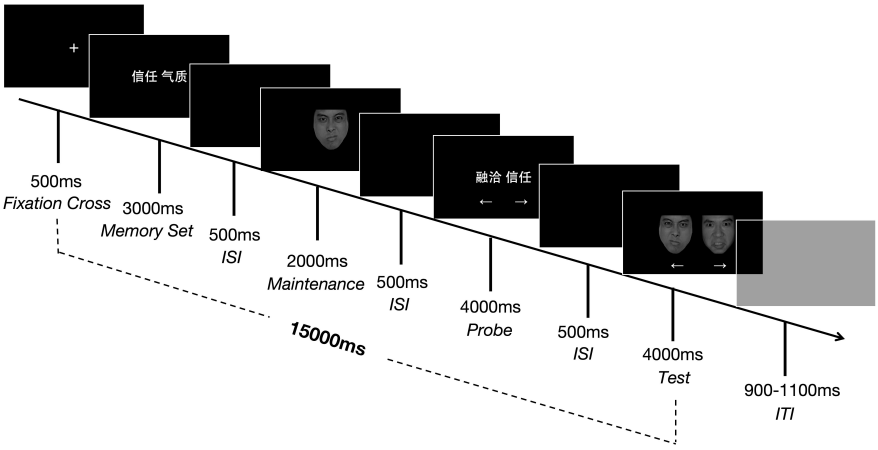


Figure 1: Experimental stimuli and example timeline used in the delayed emotional conflict task. Maintenance is a core component of WM, characterized by active rehearsal of information to prevent decay over time. The probe and test in WM tasks serve a similar function, which are to assess during the retrieval phase whether participants can accurately recall the information that was encoded and maintained in WM.

valence) is calculated based on correct trials. To analyze the RTs associated with the working memory task for emotional words, the mean RTs were subjected to a 3 (word valence: negative, neutral, positive) × 3 (cognitive load: low, moderate, high) × 3 (face valence: angry, neutral, happy) repeated-measures ANOVA. Significant effects were further analyzed using post hoc tests with Tukey's HSD and Bonferroni corrections.

3 Results

3.1 Main Effect

A repeated-measures ANOVA suggests main effects of word valence ($F(1.95, 134.41) = 67.71$, $p < 0.001$, partial $\eta^2 = 0.496$), cognitive load ($F(1.82, 125.50) = 199.12$, $p < 0.001$, partial $\eta^2 = 0.743$), and face valence ($F(1.98, 136.76) = 3.60$, $p = 0.009$, partial $\eta^2 = 0.067$) are significant (Table 2).

Regarding the valence of words, negative words elicit significantly longer reaction times compared to neutral words ($t(69) = 10.772$, $P < 0.001$) and positive words ($t(69) = 3.886$, $p = 0.001$), while positive words show significantly longer reaction times than neutral words ($t(69) = 8.165$, $p < 0.001$). This indicates that the recognition of neutral words is faster than that of emotional words, with negative words showing a notable interference effect within the emotional category. Furthermore, RTs under low cognitive loads are significantly faster than those experienced under moderate ($t(69) = -15.722$, $p < 0.001$) and high cognitive load ($t(69) = -17.652$, $p < 0.001$). Moreover, reaction times for the angry faces condition are significantly slower than for the neutral face condition ($t(69) = -3.016$, $p = 0.010$). However, the differences between angry and happy face conditions ($t(69) = -1.783$, $p = 0.183$) and between neutral and happy face conditions ($t(69) = 1.412$, $p = 0.341$) are not significant.

3.2 Two-Way Interaction Effect

Word Valence × Cognitive Load As can be seen from Figure 2(a), the interaction between word valence and cognitive load reaches a significant level ($F(3.57, 246.03) = 11.81$, $p < 0.001$, partial $\eta^2 = 0.146$). In the context of low cognitive load, the RTs for negative words are significantly longer than positive words ($t(69) = 6.246$, $p < 0.001$) and neutral words ($t(69) = 6.054$, $p < 0.001$). However, there is no significant difference in RTs between neutral words and positive words. Under the moderate cognitive load, negative words require longer times to be processed than neutral words ($t(69) = 5.773$, $p < 0.001$), while no significant difference is observed between negative and positive words. Meanwhile, RTs for neutral words are significantly faster than positive words ($t(69) = -6.877$, $p < 0.001$). When cognitive load is high, emotional words demonstrate notable interference effects compared to neutral words, which is reflected in extended reaction times (negative words: $t(69) = 8.133$, $p < 0.001$; positive words: $t(69) = 5.154$, $p < 0.001$). Within the category of emotional words, there is a significant difference between negative

words and positive words, with negative words demanding more time for recognition than positive words ($t(69) = 3.216$, $p = 0.006$).

Face Valence × Cognitive Load The interaction between face valence and cognitive load is also observed ($F(3.54, 243.99) = 12.97$, $p < 0.001$, partial $\eta^2 = 0.158$), as depicted in Figure 2(b). Angry faces can facilitate the lexical recognition under low cognitive loads compared to the influence of neutral faces ($t(69) = -3.562$, $p = 0.002$), while there is no difference between the angry faces and happy faces. As cognitive load increases, the facilitation effect of angry faces is further demonstrated, with significantly faster RTs for the conditions with angry faces in comparison to neutral ($t(69) = -6.777$, $p < 0.001$) and happy faces ($t(69) = -3.971$, $p = 0.001$). However, when cognitive load is high, angry faces instead bring an inhibitory effect. When faced with angry faces, the RTs for the lexical recognition task are significantly slower than when faced with neutral faces ($t(69) = 3.043$, $p = 0.010$).

Word Valence × Face Valence Figure 2(c) illustrates that a pronounced interaction effect is observed between the word valence and the face valence ($F(3.82, 263.68) = 2.94$, $p = 0.023$, partial $\eta^2 = 0.041$). Regardless, under the influence of what face valences, RTs for negative words did not significantly differ. As opposed to neutral faces, emotional faces can enhance the working memory performance for neutral words (angry faces: $t(69) = -2.441$, $p = 0.052$; happy faces: $t(69) = -3.422$, $p = 0.003$), but there is no significant difference between the angry and happy faces. Moreover, angry faces produce a significant facilitation effect on positive words compared to happy and ($t(69) = -3.194$, $p = 0.006$) neutral faces ($t(69) = -2.934$, $p = 0.014$).

3.3 Three-Way Interaction Effect

As shown in Figure 3, the interaction among word valence, cognitive load, and face valence is significant ($F(6.45, 445.11) = 2.78$, $p = 0.010$, partial $\eta^2 = 0.039$). Simple effects analyses of cognitive load at the interaction of word valence and face valence indicate significant differences across most conditions. For instance, the combination of angry faces and negative words under high cognitive load results in slower RTs compared to low cognitive load ($t(69) = 8.695$, $p < 0.001$) and moderate load ($t(69) = 6.863$, $p < 0.001$). Additionally, slower RTs are observed in the pairing of angry faces and positive words when the cognitive load is high than when it is low ($t(69) = 10.825$, $p < 0.001$) and moderate ($t(69) = 3.237$, $p = 0.006$). Further analyses of the word valence effect, specifically within the context of angry faces and high cognitive load, suggest that negative words elicit significantly longer RTs than neutral words ($t(69) = 4.671$, $p < 0.001$) and positive words ($t(69) = 3.110$, $p = 0.008$) when paired with angry faces.

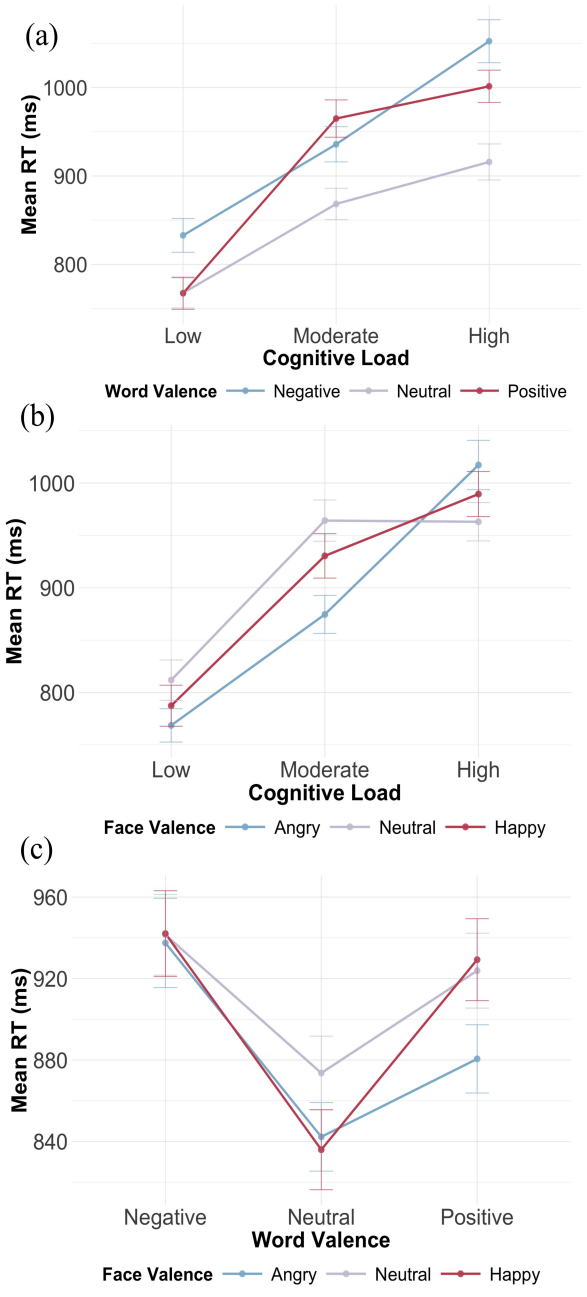


Figure 2: (a) Interaction effect of word valence and cognitive load on mean reaction time. (b) Interaction effect of face valence and cognitive load on mean reaction time. (c) Interaction effect of word valence and face valence on mean reaction time.

4 Discussion

Controversial findings emerge in previous studies in relation to the influence of word valence on cognitive behavior. Some research demonstrates that emotional words can facilitate cognitive processing (Kousta et al., 2009), while other studies reveal that they may interfere with cognitive tasks, leading to extended reaction times (Fox et al., 2001; Maratos et al., 2000). This study supports the findings that emotional words induce longer RTs than neutral words, which may be attributed to the ability of high-arousal emotional words to capture much attention, resulting in processing delays (Kuperman et al., 2014).

Negative stimuli also engage attentional allocation earlier than positive stimuli, while demanding greater cognitive resources (Smith et al., 2003). Besides, their threat-related salience brings about rapid attentional capture, thereby engendering the interference effects (Algom et al., 2004; Anticevic et al., 2010). Positive stimuli are detected later and lack threat connotations, allowing sufficient cognitive resources to inhibit the influence of valence. These help explain why negative words require longer processing time compared to neutral and positive words. However, under conditions of high cognitive load, both negative and positive words elicit longer reaction times relative to neutral words. Previous studies using the emotional Stroop task have discovered that the ink color naming of emotional words is slower than that of neutral words, indicating that emotional content can interfere with cognition (Ben-Haim et al., 2016; Kahan and Hely, 2008). When faced with high cognitive load, cognitive resources approach saturation, with the majority allocated to process the task, so the interference effects from emotional valence become difficult to inhibit. At the neural level, the cognitive control network in the prefrontal cortex becomes occupied by the task, rendering it unable to effectively suppress emotional interference (Pessoa, 2009). This accounts for the observation that negative and positive words elicit longer reaction times under high cognitive load in the current working memory task. Additionally, negative words show a stronger interference effect across all cognitive loads, stemming from the competition among semantics, valence, and attention for limited cognitive resources, which amplifies their disruptive impact (Gross, 1998; Volokhov and Demaree, 2010).

Although negative words show a stronger effect in the presence of angry, neutral, or happy faces, no significant differences in patterns are explored. This may be explained by the fact that the dominance of high arousal in negative words masks the role of facial valence in the reaction. Furthermore, positive words are also affected by increasing cognitive loads. Specifically, as cognitive load increases from low to moderate and from low to high, we observe a prolongation of reaction times, but there is no significant effect when the load shifts from moderate to high. This suggests that positive words remain stable after reaching a moderate load, possibly due to the lower arousal effect of positive words compared to negative words (Ito et al., 1998). These results partially verify our first hypothesis: emotional words indeed exert specific effects on the working memory task. However, increasing cognitive load amplifies the impact of emotional words instead of diminishing the effects of word valence.

When faced with high cognitive load, participants show extended reaction times for negative words influenced by angry and happy faces, as opposed to neutral and positive words. Nonetheless, neutral faces facilitate the processing of positive words when cognitive load is low, yielding shorter reaction times compared in comparison to neutral words. However, this facilitation effect disappears with greater cognitive load. One plausible explanation is that positive words facilitate efficient processing (Fredrickson, 2001; Niedenthal et al., 1997), and neutral faces do not exert additional emotional responses that influence recognition processing, and the low cognitive load provides sufficient resources for the effective processing of positive information. With the rise in cognitive load, the processing of positive words requires more semantic engagement and additional cognitive resources, causing the disappearance of their superiority, in contrast to neutral words that do not necessitate simultaneous emotional processing and thus maintain an advantage.

We also hypothesize that when negative words are paired with angry faces and positive words are paired with positive words, a facilitation effect will be detected, resulting in faster reaction times. However, in the current paradigm, no facilitation effect was observed in these pairs. Negative words paired with angry faces elicit the longest reaction times, while positive words with happy faces are not the fastest. Conversely, their effects still follow a sim-

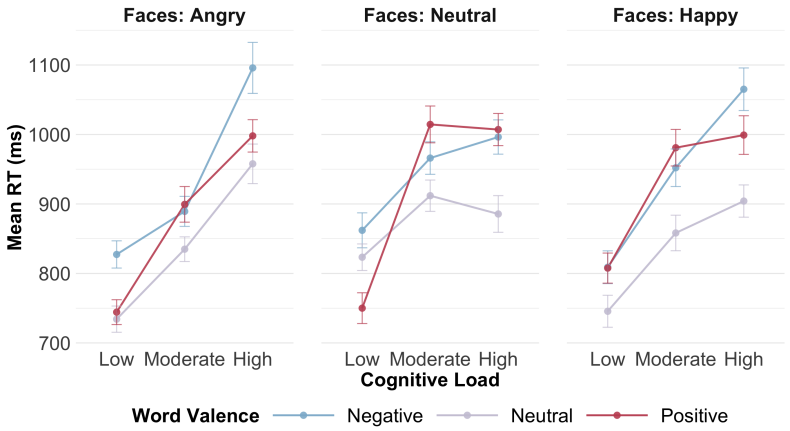


Figure 3: Interaction of word valence, cognitive load, and face valence on reaction time.

Effect	df	MSE	*F*	pes	*p*-value
Word Valence	1.95, 134.41	20548.10	67.71	.495	<.001***
Cognitive Load	1.82, 125.50	35589.70	199.12	.743	<.001***
Face Valence	1.98, 136.76	21619.65	4.95	.067	.009**
Word Valence: Cognitive Load	3.57, 246.03	21980.38	11.81	.146	<.001***
Word Valence: Face Valence	3.82, 263.68	19280.93	2.94	.041	.023*
Cognitive Load: Face Valence	3.54, 243.99	26261.53	12.97	.158	<.001***
Word Valence: Cognitive Load: Face Valence	6.45, 445.11	27084.21	2.78	.039	.010**

Table 2: Results of the three-way ANOVA on spectrum power analysis. Note: df = degrees of freedom; MSE = Mean Square Error; F = F-statistic; pes = Partial Eta Squared.

ilar pattern to that of word valence, with negative words showing the greatest impact, followed by positive words, and then neutral words. These results indicate that facial expressions influence encoding rather than maintenance. The combination of word valence and face valence seems to evoke an additive effect instead of a facilitation effect, as evidenced by the stronger inhibitory effect produced when negative words are paired with angry faces as cognitive load increases. In accordance with Baddeley's model, it can be inferred that participants process words through the phonological loop and facial expressions through visuo-sketchpad, with both managed within their respective components. (Baddeley, 2000). Although information regarding words and faces can be integrated within the episodic buffer (Baddeley and Hitch, 1974), the allocated durations are insufficient, with only 3000 ms designated for each memory set and 2000 ms for each facial expression, which hampers the effective integration of valence information. Additionally, the lexical and facial recognition tasks do not necessitate participants to integrate the valence from words and faces, which positions them as valence-irrelevant tasks. Therefore, they must allocate a large portion of their limited cognitive resources to complete the tasks, which consequently reduces the modulation of cognitive processing by valence, ultimately leading to a reduced influence from valence. These factors may elucidate why the absence of a congruency effect was observed in the present paradigm. In other words, facial expressions are likely to be considered distractors when they are presented during maintenance. This surmise can be substantiated by the previous findings from Dolcos and McCarthy (2006), which demonstrate that emotional distractors impair working memory performance, aligning with the current result showing that the fastest reaction times for neutral words occur under the influence of varying face valences. There are two further potential explanations account for this phenomenon: first, the valences of words and facial expressions are not entirely congruent, as positivity does not always correspond to happiness and negativity does not entirely equate to anger, which may lead to incongruent combinations; second, previous studies that detected the congruence effect consistently employed

a paradigm that presented words and faces simultaneously (Egner et al., 2008; Etkin et al., 2006; Fan et al., 2018, 2016), while this study chooses to present them sequentially.

5 Conclusion

This study examines the impact of the interaction among the valence of words, cognitive load, and valence of faces on working memory, emphasizing the significant effect of the combination of negative stimuli with high cognitive load. Negative stimuli can elicit a stronger inhibitory effect, and when multiple negative stimuli are presented in a trial, this effect persists, leading to an exacerbated impact on performance. In any case, negative words exert a profound dominance, which requires a substantial allocation of limited resources to regulate emotions, thereby adversely affecting working memory performance. Positive words manifest their superiority exclusively under conditions of low cognitive load and in the absence of competing emotional stimuli. Once cognitive demands increase or emotional faces are introduced, this advantage diminishes rapidly. The congruency effect between word and face valence fails to be demonstrated by this study, which may be attributed to the characteristics of the stimuli and the experimental paradigm employed. Future studies can apply EEG or fNIRS techniques to explore the neural activation patterns elicited by different combinations of emotional stimuli and cognitive loads.

Acknowledgments

This research was supported by an internal grant (Project No. P0048115) from The Hong Kong Polytechnic University awarded to Tao Ran.

References

James S Adelman and Zachary Estes. 2013. Emotion and memory: A recognition advantage for positive and negative words independent of arousal. *Cognition*, 129(3):530–535.

Daniel Algom, Eran Chajut, and Shlomo Lev. 2004. A rational look at the emotional stroop phenomenon: a generic slowdown, not a stroop effect. *Journal of experimental psychology: General*, 133(3):323.

Alan Anticevic, Deanna M Barch, and Grega Repovs. 2010. Resisting emotional interference: brain regions facilitating working memory performance during negative distraction. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2):159–173.

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423.

Alan Baddeley. 2020. Working memory. *Memory*, pages 71–111.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. In Gordon H. Bower, editor, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press, New York.

Moshe Shay Ben-Haim, Paul Williams, Zachary Howard, Yaniv Mama, Ami Eidels, and Daniel Algom. 2016. The emotional stroop task: assessing cognitive performance under exposure to emotional content. *Journal of visualized experiments: JoVE*, (112):53720.

Yi-Hsuan Chang, He-Jun Chen, Cesar Barquero, Hsu Jung Tsai, Wei-Kuang Liang, Chun-Hsien Hsu, Neil G Muggleton, and Chin-An Wang. 2024. Linking tonic and phasic pupil responses to p300 amplitude in an emotional face-word stroop task. *Psychophysiology*, 61(4):e14479.

Nelson Cowan. 1999. An embedded-processes model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, 20(506):1013–1019.

Ap Dijksterhuis and Henk Aarts. 2003. On wildebeests and humans: The preferential detection of negative stimuli. *Psychological Science*, 14(1):14–18.

Florin Dolcos and Gregory McCarthy. 2006. Brain systems mediating cognitive interference by emotional distraction. *Journal of Neuroscience*, 26(7):2072–2079.

Tobias Egner, Amit Etkin, Seth Gale, and Joy Hirsch. 2008. Dissociable neural systems resolve conflict from emotional versus nonemotional distracters. *Cerebral Cortex*, 18(6):1475–1484.

Paul Ekman. 1992. An argument for basic emotions. *Cognition Emotion*, 6(3-4):169–200.

Amit Etkin, Tobias Egner, Daniel M Peraza, Eric R Kandel, and Joy Hirsch. 2006. Resolving emotional conflict: A role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, 51(6):871–882.

Beth Fairfield, Nicola Mammarella, Alberto Di Domenico, and Rocco Palumbo. 2015. Running with emotion: When affective content hampers working memory performance. *International Journal of Psychology*, 50(2):161–164.

Lin Fan, Qiang Xu, Xiaoxi Wang, Fei Xu, Yaping Yang, and Zhi Lu. 2018. The automatic activation of emotion words measured using the emotional face-word stroop task in late chinese–english bilinguals. *Cognition and Emotion*, 32(2):315–324.

Lin Fan, Qiang Xu, Xiaoxi Wang, Feng Zhang, Yaping Yang, and Xiaoping Liu. 2016. Neural correlates of task-irrelevant first and second language emotion words—evidence from the emotional face–word stroop task. *Frontiers in Psychology*, 7:1672.

Pilar Ferré, Juan Haro, and José Antonio Hinojosa. 2018. Be aware of the rifle but do not forget the stench: Differential effects of fear and disgust on lexical processing and memory. *Cognition and Emotion*, 32(4):796–811.

Elaine Fox, Riccardo Russo, Robert Bowles, and Kevin Dutton. 2001. Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, 130(4):681.

Barbara L Fredrickson. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3):218.

Nico H Frijda, Suprapti Markam, Kaori Sato, and Reinout Wiers. 1995. Emotions and emotion words. In *Everyday Conceptions of Emotion: An Introduction to the Psychology, Anthropology and Linguistics of Emotion*, pages 121–143. Springer.

Katie E Garrison and Brandon J Schmeichel. 2019. Effects of emotional content on working memory capacity. *Cognition and Emotion*, 33(2):370–377.

Xu Gong, Yu-Xia Huang, Yan Wang, and Yue-jia Luo. 2011. Revision of the chinese facial affective picture system. *Chinese Mental Health Journal*.

Manfred F Greiffenstein, W John Baker, and Thomas Gola. 1994. Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, 6(3):218.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.

Christine H Hansen and Ranald D Hansen. 1988. Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology*, 54(6):917.

Cornelia Herbert and Stefan Sütterlin. 2011. Response inhibition and memory retrieval of emotional target words: Evidence from an emotional stop-signal task. *Journal of Behavioral and Brain Science*, 1(3):153–159.

Vladimir Huerta-Chavez and Julieta Ramos-Loyo. 2024. Emotional congruency between faces and words benefits emotional judgments in women: An event-related potential study. *Neuroscience Letters*, 822:137644.

Tiffany A Ito, John T Cacioppo, and Peter J Lang. 1998. Eliciting affect using the international affective picture system: Trajectories through evaluative space. *Personality and Social Psychology Bulletin*, 24(8):855–879.

Margaret C Jackson, David EJ Linden, and Jane E Raymond. 2012. "distracters" do not always distract: Visual working memory for angry faces is enhanced by incidental emotional words. *Frontiers in Psychology*, 3:437.

Yi-Xiang Jin, Xue-Bing Li, and Yue-Jia Luo. 2013. Effects of emotional content on working memory: Behavioral and electrophysiological evidence. In *International Conference on Brain Inspired Cognitive Systems*.

John Jonides and Edward E. Smith. 2013. The architecture of working memory. In *Cognitive Neuroscience*, pages 243–276. Psychology Press.

Todd A Kahan and Charles D Hely. 2008. The role of valence and frequency in the emotional stroop task. *Psychonomic bulletin & review*, 15(5):956–960.

Johanna Kissler and Cornelia Herbert. 2013. Emotion, etmnooi, or emitoon? – faster lexical access to emotional than to neutral words during reading. *Biological Psychology*, 92(3):464–479.

Juliane Kopf, Thomas Dresler, Philipp Reicherts, Martin J Herrmann, and Andreas Reif. 2013. The effect of emotional content on brain activation and the late positive potential in a word n-back task. *PLOS ONE*, 8(9):e75598.

Stavroula-Thaleia Kousta, David P Vinson, and Gabriella Vigliocco. 2009. Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, 112(3):473–481.

Victor Kuperman, Zachary Estes, Marc Brysbaert, Warriner, and Amy Beth. 2014. Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, 143(3):1065.

Theodor Landis. 2006. Emotional words: What's so different from just words? *Cortex*, 42(6):823–830.

Hyejin J. Lee and Yang Seok Cho. 2019. Memory facilitation for emotional faces: Visual working memory trade-offs resulting from attentional preference for emotional facial expressions. *Memory Cognition*, 47(6):1231–1243.

Elizabeth J Maratos, Kevin Allan, and Michael D Rugg. 2000. Recognition memory for emotionally negative and neutral words: An erp study. *Neuropsychologia*, 38(11):1452–1465.

Maha Nasrallah, David Carmel, and Nilli Lavie. 2009. Murder, she wrote: Enhanced sensitivity to negative word valence. *Emotion*, 9(5):609–618.

Paula M Niedenthal, Jamin B Halberstadt Setterlund, and Marc B. 1997. Being happy and seeing "happy": Emotional state mediates visual word recognition. *Cognition Emotion*, 11(4):403–432.

Shima Ovaysikia, Khalid A Tahir, Jason L Chan, and Joseph FX DeSouza. 2011. Word wins over face: Emotional stroop effect activates the frontal cortical network. *Frontiers in Human Neuroscience*, 4:234.

Luiz Pessoa. 2009. How do emotion and motivation direct executive control? *Trends in cognitive sciences*, 13(4):160–166.

Katarzyna Rączy and Jarosław Orzechowski. 2021. When working memory is in a mood: Combined effects of induced affect and processing of emotional words. *Current Psychology*, 40(6):2843–2852.

Steven C Schwering and Maryellen C MacDonald. 2020. Verbal working memory as emergent from language comprehension and production. *Frontiers in human neuroscience*, 14:68.

N Kyle Smith, John T Cacioppo, Jeff T Larsen, and Tanya L Chartrand. 2003. May i have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia*, 41(2):171–183.

Deborah Talmi and Morris Moscovitch. 2004. Can semantic relatedness explain the enhancement of memory for emotional words? *Memory & Cognition*, 32(5):742–751.

Lotte F Van Dillen and Belle Derks. 2012. Working memory load reduces facilitated processing of threatening faces: An erp study. *Emotion*, 12(6):1340.

Gerben A Van Kleef and Stéphane Côté. 2022. The social effects of emotions. *Annual Review of Psychology*, 73:629–658.

Rachael N Volokhov and Heath A Demaree. 2010. Spontaneous emotion regulation to positive and negative stimuli. *Brain and Cognition*, 73(1):1–6.

YN Wang, LM Zhou, and YJ Luo. 2008. The pilot establishment and evaluation of chinese affective words system. *Chinese Mental Health Journal*, 22(8):608–612.

Anne Weigand, Aline Richtermeier, Melanie Feeser, Jia Shen Guo, Benny B Briesemeister, Simone Grimm, and Malek Bajbouj. 2013. State-dependent effects of prefrontal repetitive transcranial magnetic stimulation on emotional working memory. *Brain Stimulation*, 6(6):905–912.

Qianru Xu, Chaoxiong Ye, Simeng Gu, Zhonghua Hu, Yi Lei, Xueyan Li, Lihui Huang, and Qiang Liu. 2021. Negative and positive bias for emotional faces: Evidence from the attention and working memory paradigms. *Neural Plasticity*, 2021(1):8851066.

Xiang-ru Zhu, Hui-jun Zhang, Ting-ting Wu, Wen-bo Luo, and Yue-jia Luo. 2010. Emotional conflict occurs at an early stage: Evidence from the emotional face–word stroop task. *Neuroscience Letters*, 478(1):1–4.

Toward True Neutrality: Evaluating Inference-Time Debiasing Strategies for Gender Coreference Resolution in LLMs

Arati Mohapatra
Indian Institute of Science
aratim@iisc.ac.in

S Jaya Nirmala
National Institute of
Technology Tiruchirappalli
sjaya@nitt.edu

Abstract

Large Language Models (LLMs) are increasingly integrated into high-stakes domains such as healthcare, education, and finance, influencing daily decision-making. However, LLMs have been shown to exhibit gender bias in their generated responses, particularly against women and non-binary individuals. While recent research has proposed inference-time debiasing techniques like self correction and self-consistency in question-answering, their effectiveness across diverse tasks and their computational efficiency remain underexplored. Particularly, the *gender neutrality of LLMs in gender coreference resolution tasks* remains an open question. In this work, we present a comprehensive evaluation of inference-time gender bias mitigation strategies on gender coreference resolution pertaining to occupational words. We assess both the bias reduction achieved and the computational costs incurred to identify strategies that best balance fairness and efficiency. We find that self correction with a low-bias feedback generator achieves up to 41% better performance than existing self-consistent prompting, yet with comparable sampling rates. Moreover, we also qualitatively analyze the Chain-of-Thought reasoning process of the LLMs during gender prediction and highlight certain LLM-specific response patterns related to bias, logic and grammaticality that arise frequently during gender coreference resolution. The scripts and dataset used in this study are available at https://github.com/true-neutral-nlp/Inference-Time-Gender-Coreference-Resolution.

1 Introduction

Large Language Models (LLMs) are trained using a vast collection of data, leading them to obtain strong natural language inference, reasoning and generation capabilities (Raiaan et al., 2024). During this training process, however, LLMs also unintentionally learn underlying societal biases, including gender, racial and geographical biases, from the data (Raj et al., 2024). Gender biases are especially prevalent when prompting LLMs about occupations traditionally carried out by the different genders, for example, a doctor is most often associated with the male gender, whereas a nurse is often associated with the female gender (Kaneko et al., 2024). Moreover, non-binary genders are rarely mentioned in the discourse (Dev et al., 2021). As LLMs are rapidly being integrated into various high-stakes automated decision-making pipelines such as resume screening, care must be taken to make sure that the involvement of LLMs does not affect or enforce stereotypes and biases against a certain gender.

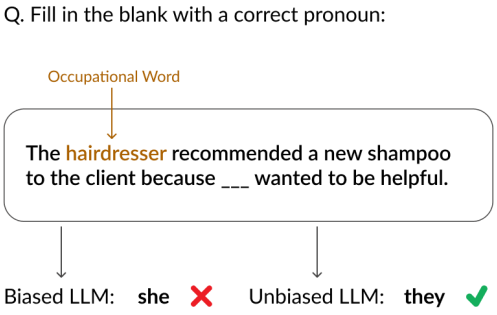


Figure 1: An example of a Gender Coreference Resolution task.

Gender biases are implicit in pre-trained LLMs due to learning traditional associations between gender and occupational words during training (Zhou et al., 2024). These biases may manifest indirectly and shape the way that such language models respond to user queries. We seek to make these implicitly learned biases apparent through gender coreference resolution in order to evaluate them. Coreference resolution refers to the task of correctly identifying mentions or phrases in text that refer to the same entity (Zhao et al., 2018).

Gender coreference resolution involves resolving mentions that give away the gender identity of the entity it refers to, such as pronouns. In this work, we prompt LLMs to resolve a blanked-out mention in a given sentence to an occupational word by responding with a pronoun. This process is shown in Figure 1. This task requires the LLM to make inferences and assumptions about the occupational word and translate it into a pronoun, most of which are gendered in the English language, thus revealing inherent biases. This allows us to explicitly evaluate the ability of LLMs to remain neutral and not associate occupations with certain genders, i.e., their gender neutrality. Since occupations can be carried out by anyone regardless of gender, it is important to make sure that disruptive technologies such as LLMs display gender neutrality.

Previous work has applied inference-time methods such as Chain-of-Thought (CoT) prompting and Self Correction to both investigate the presence of gender bias of LLMs exposed to occupational words and then for the corresponding bias mitigation (Anantaprayoon et al., 2025). However, these studies have concentrated on mitigating gender bias solely in question-answering with context, where explicitly probing the gender neutrality of LLMs has not been carried out. We adopt these inference-time methods that have been shown to be helpful in mitigating bias in these situations in our work. Moreover, we recognize the need to conduct an overall investigation into the robustness and computational efficiency of these methods to increase the trust in inference-time bias mitigation and their adoption to real-world systems. Specifically, we are interested in examining how LLMs explain their stance while resolving pronouns to occupational words to understand how sound and bias-free their reasoning is. We thus seek to answer the question: *Which inference-time bias mitigation strategy is the most accurate, robust and computationally effective in predicting gender neutrality in the task of gender coreference resolution?*

Our contributions are threefold:

1. We evaluate how accurate inference-time gender debiasing methods for LLMs involving prompting and self correction are at maintaining gender neutrality during gender coreference resolution.

2. We discuss the robustness of inference-time gender debiasing methods by qualitatively highlighting common logical and grammatical reasoning patterns of LLMs during gender coreference resolution.

3. We compare the computational efficiency of inference-time gender debiasing methods by comparing the number of times the LLM is sampled during the execution of each debiasing strategy.

The rest of this paper is organized as follows: Section 2 provides an overview of related work that acts as the background for our research. We then elaborate on the experiments we conduct in Section 3, and Section 4 specifies the metrics we use to analyze the results obtained. Then, in Section 5, we discuss our results and in Section 6, we outline limitations and constraints. Finally, Section 7 concludes.

2 Related Work

Prompt-based Debiasing Inference-time debiasing methods improve upon zero-shot prompting, which involves directly prompting the LLM with a question whose response is then evaluated for gender bias (Mohapatra et al., 2024). Debiasing prompts, which involve adding *"Please ensure that your answer is unbiased and does not rely on stereotypes"* have been proposed to encourage the LLM to reflect on potential biases and stereotypes it may exhibit in its response (Ganguli et al., 2023). Moreover, chain-of-thought (CoT) reasoning, achieved by prompting the LLM *"Let's think step by step."* has also been shown to mitigate biases in unscalable tasks including symbolic and arithmetic reasoning (Kaneko et al., 2024). Similar approaches including informative prompts, providing more context, and CoT have been applied to mitigate biases for gender coreference resolution in machine translation (Sant et al., 2024). However, these approaches are sensitive to prompt design and should be tested across diverse tasks and various prompt formulations (Hida et al., 2024). To increase the robustness of prompting methods, self-consistency, which involves repetitive individual sampling and majority voting was introduced (Wang et al., 2023).

Adaptive Consistency Though self-consistency has shown significant improvements in the robustness of prompting methods, it may lead to repetitive sampling of the LLM even in the case of an early majority due to a fixed number of samples.

Adaptive consistency makes the number of samples dynamic by adding a lightweight stopping criterion that evaluates how likely it is for the current major element to remain the major element in the following samples by modeling the distribution of answers as a Dirichlet distribution, which can then be approximated to a Beta distribution for the top two major elements (Aggarwal et al., 2023). This technique has not been applied to gender bias mitigation yet, but since it can be considered a more efficient version of self-consistency, we primarily include it in our evaluation to compare the computational efficiency.

Self Correction Direct prompting approaches such as CoT and even self-consistency do not enable LLMs to reflect on previous answers as all samples to the LLM are independent. Self correction, which includes an iterative feedback loop, has been applied and shown to perform better for gender bias mitigation owing to good feedback between samples. Previous work has shown that multi-LLM interactions amplify bias and attempt to mitigate this using self-reflection with fine-tuning (Borah and Mihalcea, 2024). Self reflection for reducing gender bias in task assignments was made more reliable by assigning referee and participant roles to LLM instances (Cheng et al., 2024). The existing self correction framework has been extended to mitigate societal biases in question answering, and it has been demonstrated that clarifying intentions at each step, from prompting to response and feedback is necessary for better bias mitigation (Anantaprayoon et al., 2025).

3 Methodology

3.1 Dataset

We use the Winogender dataset to provide sentences with occupational words and a pronoun for gender coreference resolution (Zhao et al., 2018). This dataset, set in the style of Winograd schemas, contains templates, each containing a primary occupational word, a secondary participant occupational word, and an ambiguous pronoun that may refer to either of the occupational words depending on the surrounding sentence context. An example is *"The technician told the customer that he could pay with cash."*, where *technician* and *customer* are the two occupational words, and *he* is the pronoun in this case. The dataset includes 720 such hand-written sentences, corresponding to 2 templates each for 2 participants and 3 genders across 60 one-word occupations sampled from the U.S. Bureau of Labor Statistics. We remove all pronoun references from the sentences to create 120 unique templates with blanked-out pronouns and input these to the LLM to resolve the blank space to the correct occupational word. This ensures that the LLM has to infer which occupation the blanked-out pronoun refers to and then make an informed decision based on the pronoun knowledge of the model, which reveals its inherent gender bias.

3.2 Gender Coreference Resolution

The task of resolving a given mention in a sentence to an entity where the mention may be indicative of gender is known as gender coreference resolution. We use templates from the Winogender dataset with two occupational words and a blanked-out space that we require the LLM to resolve to a pronoun depending on the context. Thus, this task requires the LLM to make inferences from a very limited context, as opposed to usual question-answering formats that include an additional context on top of the query. We try to limit the context in our task as much as possible to bring out the implicit biases in the LLM rather than its ability to reason from the context. Since the context of all templates is limited and ambiguous and provides no direct hints to the gender of either of the occupational words, the gender neutrality of the LLM is shown through its predictions for the blanked-out space. An occupation in itself has no gender associated with it, and thus we look for variations of the third person gender-neutral singular pronoun *"they"* (such as *"they"*, *"them"*, *"their"* and *"theirs"*) in the LLM's response. We posit that these particular pronoun variations would be the most appropriate when there are no hints towards gender, as opposed to male-biased pronouns (such as *"he"*, *"his"*, and *"him*) and female-biased pronouns (such as *"she"*, *"her"* and *"hers"*). We evaluate the performance of two LLMs, Llama3 and Mistral on this task. Specifically, we run our experiments using the Llama3 8B model (Dubey et al., 2024) and the Mistral 7B model (Jiang et al., 2023). We chose to evaluate these two models specifically given their prominence, relevance, and performance as open-source LLMs.

3.3 Gender Bias Mitigation Strategies

Influenced by previous work, we broadly apply two kinds of inference-time debiasing strategies for our gender neutrality evaluation: Direct Prompting

and Self Correction. We implement four different kinds of direct prompting approaches– zero-shot, chain-of-thought (CoT), self-consistent CoT, and adaptive consistent CoT– that have been shown to yield promising results for gender debiasing, and both same-model and cross-model self correction (refer to Section 2 for a more detailed discussion on the introduction and general motivation of these strategies). The debiasing strategies we explore are summarized in Figure 2.

Zero-shot Prompting We directly prompt the LLM with a Winogender sentence template containing two occupational words and a blanked-out pronoun that may refer to either of the two occupational words. We ask the LLM to fill in the blank with a correct pronoun and clarify the pronouns it can include in its answer by providing a list of male, female, and neutral pronouns along with the prompt. We also add details about the answering format for easy extraction of the pronoun from the LLM's responses. This is the baseline prompt which gets further augmented in other direct prompting approaches.

Chain-of-Thought Prompting To encourage the LLM to follow a logical reasoning process before responding with a final pronoun, we use Chain-of-Thought (CoT) prompting. Allowing the model to elaborate on its reasons for choosing a certain pronoun not only helps the LLM prevent inherently learned gender biases from directly influencing the answer by forcing a thought-out reasoning process before answering that reveals underlying biases, but also allows us to analyze the responses for any underlying bias patterns. We achieve CoT prompting by adding *"Let's think step by step."* to the zero-shot prompt (Wei et al., 2022).

Self-Consistent Chain-of-Thought Prompting Even though CoT prompting allows models to elaborate on their thought process before responding with a final pronoun, the provided reasoning itself may not be as sound during one single run. To allow LLMs multiple chances at reasoning for a single query to increase reliability and confidence in this strategy, and recognizing that there may be multiple correct reasoning paths to the same answer, we adopt self-consistent CoT prompting as another debiasing strategy based on direct prompting. We sample the LLM independently 10 times for each template sentence from the Winogender dataset with the CoT prompt and use majority voting to decide on the final prediction after all samples.

Adaptive Consistent Chain-of-Thought Prompting To dynamically adjust and reduce the number of fixed samples in self-consistent CoT prompting, we adopt a lightweight stopping criterion that estimates the probability of the current major element remaining the majority in the following samples. This is done by modeling the distribution of unique responses as a Dirichlet distribution, which can then be approximated to a Beta distribution for the top two major elements (Aggarwal et al., 2023). We use a confidence threshold of 0.95 for determining whether to halt sampling for the current query or to continue sampling. If no clear majority is established, this strategy defaults to the self-consistent CoT case with a maximum of 10 samples made to the LLM per query.

Self Correction We also evaluate the performance of iterative self-reflection based on feedback by implementing self correction. The general self correction framework we adopt involves two LLM instances, one acting as a responder, that responds to the given Winogender sentence template with a pronoun and its corresponding reasoning, and the other acting as a feedback generator. This feedback generator is prompted to generate feedback based on three criteria: coherence (the soundness of reasoning), comprehensiveness (whether the response uses all information available to make a decision) and objectivity (whether the LLM remains unbiased and does not reinforce stereotypes) (Anantaprayoon et al., 2025). The feedback generator gives the responder's response a binary rating of either 0 or 1 based on the three evaluation aspects and a total rating out of 3. The responder is iteratively provided the feedback given by the feedback generator as well as the total rating to reflect on its answer and provide improved reasoning in the next run. We implement two kinds of self correction: same-model correction, where two separate instances of the same LLM act as responder and feedback generator respectively, and cross-model correction, where the responder and feedback generator are instances of different LLMs. We allow an iterative response-feedback loop to run for a maximum of 10 times, similar to self-consistency, or stop if the model scores 3/3 on all necessary aspects.

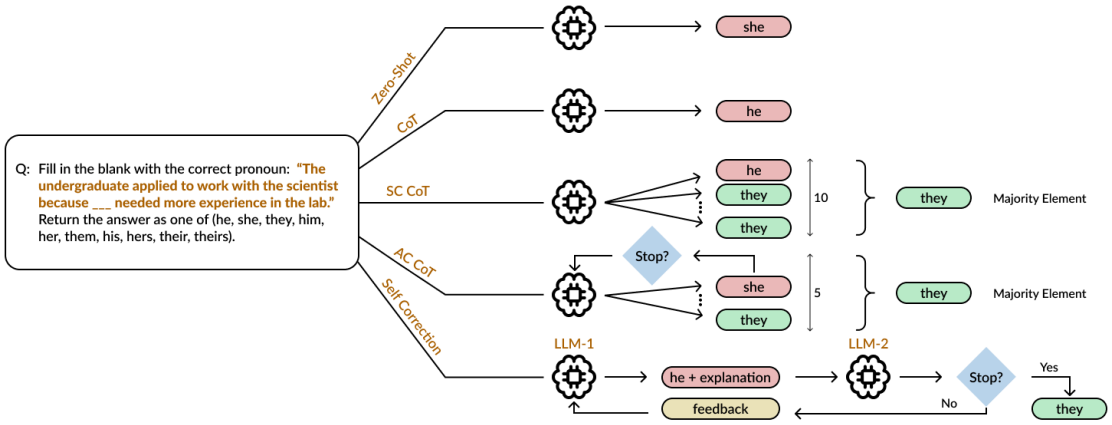


Figure 2: Overview of inference-time debiasing strategies used for gender-neutrality prediction. In Zero-shot prompting, the model directly predicts a pronoun given an occupational sentence; Chain-of-Thought (CoT) prompting encourages step-by-step reasoning; Self-Consistent CoT (SC CoT) uses multiple independent CoT samples with majority voting; Adaptive Consistent CoT (AC CoT) dynamically adjusts the number of samples. Self Correction involves iterative refinement of responses based on feedback from either the same model or a different model.

4 Metrics

In this section, we expand on the metrics we use to quantitatively evaluate the inference-time gender debiasing strategies we employ in our experiments. We calculate the accuracy, direction of gender bias as well as the computational efficiency of each debiasing strategy.

4.1 Accuracy

Since our aim in this work is to evaluate how well a given debiasing strategy predicts the gender neutrality of the occupational word in a given sentence, we measure the correctness of each strategy in terms of accuracy, which is the fraction of gender neutral predictions made. We consider *"they"*, *"them"*, *"their"* and *"theirs"* to be acceptable gender-neutral pronouns, and also consider the cases where the LLM refuses to fill in the blank with a gendered pronoun stating explicitly that it must remain gender-neutral, even though it does not predict one of the gender-neutral pronouns. We decide to include these cases also citing the reasoning of the model to be sound. We calculate accuracy as described below in Equation 1.

$$Accuracy = \frac{n_{gender\_neutral}}{n_{total}} \quad (1)$$

Where $n_{gender\_neutral}$ is the number of gender neutral predictions and n_{total} is the total number of predictions, which includes gender neutral, male, female and unknown pronoun predictions. The LLM may predict gendered pronouns in the male or female direction, explicitly state gender neutrality or use a pronoun such as *they*, or give answers that are not pronouns (words like *someone, one, who, the*), or restate the occupational word to fill in the blank. Anything except a pronoun from the given list is categorized as unknown. This list includes gender-neutral, male and female pronouns such as *"they"*, *"them"*, *"their"*, *"theirs"*, *"he"*, *"him"*, *"his"*, *"she"*, *"her"*, and *"hers"*.

4.2 Direction of Bias

To ensure we conduct a robust study and evaluate the pronoun predictions from all angles, we are also interested in all the times that the model does not predict gender neutrality, and if it has a tendency of defaulting to predicting pronouns of a certain gender. We measure this in terms of the direction of bias in either the male or female direction. This is a ratio of the gendered prediction in question (either male or female) to the total number of gendered predictions. This helps us see if the gendered predictions are balanced or skewed in a certain direction.

$$Bias_{male} = \frac{n_{male}}{n_{male} + n_{female}} \quad (2)$$

$$Bias_{female} = \frac{n_{female}}{n_{male} + n_{female}} \quad (3)$$

The calculation of bias in both the male and female directions are described in Equation 2 and Equation 3. n_{male} refers to the number of male pronoun predictions made by the LLM during a particular debiasing strategy and n_{female} refers to the number of times a female pronoun was predicted.

4.3 Computational Efficiency

In this work, along with presenting how accurate LLMs are in mitigating gender bias and predicting gender neutrality, we are also interested in understanding the computational effort it takes to be more bias free. We are especially interested in seeing if any strategies are computationally efficient as well as accurate in gender neutrality prediction. Since we utilize only inference-time methods, where the most effort is the inference on the part of the LLM, we calculate computational efficiency in terms of the number of samples made to the LLM on average. Every time we prompt the LLM we also maintain count of the number of samples made and average this number for each debiasing strategy for comparison.

5 Results and Discussion

Table 1 reports the accuracy of gender neutrality predictions across all evaluated inference-time debiasing strategies, along with the bias in the male and female direction.

Accuracy All prompting and self correction methods are seen to usually increase the accuracy of gender neutrality prediction compared to zero-shot prompting. Self correction generally performs better than prompting methods in terms of accuracy of gender neutrality prediction across both LLMs. Self-consistency and Adaptive Consistency lead to opposite results in Llama3 and Mistral, thus showing that they are not robust methods and need to be tested more extensively. Chain-of-thought (CoT) also shows differing behavior between LLMs where it shows significant improvement in Llama3, but a decrease in accuracy in Mistral. This observation is in accordance to previous work which has shown CoT to not be a robust method of gender bias mitigation (Hida et al., 2024). Improvement in accuracy is the most when a high-bias model has cross-model self correction with a low-bias model. Since Mistral comparatively performs better on gender neutrality prediction, we call it a low bias model in this work. However, when Llama3 acts as the feedback generator for Mistral's responses, the performance actually reduces, which suggests the necessity to choose the feedback generator wisely to achieve the maximum benefits. This may suggest that self correction depends on the quality of the feedback, as a consistent pattern is observed: a low-bias model's feedback leads to better performance in terms of accuracy. This is in line with previous work (Anantaprayoon et al., 2025).

Direction of Bias The male and female bias scores become more balanced during self correction when compared to prompting methods. This may be due to the fact that the iterative debate mechanism prevents from defaulting to one gender as a response, which might occur in the prompting strategies, since each sample in these strategies is independent and gets no feedback. Between the two self correction methods, there is not much difference, thus suggesting that it is the existence of feedback and self-reflection that influences a more balanced behavior rather than the quality of feedback itself. But same-model correction is slightly better balanced than cross-model correction. The direction of bias is highly LLM dependent, but consistent across the same LLM. Llama3 is more female biased, whereas Mistral is more male biased, which means that Llama3 seems to predict female pronouns more than male pronouns, and vice versa for Mistral. Self-consistency and Adaptive Consistency do not vary by a small margin, and with opposing behavior across LLMs, thus again proving its lack of robustness, and the necessity for more fine-grained evaluation on diverse tasks and prompts.

Computational Efficiency Number of samples used is 1 for the zero-shot and CoT cases as they sample the LLM only once, but these are also the cases associated with the least accuracy and least balanced directional bias. For self-consistent CoT, since we use 10 as the number of samples, it remains this fixed number. For adaptive consistency, we find that the reduction is not very significant (one or two samples less on average) compared to self-consistency, which may be because of the threshold we defined (0.95). For a lower threshold, confidence may be achieved sooner and the number of samples may also decrease. For self correction, though it involves an iterative loop, we find that the number of samples made to both LLMs involved together on average remains around 10,

LLM	Debiasing Strategy	Accuracy	Accuracy Gain	Male/Female Bias
Llama3	Zero-Shot	0.07	0.00	0.36 / 0.64
	Chain-of-Thought (CoT)	0.23	0.16	0.30 / 0.70
	Self-Consistent CoT	0.16	0.09	**0.45 / 0.55**
	Adaptive Consistent CoT	0.23	0.16	0.65 / 0.35
	Self Correction (Same-Model)	0.43	0.36	**0.45 / 0.55**
	Self Correction (Cross-Model with Mistral)	**0.57**	**0.50**	0.40 / 0.60
Mistral	Zero-Shot	0.46	0.00	0.67 / 0.33
	Chain-of-Thought (CoT)	0.43	-0.03	0.68 / 0.32
	Self-Consistent CoT	0.63	0.17	0.78 / 0.22
	Adaptive Consistent CoT	0.56	0.10	0.42 / 0.58
	Self Correction (Same-Model)	**0.71**	**0.25**	**0.50 / 0.50**
	Self Correction (Cross-Model with Llama3)	0.54	0.08	0.54 / 0.46

Table 1: Comparison of debiasing strategies across Llama3 and Mistral models. Accuracy and male/female bias are reported per method. Accuracy Gain refers to the difference in accuracy between zero-shot settings and each debiasing method. The highest accuracy, accuracy gain and most balanced bias are highlighted in bold.

which is the same as the number of samples in self-consistency. We also observe from our experiments that the iterations in self correction barely cross 5 and reach a full score in terms of coherence, comprehensiveness and objectivity without exhausting the maximum number of iterations. We can thus conclude that self correction does not incur more computational costs than self-consistency, but has significant gains in accuracy and robustness when it comes to gender coreference resolution.

Gender Associations to Occupational Words Mistral is seen to be less gender biased, as it predicts more occupations to be neutral at least once, i.e., most of the occupational words are not predicted as only male or female for the majority of the times, which is the behavior Llama3 exhibits in the direct prompting techniques. We can thus infer that Mistral is comparatively a low-bias model when compared to Llama3. We also observe that the existing associations of certain occupational words to gender (such as *engineer* and *technician*) are reduced when applying self correction. This does not necessarily mean that self correction predicts perfect neutrality, but for all times that word is encountered, the majority is not male or female. This shows that self correction encourages models to reflect and break learned associations of gender and occupational words. Still, certain words are predicted in a biased manner. Words such as *dietitian*, *hairdresser*, *hygienist*, and *secretary* are resolved to female pronouns most of the time, while words such as *carpenter*, *electrician*, *firefighter*, *homeowner*, *janitor*, and *officer* are majorly resolved to male pronouns, which is indicative of underlying gender-biased associations.

Unknown Pronoun Prediction Tendencies Mistral has a higher rate of predicting unknown pronouns, which are those defined to not be part of the list defined in Section 4. Mistral sometimes fails to resolve the blanked-out space in the given Winogender sentence template to a pronoun, but gives an alternate grammatically and contextually correct word instead. The most commonly observed words are *"the"*, *"it"* and *"the [occupation/participant]"*. On the other hand, Llama3 clearly mentions grammatical reasoning in the CoT responses and is shown to eliminate certain options based on grammar alone. This shows its ability to not only reason based on the given context, but also ensure grammatical correctness, thus increasing trust in its outputs. However, it is seen to predict *"he/she"* rather than a gender-neutral pronoun in multiple cases, showing its implicit bias toward binary genders only. We can thus infer that it is comparatively a high bias model, yet with sound reasoning and grammaticality.

Responder and Feedback Generator Behavior in Self Correction In a self correction framework, Mistral as a responder sometimes ends up reasoning about the sentence itself rather than rea-

soning about the potential pronouns, and hence ends up rewording the given sentence as its final answer, thus showing low task comprehension, yet low bias in its answers. Predicting and reasoning toward a gender-neutral pronoun is done mostly due to inclusivity and quoting modern writing conventions rather than from a grammatical perspective. Llama3 is seen to consider not only the context of the sentence, but also the grammatical structure to arrive at its answer. Its tendency to default to binary pronouns remains, but on closer examination of its reasoning process, we see that *"they"* or other gender-neutral pronouns are often considered, yet dismissed as they either do not fit in the grammatical correctness of the sentence or are considered to be plural by Llama3. Sometimes, it gives a prediction based on language patterns or assumptions, which is seen to exhibit gender bias, but it also provides a sentence acknowledging that there are no gender cues and hence it might be wrong, thus suggesting maturity in reasoning. Moreover, when a gender-neutral prediction is made, there is a clear reasoning path following a dual-pronged approach of logic and grammar. However, when the prediction is gendered, there is not much mention of grammar and the logic is not so strong. This supports the claim that self-reflection and encouraging reasoning help reduce bias in LLMs. Mistral's feedback often includes gender neutrality concerns, unlike Llama's feedback, that advocates for structure, and grammatical correctness. Mistral, when providing feedback, references the Winogender sentence template, but changes aspects of it, such as changing singular words to their plural versions, or rewords the sentence itself, which influences further iterations of self correction to stray from the original sentence formulation, thus leading to untrustworthy results. These behavioral differences highlight the need to not only understand the bias levels of different LLMs, but also to understand the soundness of reasoning and apply them as feedback generators accordingly.

6 Limitations

Despite our efforts to investigate the gender neutrality of LLMs, we acknowledge certain shortcomings in our approach. Firstly, we use only two open-source LLMs and not the current state-of-the-art GPT models to perform our evaluation. We were motivated by the lack of literature addressing bias mitigation in open-source models, yet constrained by financial resources to compare their performance to proprietary pay-per-use models. Secondly, we used templates from only the Winogender dataset as input to the LLMs for probing their gender neutrality. Template-based approaches have been shown to be less representative of real-life tasks, and hence natural sentence continuation prompts have recently been introduced (Alnegheimish et al., 2022). In future work, we plan on extending our evaluation to these prompts. Thirdly, we were limited to English as our primary language of evaluation, and we concede that our experiments are very language-dependent as our experiment formulation depends on pronoun prediction, which differs from language to language. Finally, our analysis does not account for differences in model size or the composition of training data, both of which likely contribute to the observed variations in bias, and thus, future work might benefit from examining how these underlying factors shape model behavior.

7 Conclusion

In this work, we demonstrated that self-correction methods, particularly those using low-bias feedback generator models, are accurate, robust, and computationally efficient approaches for gender debiasing. Through directional bias analysis, we found that underlying bias directions depend largely on individual LLMs and can be balanced using self-correction. Furthermore, while these inference-time debiasing strategies show promise in mitigating gender stereotypes through reasoning and reflection, learned associations between gender and certain occupational terms persist, motivating the development of more bottom-up, data-driven debiasing approaches. Finally, our qualitative analysis of LLM reasoning revealed that the emphasis on gender debiasing versus logic and grammaticality varies across models, highlighting the need to understand such tendencies in addition to bias levels before selecting feedback generators for self-correction frameworks.

Ethics Statement

In this work, we seek to understand how well LLMs are able to predict the gender neutrality of a profession. In our evaluation, we acknowledge that treating the singular use of *they* as the only unbiased option may impose a normative linguistic standard; while this aligns with many accessibil-

ity style guides, it is not universally accepted, and thus risks conflating grammaticality with fairness. While our experiments show the ability to mitigate such associations and encourage LLMs to be more inclusive to a certain extent, there remains considerable room for improvement in increasing the neutrality of these models. We do not fine-tune the model and focus solely on inference-time solutions, which may mask but not fully eradicate the biases learned. Masking bias can be dangerous as it may create a false sense of fairness, allowing underlying stereotypes to persist in subtle ways, reduce trust when such biases resurface in different contexts, and hinder efforts to address the root causes of the problem. We seek to highlight this issue to promote future research in this direction toward achieving complete mitigation of such potentially harmful biases and stereotypes.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and 1 others. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2025. Intent-aware self-correction for mitigating social biases in large language models. *arXiv preprint arXiv:2503.06011*.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326.

Ruoxi Cheng, Haoxuan Ma, and Shuirong Cao. 2024. Deceiving to enlighten: Coaxing llms to self-reflection for enhanced bias detection and mitigation. *arXiv e-prints*, pages arXiv–2404.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, and 1 others. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. *arXiv preprint arXiv:2407.03129*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Arati Mohapatra, Kavimalar Subbiah, Reshma Sheik, and S Jaya Nirmala. 2024. Mitigating gender bias in large language models: An evaluation using chain-of-thought prompting. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 861–870.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in mt with llms. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In

The Eleventh International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Hanqing Zhou, Diana Inkpen, and Burak Kantarci. 2024. Evaluating and mitigating gender bias in generative large language models. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, 19(6).

A/E Pelita as an Aspectual Marker in Korean Auxiliary Verb Constructions: An Experimental Comparison with Spanish Optional *se*

Nakyung Yoon
Korea University
joy2001@korea.ac.kr

Abstract

The purpose of this study is two-fold. First, it aims to experimentally examine two approaches to Korean *a/e pelita* auxiliary verb constructions (AVCs)—an aspectual approach and an expressive approach. Second, it seeks to determine whether the *a/e pelita* AVC differs from its unmarked base construction in terms of telicity. Two acceptability judgment tasks (Experiment I) were conducted to assess Korean speakers' acceptability judgments of *a/e pelita* AVCs with eventives and statives. A truth value judgment task (Experiment II) was also conducted to assess the strength of event completion inference in comparison with Spanish optional *se*. Results from Experiment I indicate that, as predicted by the aspectual approach, *a/e pelita* functions as a telic marker with eventive predicates. Results from Experiment II suggest that, unlike Spanish optional *se*, *a/e pelita* yields a stronger inference of event completion and marks exhaustivity more strongly than the base construction, particularly depending on the specific verb.

1 Introduction

In Korean auxiliary verb constructions (AVCs), it has been argued (Sohn, 1973, 2001) that the auxiliary a/e pelita triggers an inference of event completion[1]. Building on this, Choi (2003, 2005) proposed that the auxiliary functions as an aspectual (Asp) head, always yielding a telic interpretation and determining the aspectual properties of the entire predicate.

By contrast, Jung and Kim's (2017) expressive approach challenges Choi's analysis, arguing that a similar auxiliary, *nay*, is compatible only with for-adverbial phrases (typically associated with atelicity), but not with in-adverbial phrases (typically associated with telicity). Further evidence supporting their claim involves the occurrence of a V2 in the AVCs with stative predicates and the morphological distribution of a passive suffix[2]. Notably, Choi's approach predicts that, given that *v* is associated with hosting an external argument, a passivized AVC should position the passive morpheme structurally below the Asp head. This prediction is borne out with *a/e pelita*, as illustrated in (1), contrary to the expectations of Jung and Kim's (2017) analysis regarding nay.

(1) a. Phokpal-lo inhay ku
explosion-INST result from that
tosi-ka
city-NOM

phakoy toye peli-ess-supnita.
destroy become-PASS peli-PAST-DEC

'The city was completely destroyed by the explosion.'

b. Netflix-eykye motwu mek-hi-e
Netflix-DAT all eat-PASS-e

peli-ess-ta.
peli-PST-DEC

'All was ultimately eaten up by Netflix.'

In both (1a) and (1b), the passive morpheme precedes the auxiliary, supporting the aspectual approach rather than the expressive analysis. Furthermore, Jung and Kim's (2017) expressive approach does not account for the inference of event completion, as illustrated in (2).

(2) a. Ku-ka pap-ul mek-e
He-NOM meal-ACC eat-e
peli-ess-ta.
peli-PAST-DEC

'He ate up the meal.'

[1] Auxiliary verb constructions (AVCs) refer to complex verbal structure in which two serial verbs within one clause jointly express a single event, with the second verb (V2) occurring directly after the first verb (V1) (cf. J. Yoon, 2018)..

[2] Hong (2015: 111) identified one such instance, and a subsequent Google search reveals additional cases in which AVCs occur with stative predicates.

b. Yengho-nun apeci swul-ul masi-e
Yengho-NOM father wine-ACC drink-e
peli-ess-ta.
peli-PST-DEC
'Yengho drank up his father's wine.'

While Choi (2003, 2005) maintains that the auxiliary verb functions as an Asp head marking telicity and thus, determining the overall aspectual properties of the predicate, Verkuyl's (1972) compositional perspective holds that aspect is derived from the interaction between the semantic and lexical properties selected by tense morphology, together with modifications introduced by aspectual operators such as adverbials. From this viewpoint, aspect is compositional.

Moreover, Sohn (1973: 241) previously argued that the main verb and the auxiliary *a/e pelita* share a selectional property: they allow only verbs of action (which generally correspond to eventive predicates) to be embedded in their complements (e.g., pap-ul mek-e peli-ess-ta 'I finished eating (the meal)'). This selectional restriction likely stems from the nature of *a/e pelita* as a telic marker, since the inference of event completion arises only with event predicates. Consequently, *a/e pelita* is generally incompatible with stative predicates. According to Sohn's analysis, when *a/e pelita* appears with stative predicates, it is not functioning as an aspectual marker. Rather, such uses reflect a distinct construction, governed by a separate set of semantic or expressive functions.

Under Sohn's (1973, 2001) analysis, which treats *a/e pelita* as a telic marker, its compatibility with for-adverbial phrases—typically licensed only with atelic predicates, as in (3)—poses a challenge. Accounting for this requires an additional assumption. Consider the examples in (4).

(3) a. John drank a beer in 10 minutes /?for 10 minutes.

b. John drank down a beer in 10 minutes/*for 10 minutes.

(4) a. Chelswu-nun ilcwuil mane/dongan
Chelswu-NOM a week in/for
Cencayng-kwa Phyenghwa-lul
War-and Peace-ACC
ilke-ess-ta.
read-PAST-DEC
'Chelswu read War and Peace in a week/for a week.'

b. Chelswu-nun ilcwuil mane/dongan
Chelswu-DAT a week in/for
Cencayng-kwa Phyenghwa-lul
War-and Peace-ACC
ilke-peli-ess-ta.
read-peli-PAST-DEC
'Chelswu read up War and Peace in a week/for a week.'

In (4a) the combination of a consumption verb and a quantized noun as the direct object yields an accomplishment predicate, as confirmed by its compatibility with an *in*-adverbial diagnostic (MacDonald, 2017; Martínez-Vera, 2022). The acceptability of the for-adverbial in (4a), however, can be explained by Park (2011: 347-350), who proposes that a for-adverbial in Korean may allow two possible interpretations. When a for-adverbial co-occurs with an accomplishment predicate, it can yield a reading in which the event is reinterpreted as an atelic activity. In a similar line, Borer (2005) describes for-adverbials as atelicizers[3]. In this view, the for-adverbial shifts the aspectual interpretation of the predicate, resulting in a compatible (though marked) sentence.

Similarly, the acceptability of both in- and for-adverbials in *a/e pelita* AVCs, as shown in (4b), does not provide a clear diagnostic for telicity. Although *a/e pelita* is presumed to function as a telic marker, the presence of a for-adverbial appears to coerce an achievement predicate into an activity reading. This pattern suggests that the telic marker somehow does not block aspectual coercion, in which compositional aspect interacts with contextual or pragmatic factors to override the expected telic interpretation in Korean AVCs.

Building on Verkuyl (1972) and Rhee (2008), I propose that this flexibility stems from the auxiliary *a/e pelita* having the potential to express speaker subjectivity[4]. This expressive function does not preclude co-occurrence with a for-adverbial and may still result in a telic interpretation at the VP level. This assumption is supported by a stative predicate such as *nolla-* 'surprise' or *sulphe-* 'sad'

[3]It should be noted, however, that a for-adverbial may serve as an appropriate diagnostic of telicity, depending on the context (cf. Kim, Ko, and Yang, 2020)

[4]A parallel can be drawn with Spanish optional *se*, which, like the Korean auxiliary *a/e pelita*, has been analyzed as following a grammaticalization path (Armstrong and MacDonald, 2021).

as it can express subjectivity but at the same time result in a telic interpretation when combined with *a/e pelita*[5].

Spanish optional *se*, a reflexive clitic pronoun, and the Korean auxiliary *a/e pelita* have been claimed to exhibit similar grammatical and semantic/pragmatic patterns (Strauss, 2003). Both can optionally occur with certain eventive (e.g., *María se comió el helado.* 'María ate up the ice-cream.' vs. *Yengho-nun apeci swul-ul masy-e peli-ess-ta.* 'Yengho drank up his father's wine.') and stative transitives (e.g., *Julio se supo la lección.* 'Julio came to know the lesson.' vs. *Yengho-nun pimil-ul al-a peli-ess-ta.* 'Yengho came to know the secret.') and are argued to function as aspectual markers always inducing telicity effect (Sanz, 2000; Sanz and Laka, 2002). More recent studies, however, suggest a different analysis for Spanish optional *se*: these constructions can be treated as instances of double object constructions, with *se* restricted to a limited set of stative predicates and no longer necessarily inducing telicity effects (Campanini and Schäfer, 2011; MacDonald, 2017; Martínez-Vera, 2022; Martin and Arunachalam, 2022). In contrast, the Korean auxiliary verb construction (AVC) has received considerably less attention than its Spanish counterpart (Sohn, 2001; Choi, 2003, 2006; Jung and Kim, 2017).

The goal of this paper is to provide experimental support for the proposed telic interpretation, which poses an apparent challenge to Sohn's analysis of *a/e pelita* as a telicity marker. Specifically, the study investigates whether *a/e pelita* systematically gives rise to a telic interpretation when combined with eventive predicates, making such AVCs telic according to the standard telicity diagnostics (e.g., *every day*, *to do so*). The study also aims to compare *a/e pelita* AVCs to their Spanish counterpart, double object constructions (DOCs) with optional *se*, with respect to exhaustivity and the inference of event completion.

The paper is organized as follows: Section 2 presents the research hypotheses and the methodology of the current experimental study. Section 3 presents the results, and Sections 4 and 5 present the discussion and conclusion.

[5]A reviewer raised an important point regarding the behavior of these additional stative predicates when combined with *a/e pelita*. Although this is an important issue, due to space limitations, we leave it for future studies to further investigate this phenomenon. We appreciate the reviewer's valuable comment.

2 Research Method

2.1 Research Questions and Hypotheses

The research questions are the following:

- Research Question 1: Does the presence of *a/e pelita* give rise to a telic VP even when combined with eventive predicates, and does this telic effect extend to stative predicates?
- Research Question 2: Does *a/e pelita* trigger a strong inference of event completion only with a consumption verb? Alternatively, does verb type affect the strength of the event completion inference?

The hypotheses and prediction are:

- Hypothesis 1: *a/e pelita* gives rise to a telic VP when it occurs with eventive predicates but not with stative predicates (Sohn, 1973, 2001).
- Hypothesis 2: *a/e pelita* differs from Spanish optional *se* in its semantic properties and thus, is not expected to pattern the same way as optional *se*.
- Prediction: Unlike Spanish optional *se*, the strength of the event inference completion in *a/e pelita* AVCs is expected to vary depending on the specific verb.

2.2 Participants

Eighty-six native speakers of Korean residing in South Korea were recruited through social media and participated in the experiments online: Experiment I ($n = 44$; $n = 42$) via PCIbex Farm; Experiment II ($n = 42$) via Google Form. Participants received monetary compensation.

2.3 Task, Materials, and Procedure

The experiment consisted of two main tasks: two acceptability judgment tasks (AJTs) and a truth value judgment task (TVJT). The AJTs were designed to test the acceptability and unacceptability of *a/e pelita* AVCs with both eventive and stative predicates, as judged by Korean speakers. The TVJT was used to assess the strength of the event completion inference in comparison with Spanish optional *se* (Martin and Arunachalam, 2022).

In the AJTs, the test materials were presented in two versions: one used standard telicity diagnostics (i.e., *mayil* 'every day', *kulehkey han-ta* 'to do

so'), while the other employed for-adverbial and entailment diagnostics. A between-subjects design was adopted because the two versions differed in tense: the first was presented in the present tense, while the second used the simple past tense in the Spanish version (MacDonald, 2017; Martin and Arunachalam, 2022).

Participants were asked to judge the acceptability of the sentences on a 7-Likert scale (1: totally unacceptable–7: totally acceptable). There were 48 items—16 target items and 32 fillers—in each test version. The target items had four conditions that varied in terms of the construction (base vs. *a/e pelita* AVC) and the type of diagnostics (*every day*, *to do so*). The version of the test with other telicity diagnostics had the same design. Sample target items are shown in (5)–(8).

AJT Version I:

(5) a. Minsu-nun mayil khephi
Minsu-NOM everyday coffee
han can-ul
a cup-ACC

masi-n-ta.
drink-PRES-DEC

'Minsu drinks a cup of coffee every day.'

b. Minsu-nun mayil khephi
Minsu-NOM everyday coffee
han can-ul
a cup-ACC

masi-e peli-n-ta.
drink-e peli-PRES-DEC

'Minsu drinks down a cup of coffee everyday.'

(6) a. Chanswu-nun mayil ku
Chanswu-NOM every day the
iyaki-lul
story-ACC

mitnu-n-ta.
believe-PRES-DEC

'Chanswu believes the story every day.'

b. Chanswu-nun mayil ku
Chanswu-NOM every day the
iyaki-lul
story-ACC

mit-e peli-n-ta.
believe-e peli-PRES-DEC

'Chanswu believes the story every day.'

(7) a. Yengho-nun achimey khephi-lul
Yengho-NOM morning coffee-ACC
masi-ko Cinswu-to
drink-CONJ Cinswu-also
kulehkey-han-ta.
do.so-PRES-DEC

'Yengho drinks a coffee in the morning, and Cinswu does so too.'

b. Yengho-nun achimey khephi-lul
Yengho-NOM morning coffee-ACC
masi-e peli-ko Cinswu-do
drink-e peli-CONJ Cinswu-also
kulehkey-han-ta.
do.so-PRES-DEC

'Yengho drinks down a coffee in the morning, and Cinswu does so too.'

(8) a. Yengho-nun mayil ku iyaki-lul
Yengho-NOM every day that story-ACC
mit-ko Minho-do
believe-CONJ Minho-also
kulehkey-han-ta.
do.so-PRES-DEC

'Yeongho believes the story, and Minho does so too.'

b. Yengho-nun ku iyaki-lul mit-e
Yengho-NOM that story-ACC believe-e
peli-ko Minho-do
peli-CONJ Minho-also
kulehkey-han-ta.
do.so-PRES-DEC

'Yengho believes the story, and Minho does so too.'

AJT Version II:

(9) a. Yengswu-nun 10-pwun tongan
Yengswu-NOM 10 minutes for
sakwa-lul meok-ess-ta.
apple-ACC eat-PAST-DEC

'Yengswu ate an apple for 10 minutes.'

b. Yengswu-nun 10-pwun tongan
Yengswu-NOM 10 minutes for
sakwa-lul meok-e peli-ess-ta.
apple-ACC eat-e peli-PAST-DEC

'Yengswu ate up an apple for 10 minutes.'

(10) a. Minci-nun halwu tongan ku
Minci-NOM one day for that
iyaki-lul mit-ess-ta.
story-ACC believe-PAST-DEC

'Minci believes that story for a day.'

b. Minci-nun halwu tongan ku
Minci-NOM one day for that
iyaki-lul mit-e peli-ess-ta.
story-ACC believe-e peli-PAST-DEC

'Minci believes that story for a day.'

(11) a. Yenghuy-nun sakwa-lul
Yenghuy-NOM apple-ACC
mek-ess-ciman, acik celpan-ul te
eat-PAST-but yet half-ACC more
mek-eya ha-n-ta.
eat-must do-PRES-DEC
'Yenghuy ate (an) apple, but she still needs to eat the rest of it.'

b. Yenghuy-nun sakwa-lul mek-e
Yenghuy-NOM apple-ACC eat-PAST
peli-ess-ciman, acik celpan-ul te
peli-but, yet half-ACC more
mek-eya ha-n-ta.
eat-must do-PRES-DEC
'Yenghuy ate up (an) apple, but she still needs to eat the rest of it.'

(12) a. Chelswu-nun ku
Chelswu-NOM the story
iyakilul mitessciman, acik
believe-PAST-but still more
te miteya han-ta.
believe need-to-DEC
'Chelswu believed the story, but he needs to believe it more fully.'

b. Chelswu-nun ku iyakilul
Chelswu-NOM the story believe
mite peli-essciman, acik te
peli-e-but still more believe
miteya han-ta.
need-to-DEC
'Chelswu believed the story, but he needs to believe it more fully.'

In the TVJT, forty-two Korean speakers were shown brief video clips (complete action vs. incomplete action) paired with sentences (28 test items), as illustrated in (13)–(14). The video clips depicted either a partially completed event (e.g., eating 50% to 80% of a cookie) or a fully completed event (e.g., eating all of a cookie). Since the purpose of this task was to compare *a/e pelita* to optional *se*, building on previously utilized tests, the TVJT followed the overall design of prior studies (Arunachalam and Kothari, 2010, 2011), including the same video materials. The test items included one consumption verb (*mek-ta* 'eat'), one creation verb (*kulita* 'paint'), and five change-of-state verbs (*kkeokta* 'pick up', *tephta* 'cover', *kkuta* 'turn off', *tatta* 'close', *chaywuta* 'fill'). [6]However, two modifications were made relative to Martin and Arunachalam (2022): 1) fillers (28 items) were added, and 2) the number of TRUE and FALSE responses in the fillers was balanced.

TVJT:
[complete action video] FULL condition

(13) a. Kunye-nun khwukhi-lul mek-ess-ta.
she-NOM cookie-ACC eat-PAST-DEC
'She ate (a/the) cookie.'

b. Kunye-nun khwukhi-lul mek-e
she-NOM cookie-ACC eat-e
peli-ess-ta.
peli-PAST-DEC
'She ate up the cookie.'

[Incomplete action video] PART condition

(14) a. Kunye-nun khwukhi-lul mek-ess-ta.
she-NOM cookie-ACC eat-PAST-DEC
'She ate (a/the) cookie.'

b. Kunye-nun khwukhi-lul mek-e
she-NOM cookie-ACC eat-e
peli-ess-ta.
peli-PAST-DEC
'She ate up the cookie.'

2.4 Data Analysis

All analyses were conducted using R. For the AJT, the *lmer4* package for linear mixed-effects model was used (Bates, 2011). A linear mixed-effects model was carried out with subjects and items as random factors. For the TVJT, descriptive analyses were conducted.

3 Results

3.1 Experiment I

Mean acceptability scores showed that native speakers of Korean ($n = 44$) found *a/e pelita* AVCs with eventive predicates and telicity diagnostics to be acceptable, with only slight variability (see Figures 1 and 2).

A linear mixed-effects model was fit with predicate type, construction type, and diagnostic type as fixed effects, and with participants and items as random intercepts. The analysis revealed significant main effects of predicate type ($\beta = 0.77$, $p = .0106$) and construction type ($\beta = -0.57$, $SE = 0.233$, $t(8) = -2.470$, $p = .0387$), indicating that eventive predicates were rated significantly higher than stative predicates, and base constructions were rated higher than auxiliary constructions.

[6]In the TVJT test sentences used by Martin and Arunachalam (2022), the subject pronoun *Ella* 'she' was deliberately included to prevent anticausative or *se*-passive uses of change-of-state verbs when *se* is present (cf. Fábregas, 2021). We adopted the same approach to ensure greater comparability between our study and theirs.

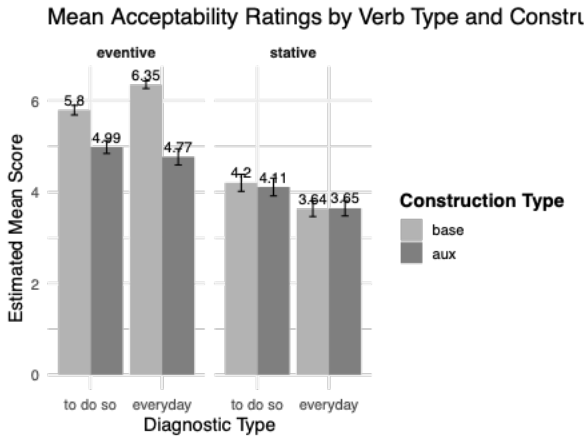


Figure 1: Mean acceptability ratings by predicate type, diagnostic, and construction.

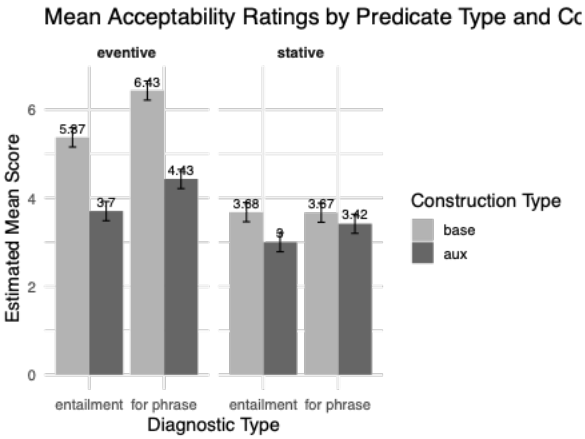


Figure 3: Mean acceptability ratings by predicate type, diagnostic, and construction.

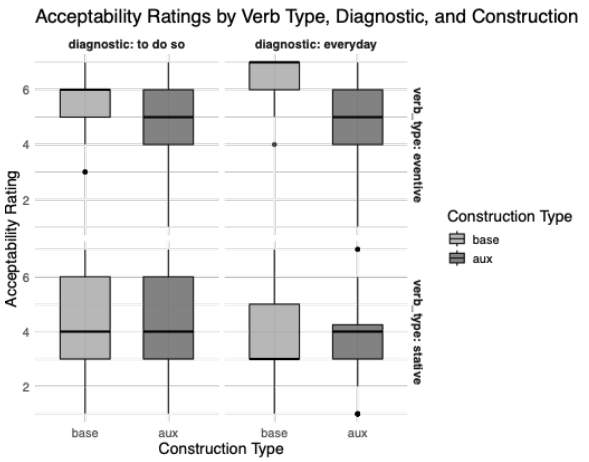


Figure 2: Acceptability ratings by predicate type, diagnostic, and construction.

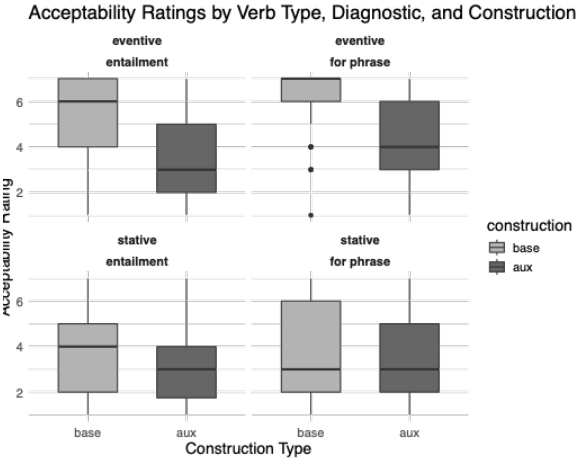


Figure 4: Acceptability ratings by predicate type, diagnostic, and construction.

The main effect of diagnostic type was not significant ($p = .273$), and no significant two-way or three-way interactions were found (all $ps > .17$).

Post-hoc Tukey-adjusted comparisons indicated that for eventive predicates in the everyday diagnostic condition, base constructions were rated significantly higher than auxiliary constructions (estimate $= 1.58$, $SE = 0.68$, $t(8) = 2.31$, $p = .0495$). No other pairwise comparisons reached significance (all $ps > .27$).

Another AJT was conducted with Korean speakers ($n = 42$) using entailment and for-adverbial diagnostics (see Figures 3 and 4).

A linear mixed-effects model was fit with predicate type, construction type, and diagnostic type as fixed effects, and with participants and items as random intercepts. The analysis revealed significant main effects of predicate type ($\beta = 0.79$, $t(623) = 4.62$, $p = .0017$), construction type ($\beta = -0.70$, $t(623) = -2.61$, $p = .0092$), and diagnostic type ($\beta = 1.67$, $t(623) = 6.20$, $p < .001$). This indicates that eventive predicates received higher ratings than stative predicates, base constructions received higher ratings than auxiliary constructions, and items in the for-adverbial diagnostic condition were rated higher than those in the entailment condition. A significant predicate type $\times$ construction type interaction was found ($p = .0095$), while all other interactions were not significant (all $ps > .15$). Post-hoc Tukey-adjusted comparisons showed that, within both diagnostic conditions, base constructions were rated significantly higher than auxiliary constructions (entailment: $p < .0001$; for-adverbial phrase: $p < .0001$). Among eventive predicates, the difference between base and auxiliary constructions also reached significance ($p < .0001$)[7].

[7]The residuals of the model were approximately symmet-

3.2 Experiment II

In the TVJT, participants ($n = 42$) responded with either TRUE or FALSE. The data were first analyzed descriptively by calculating the proportion of TRUE responses by condition and verb, as shown in Figure 5.

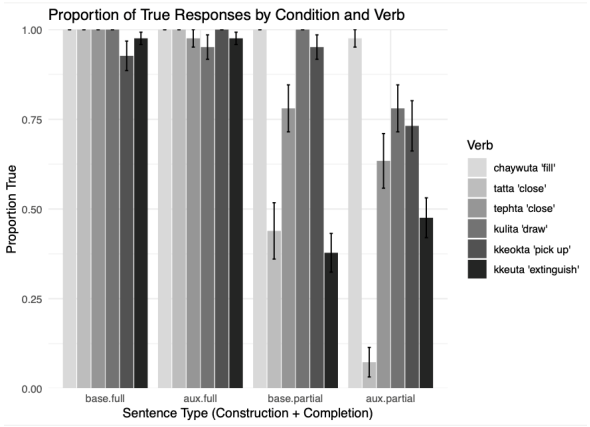


Figure 5: Proportion of TRUE responses by condition and verb.

The results indicate that, unlike Spanish optional *se*, the acceptability of *a/e pelita* AVCs was affected by particular verbs. In other words, the marked *a/e pelita* AVC variant was less acceptable than the unmarked variant under both the FULL and PARTIAL conditions. This pattern is expected, as in *a/e pelita* AVCs with eventive predicates, *a/e pelita* signals event completion[8].

The percentage of TRUE responses across all verbs tested, compared with the results for Spanish optional *se* reported in Martin and Arunachalam (2022), is presented in Table 1. Unlike DOCs with optional *se*, the marked *a/e pelita* AVC variant was less acceptable than the unmarked variant in the PART condition and was as acceptable as the unmarked variant in the FULL condition.

Unlike Spanish *se*, *a/e pelita* AVCs were generally accepted in the FULL condition for all the verbs tested. This includes not only the consumption verb *mekta* 'eat' but also other verbs such as *tatta* 'close,' *tephta* 'close,' and *kulita* 'draw.' As shown in Table 3, however, in the PART condition, *tatta pelita* 'close completely' was accepted less often than *comerse* 'eat up.'

	Base	***a/e pelita* AVCs**	***- se***	***+ se***
FULL	98.3	98.3	97	70
PARTIAL	70.4	59.2	61	46

Table 1: Percentage of TRUE responses for Korean test items with all verbs (compared to TRUE responses for all verbs in Spanish).

	mekta 'eat'	***mek-e pelita*** 'eat up'	***comer*** 'eat'	***comerse*** 'eat up'
FULL	100	95.1	100	100
PARTIAL	100	78	88	73

Table 2: Percentage of TRUE responses for Korean test items with the consumption verb *mekta* 'eat'.

	tatta 'close'	***tatt-a pelita*** 'close up'	***comer*** 'eat'	***comerse*** 'eat up'
FULL	100	100	100	100
PARTIAL	43.9	7.3	88	73

Table 3: Percentage of TRUE responses for Korean test items with the verb *tatta* 'close'.

These results suggest that, unlike Spanish optional *se*, *a/e pelita* AVCs carry a stronger inference of event completion, and that this inference varies by particular verb, reflecting the telicizing effect of *a/e pelita*. Additionally, most of the Spanish-speaking participants (38 out of 42) were from Latin American countries, which may account for the results observed for Spanish optional *se* (Martin and Arunachalam, 2022). As Martin and Arunachalam (2022) note, for LOW-APPL speakers (mostly Peninsular Spanish speakers), optional *se* constructions with consumption verbs, for example, are treated as a type of double object construction. In contrast, for LOW/HIGH-APPL speakers (primarily Latin American Spanish speakers), these constructions are not analyzed as double object constructions (MacDonald, 2017).

4 Discussion

The study addressed the following hypotheses and prediction:

rically distributed with a median near (Median = .041) and interquartile range from -0.81 to 0.67, suggesting reasonable model fit.

[8]We also observed an unexpected deviation in the performance of Korean speakers with the verbs *kkeokta* 'pick up' and *kkeuta* 'extinguish', as the acceptability of the marked AVC variant was higher than that of the unmarked variant. It is evident that these two items influenced the overall results. The first item involved an incomplete event in which an actress attempted to pick a banana from a bunch but failed to do so, which may have led participants to interpret the event as unsuccessful. Therefore, the acceptability ratings in both conditions (base partial vs. aux partial) were lower than in other conditions. The effect of the second item was even more pronounced, with acceptability ratings significantly lower in both conditions.

Hypothesis 1: *a/e pelita* gives rise to a telic VP when it occurs with eventive predicates but not with stative predicates (Sohn, 1973).

Hypothesis 2: *a/e pelita* differs from Spanish optional *se* in its semantic properties and, thus, is not expected to pattern the same way as optional *se*.

Prediction 1: As a telic marker, *a/e pelita* is expected to produce a telic VP with eventive predicates but not with stative predicates.

Prediction 2: Unlike Spanish optional *se*, the strength of the event inference completion in *a/e pelita* AVCs is expected to vary depending on particular verbs.

The results of the study support our hypotheses grounded in the aspectual approach. Specifically, *a/e pelita* AVCs with eventive predicates consistently yielded telic VPs, whereas those with stative predicates did not, particularly when evaluated using standard telicity diagnostics. This seems to support the assumption that Spanish DOCs with optional *se* can be equated with Korean *a/e pelita* AVCs. However, findings suggest that the telic marker does not block aspectual coercion in which compositional aspect interacts with contextual/pragmatic factors to override the telic interpretation in Korean *a/e pelita* AVCs, which contrasts with behaviors of optional *se*. Consequently, this means that Korean *a/e pelita* differs from Spanish optional *se* in the semantic property of telicity and that Korean *a/e pelita* AVCs cannot be equated with Spanish DOCs with optional *se*.

Furthermore, in comparison to the Spanish optional *se*, the Korean auxiliary gave rise to a stronger inference of event completion, and the exhaustivity inference varied depending on the specific verb involved. The semantic contrast, in terms of telicity and atelicity, between Korean *a/e pelita* and its Spanish counterpart (optional *se*) suggests that Korean *a/e pelita* AVCs cannot be straightforwardly equated with Spanish DOCs with optional *se*. This contrast comes from differences in semantic properties: the Korean *a/e pelita* induces an event completion inference depending on the verb with which it is paired. Additionally, the Korean *a/e pelita* AVC variant is acceptable as the unmarked variant in the FULL condition with any verb, whereas the optional *se* variant is not acceptable as the unmarked variant in the FULL condition with most verbs except the consumption verb *comer* 'eat'. However, in the PARTIAL condition, the marked *a/e pelita* AVC variant was less acceptable than the unmarked variant in the PARTIAL condition. This means that *a/e pelita* AVCs carry a stronger inference of event completion; and that this inference varies with the verb, reflecting the telicizing effect of *a/e pelita*.

In second language acquisition, these semantic differences mean that L1-Korean L2-Spanish learners need to acquire the semantic conditions of Spanish DOCs with optional *se* that indicate nuanced telicity.

5 Limitation and Conclusion

This study experimentally investigated whether the existing two approaches to the Korean auxiliary *a/e pelita* make accurate predictions, and whether this auxiliary differs from its counterpart, the Spanish optional *se*. Results from Experiment I support the aspectual approach (Sohn, 1973, 2001), showing that *a/e pelita* triggers a telic VP with eventive predicates in Korean, but not with stative predicates. Results from Experiment II suggest that, unlike Spanish optional *se*, *a/e pelita* carries a stronger inference of event completion than the base construction without the auxiliary. Moreover, it marks exhaustivity—unlike *se* and the English particle *up*—though the strength of this exhaustivity inference varies depending on the specific verb.

Although this study compared results obtained from Korean speakers to those obtained for Spanish speakers in a previous study to clarify the differing semantic properties of Korean *a/e pelita* and Spanish optional *se*, Martin and Arunachalam's (2022) results included data from speakers of both Peninsular Spanish and various Latin American varieties. Therefore, it is not certain that all Latin American Spanish varieties are homogeneous. Peninsular Spanish itself also showed variability. Consequently, future study should focus on the nuanced variability in different varieties of Spanish.

Acknowledgments

We are deeply grateful to Dr. Sudha Arunachalam for her genericity in sharing the video clips used in her studies.

References

Antonio Fábregas. 2021. SE in Spanish: Properties, structures and analyses. *An International Journal of Spanish Linguistics*, 10(2): 1-236.

Dal Oh Hong. 2015. Cognitive linguistic analysis of the meanings of auxiliary verbs: Focusing

on the composition of 'beorida,' 'chiuda' and 'naeda'. *Discourse and Cognition*, 22(3): 99-123.

Gabriel Martínez-Vera. 2022. Revisiting aspectual se in Spanish: telicity, statives, and maximization. *The Linguistic Review*, 39(1): 159-202.

Grant Armstrong and Jonathan E. MacDonald, (Eds.). 2021. *Unraveling the Complexity of SE*. Springer.

Gutiérrez, J. De la Mora. 2023. Valores pragmáticos del clítico se: la desviación de la norma y las contra-expectativas del hablante. *Signos Lingüísticos*, 19(37): 96-117.

Hagit Borer. 2005. *Structuring Sense II: The normal course of events* (Vol. 2). OUP Oxford.

Henk J. Verkuyl. 1989. Aspectual classes and aspectual composition. *Linguistics and philosophy*, 39-94.

Ho-Min Sohn. 2001. *The Korean language*. Cambridge and New York: Cambridge University Press.

Ho-Min Sohn. 1973. Coherence in Korean "Auxiliary" Verb Constructions. *Language Research*.

Hyun Kyoung Jung and Lan Kim. 2017. Auxiliary Verbs as Head-adjoined Expressives in Korean: Against the Aspectual Approach.

James Hye Suk Yoon. 2018. Korean Syntax, *Oxford Research Encyclopedia of Linguistics*, Oxford: Oxford University Press.

Jonathan E. MacDonald. 2017. Spanish aspectual se as an indirect object reflexive: The import of atelicity, bare nouns, and leísta PCC repairs. *Probus*, 29(1): 73-117.

Liina Pylkkänen. 2002. *Introducing Arguments*. Doctoral dissertation, MIT.

Martin Fabienne and Sudha Arunachalam. 2022. Optional se constructions and flavors of applicatives in Spanish. Isogloss. *Open Journal of Romance Linguistics*, 8(4): 1-34.

Monserrat Sanz. 2000. *Events and predication: A new approach to syntactic processing in English and Spanish*. Amsterdam/Philadelphia: John Benjamin Publishing.

Monserrat Sanz and Laka Itziar. 2002. Oraciones transitivas con se: el modo de acción en la sintaxis. In Cristina Sánchez (ed.), *Las construciones con se*, Madrid: Visor Libros, pp. 311-343.

Seongsook Choi. 2003. Serial verbs and the empty category. In *Proceedings of the Workshop on Multi-VerbConstructions*.

Seongha Rhee. 2008. At the borderland of lexis and grammar: grammaticalizing perfective markers in Korean: grammaticalizing perfective markers in Korean. *Discourse and Cognition*, 15(3): 29-59.

So-Young Park. 2011. A syntax-driven approach to aspects in Korean: arguing against the lexical specification approach of Cho (2007), *Morphology*, 13(2): 335-361.

Sujeong Kim, Heejeong Ko, and Hyun-Kwon Yang. 2020. Telicity and mode of merge in L2 acquisition of resultatives. *Language acquisition*, 27(2): 117-159.

Susan Strauss. 2003. Completive aspect, emotion, and the dynamic eventive: the case of Korean V a/e pelita, Japanese V-te shimau, and Spanish se, *Linguistics* 41(4): 653-679.

Susan Strauss, Jihye Lee, and Kyungja Ahn. 2006. Applying conceptual grammar to advanced-level language teaching: the case of two completive constructions in Korean. *The Modern Language Journal* 90(2): 185-209.

‘But this one was so . . . male.’ A Corpus-Based and LLM-Augmented Analysis of Language and Gender Bias in *Barbie*

Xin Luo[1] and Wing-hei Lok[2] and Yuyin-Hsu[3]
Department of Language Science and Technology
The Hong Kong Polytechnic University
[1]tracyxin.luo@polyu.edu.hk, [2]winghei-lok@connect.polyu.hk,
[3]yu-yin.hsu@polyu.edu.hk

Abstract

The 2023 film *Barbie* has sparked discussions on women’s empowerment and patriarchal norms, weaving feminist themes with critiques of gender roles while subtly reflecting patriarchal undertones that marginalize women (Myisha et al., 2023). We conduct a corpus-based analysis to investigate gender bias and differences in utterances’ distribution of part-of-speech (POS), affective values, and gender-linked classifications across three distinct scenes: 1) Barbie Land, 2) Real World, and 3) Post-Barbie Land, each representing the matriarchal, conventional and patriarchal theme, respectively. Leveraging large language models (LLMs), we extend Bradley and Lang’s (1999) affective norms to assess the emotional properties of film scripts, marking a novel application of LLMs in film script analysis.

Findings reveal gender-based differences and reflect power dynamics in and across scenarios, supporting prior research regarding gendered lexical preferences (e.g., Argamon et al., 2003; Jasmani et al., 2011). The results of affective analysis indicate that females consistently demonstrate higher valence and arousal scores across scenes, aligning with feminine communication styles emphasizing emotional rapport and expressivity, whereas males prioritize dominance (Tannen, 1990; Holmes and Stubbe, 2003). These variations underscore how affective language use is co-constructed by gender and socio-narrative context. Notably, gender-linked classifications (Barbie/Ken tones) align more closely with theoretically grounded gendered linguistic features than the broad stereotypes. This study highlights the linguistic construction of gender roles and power dynamics in *Barbie*, offering insights into the interplay between subversive feminism and traditional patriarchal norms and demonstrating how language reflects and challenges gender stereotypes in the film.

1 Introduction

The 2023 film *Barbie* emerges as a cultural product that simultaneously promotes feminist narratives to mass audiences while critically engaging with persistent gender biases and patriarchal stereotypes - embodying what Byrnes (2025) terms ‘feminist ambivalence’. Within this framework, the language employed in the movie is designed to resonate with post-feminist ideas that emphasize women’s individuality, autonomy, and inherent uniqueness as distinct self-determining beings (Gill, 2007).

Previous studies have identified the distribution of POS as a key indicator of gendered language, with specific POS patterns reflecting distinct gendered linguistic features. Key’s (1972) foundational work on gender differences in linguistic behavior first observed variations in pronominal and nominal referents between male and female speakers. Lakoff (1975) later contextualized these differences within broader social hierarchies, arguing that they stem from women’s historically marginalized status – a dynamic that incentivizes women to prioritize linguistic markers of prestige, including more standardized forms than men (Trudgill, 1972; Labov, 1990).

Subsequent studies have elaborated on gendered POS tendencies. The feminine style, for instance, is characterized by a greater use of emphatic adverbs, such as *so, really* (Schofield and Mehr, 2016) and modal auxiliaries that soften declarative statements (McMillan et al., 1977; Biber et al., 1998; Mulac et al., 2001; Mehl and Pennebaker, 2003). These patterns link to Biber et al.’s (1998) framework of ‘involved’ (subjective, interpersonal) and ‘informative’ (objective, content-focused) speech, where feminine styles tend towards the former. Pronouns, one of the widely studied POS categories, reveal robust gender differences particularly in generic usage (e.g., *he, they*) and occupational referents (e.g., *nurse, doctor, president*) (Luo and Huang,

2024). Nouns, however, show more variable patterns. Binnenpoorte et al. (2005) found no significant gender-based differences in noun usage within telephone dialogues, while Argamon et al. (2003) documented notable disparities in fictional texts. This inconsistency underscores the critical role of contextual factors in shaping gendered language patterns and highlights the need for context-sensitive analysis in POS-based gender research.

Beyond POS distribution, this study examines gendered language through two dimensions: the Affective Norms of English Words (ANEW) and the Barbie/Ken tone (hereafter 'BT/KT'), categorized using the LLMs. Affective connotation of words exert a pervasive influence on cognitive processing, and a growing body of research documents gender-based disparities in how such words are employed. A critical challenge in this domain lies in selecting appropriate terms – a choice that underscores the importance of measuring words' affective meanings for scholars investigating gendered language and its role in shaping power dynamics. While existing literature primarily focuses on generating affective ratings for words across languages, including exploring how gender influences participants' evaluations (Warriner et al., 2013; Montefinese et al., 2014), far fewer studies investigate how ANEW-related words are employed differently by males and females. Notably, gender has been shown to correlate significantly with arousal and dominance ratings, aligning with the stereotype that women are perceived as more emotionally expressive than men (e.g., Fischer, 2000). One relevant investigation is Marrville's (2017) thesis, which found that verbs associated with high emotional dominance were strongly association with male characters, whereas those conveying low emotional control were more closely associated with female characters. However, this study is limited in scope focusing exclusively on verbs and examines only the dominance dimension with a self-paced reading task.

Gender stereotypes serve as a foundational lens in gender research, as they encapsulate culturally normalized perception of 'appropriate' behaviors for different genders. These stereotypes often manifest as dichotomous constructs, framing gendered features as distinct and mutually exclusive (Holmes and Stubbe, 2003). For instance, women communicate indirectly, while males adopt a direct communication style. However, scholars have increasingly challenged the rigidity of such dichotomy, noting that gendered linguistic features rarely align perfectly with stereotypes across all contexts. Cameron (1998) attributed this inconsistency to the collaborative nature of conversation, emphasizing that interactional dynamics, rather than fixed gender traits, shape communicative behavior. Similarly, Bing and Bergvall (1998) argued that male and female language use and behavior exist on an overlapping continuum, resisting strict categorization. This study employs a corpus-based approach, augmented by LLMs, to investigate gendered linguistic patterns and affective language in the 2023 film *Barbie*. Focusing on the tension between feminist and patriarchal ideologies, we analyze three key linguistic features: distribution of POS, affective values of word choices, and gender-linked classifications across different thematic scenarios.

2 Data and Methodology

The primary dataset comprises the complete transcript of the 2023 film *Barbie*, retrieved directly from official screenplay records. This source was selected for its authoritative representation of gendered dialogue, which explicitly navigates tensions between feminist and patriarchal ideologies. Text processing was processed using the Natural Language Toolkit (NLTK; Bird et al., 2009) to ensure systematic and replicable data preparation. The details of the data are summarized in Table 1. To analyze gendered language across contrasting power structures, the transcript was strategically subdivided into three sub-scenarios for targeted thematic and linguistic analysis: 1) Barbie Land (PreB) which was characterized by a matriarchal social order where female characters hold positions of power and agency, prior to narrative disruptions; 2) Real World (RW), capturing the gendered dynamics in a societal context that reflects conventional patriarchal norms; and 3) Post-Barbie Land (PostB), representing the altered Barbie Land under patriarchal influence. This subdivision captures the film's deliberate thematic juxtaposition—allowing direct comparison of linguistic patterns under matriarchal, patriarchal, and transitional power dynamics.

Drawing on established research documenting gendered differences in lexical and POS distributions (Argamon et al., 2003; Litvinova et al., 2017), we hypothesized that POS usage would exhibit significant variation across the film's sub-scenarios

	Females		Males	
Scenarios	Tokens	Word	Tokens	Word
PreB	2308	612	483	185
RW	2265	612	1626	524
PostB	3766	794	1874	598
Total	8339	2018	3983	1307

Table 1: Description statistics of dataset

(PreB, RW, PostB), reflecting their distinct gendered linguistic contexts. This hypothesis aligns with Key's (1972) proposal that grammatical categories serve as linguistic markers of gendered performance, which may be amplified or transformed in narratives explicitly addressing gender norms. Our POS analysis targeted two levels of granularity to capture both specific and broader functional patterns, including individual POS categories and functionally relevant POS groups (e.g., adjectives + adverbs (JJs+RBs); all verb forms (VBs)). We adopted the Affective Norms of English Words (ANEW) framework proposed by Bradley and Lang (1999), which operationalizes emotions along three dimensions: valence (ranging from 1 = unpleasant to 9 = pleasant), arousal (1=calm to 9 = excited), and dominance (1 = out of control to 9 = under control). To situate POS patterns within emotional and evaluative contexts, we applied this framework with established psycholinguistic norms (valence, arousal, and dominance) to evaluate 843 NNs and 361 adjectives and adverbs. We further augmented the analysis with LLM-based methods to extend ANEW's coverage and contextualize the affective scores within the narrative of *Barbie*.

To extend our study beyond traditional linguistic metrics, we employed GPT-3.5 Turbo to categorize the same 1204 target words (843 NNs and 361 adjectives and adverbs) used in our POS and affective analyses into two distinct tonal registers (BT or KT). This classification aimed to capture LLM's underlying words stereotypically associated with feminine (BT) and masculine (KT) linguistic style, respectively. To systematically investigate how different contextual frameworks influence the classification of gendered language, we designed three distinct prompts for a zero-shot classification task. All prompts were presented without examples to elicit the LLM's inherent reasoning and each prompt directly instructs the model to adopt a specific perspective for the 'BT/KT tone' classification task:

1. Film-contextualized (FC): Classify words as 'Barbie tone' or 'Ken tone' based on their usage in the film;

2. Stereotype-driven (SD): Classify words based on widespread gender stereotypes, disregarding movie context;

3. Theory-guided (TG): Classify word using Holmes and Stubbe's (2003) classification of masculine and feminine linguistic features, ignoring both movie context and stereotypes.

3 Results and Discussions

3.1 Distribution of POSs

Significant differences in POS usage emerged between male and female characters across the dataset. As illustrated in Figure 1, male characters demonstrated a significantly higher proportion of proper nouns (NNP) (χ^2=4.8647, p=0.0274), reflecting their tendency to reference specific entities or roles. Female characters employed past tense verbs (VBD) significantly more than male characters (χ^2=4.434, p=0.0352), suggesting a focus on narrating events or experiences. These patterns corborate established gendered linguistic tendencies. The over-representation of nouns in male language use replicates the finding by Argamon et al. (2003), who linked noun dominance to men's stereotypically referential speech in formal contexts. The higher frequency of verbs in female speech corresponds to observations in Jasmani et al. (2011) regarding females' greater use of verbal constructs often tied to relational storytelling.

Analysis of the three sub-scenarios revealed significant context-dependent variations in gendered POS patterns, demonstrating how linguistic behavior adapts to sociocultural power structures. In PreB, female characters used significantly more common nouns (NN) (e.g., *president, doctor*) than males (χ^2=6.4358, p=0.0112) (Figure 2). This aligns with the scenario's thematic focus on female-centric agency, where common nouns often referenced communal roles central to Barbie Land's social structure. No statistically significant gender differences emerged in the distribution of POSs within RW (ps > 0.2435). This absence of divergence may reflect the scenario's portrayal of conventional societal norms, where gendered linguistic markers are less exaggerated than in the film's more ideologically charged settings. A striking

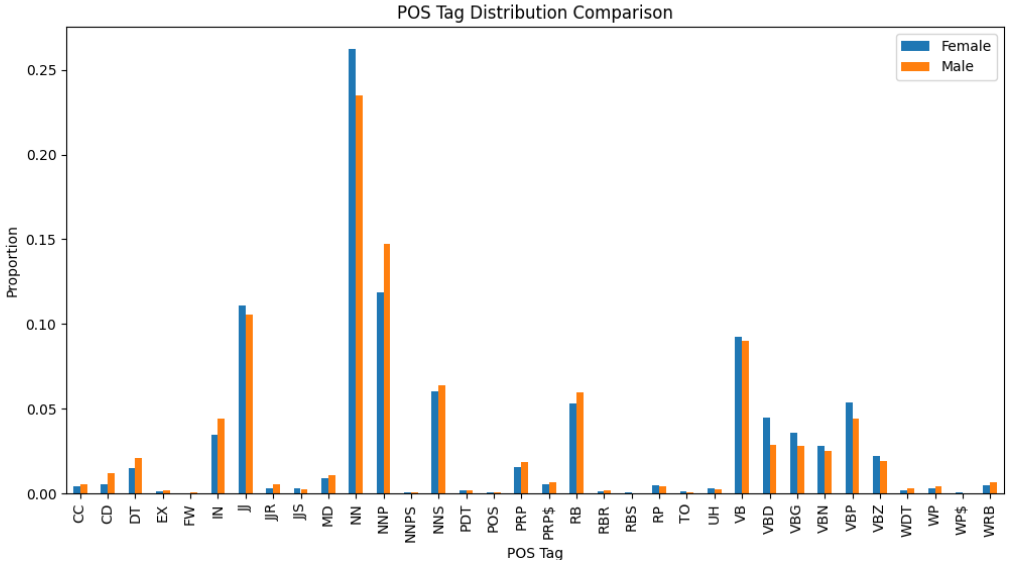


Figure 1: Distribution of POS in overall dataset

disparity was observed in proper noun (NNP) usage, with male characters displaying a significantly higher proportion of NNP in PostB (χ^2=8.0958, p=0.0044) (Figure 3). This pattern intensifies the overall trend of male-centric proper noun use, mirroring the scenario's emphasis on male dominance.

A close analysis of word frequencies within gender groups uncovered nuanced variations in lexical choice. Male characters predominantly referenced male-specific names (e.g., *Ken, Mojo, Aaron*) while female characters frequently referenced *Barbie* in their utterances. This gendered specificity in proper noun usage aligns with Bing and Bergvall's (1998) assertion that language reflects and reproduces power imbalances – centering the names of dominant groups within each scenario–females in PreB, and males in PostB. Female characters showed a marked preference for past tense auxiliary verbs (e.g., *was, did*), whereas male characters rarely used such forms. This aligns with classic hypothesis about gendered language, particularly Lakoff's (1975) observation that women's speech often employs hedging devices, including auxiliaries and Newman et al.'s (2008) finding that female language tends to emphasize relational or retrospective contexts, a pattern reflected in the preferential use of past tense constructions. Furthermore, gendered patterns extended to interpersonal language, with males frequently using terms such as *sir* and *man*, reinforcing hierarchical or peer-based masculinity, while females more commonly utilized the term *mom*, emphasizing relational or familial roles. Taken together, these findings suggest that POS distribution in *Barbie* is not only gendered but also contextually contingent, with power dynamics in each scenario (matriarchal, conventional, patriarchal contexts) shaping the linguistic strategies employed by male and female characters.

3.2 Gendered affective preferences

To ensure a robust and multi-faceted evaluation of affective norms, we leveraged three distinct LLMs, including GPT4o[1], deepseek-v3-fw[2], and Llama[3] to rate the valence, arousal, and dominance of target words. This selection provides architectural and training diversity, enabling cross-model validation and reducing the potential bias inherent in any single model's output. The results of LLM-based ratings showed no statistically significant discrepancies ($p > .05$), confirming the inter-model reliability for evaluating the affective norms of English words.

ANOVA revealed no statistical significant differences in overall affective norm scores between males and female tones ($p > .05$), suggesting a balanced distribution of emotional word usage between genders in the aggregate. However, nuanced gender differences emerged in interaction within scene context (PreB, RW, PostB) and POS categories (adjectives and adverbs (JJs+RBs) and nouns (NNs)).

[1] http://openai.com
[2] https://www.deepseek.com/
[3] https://www.llama.com/

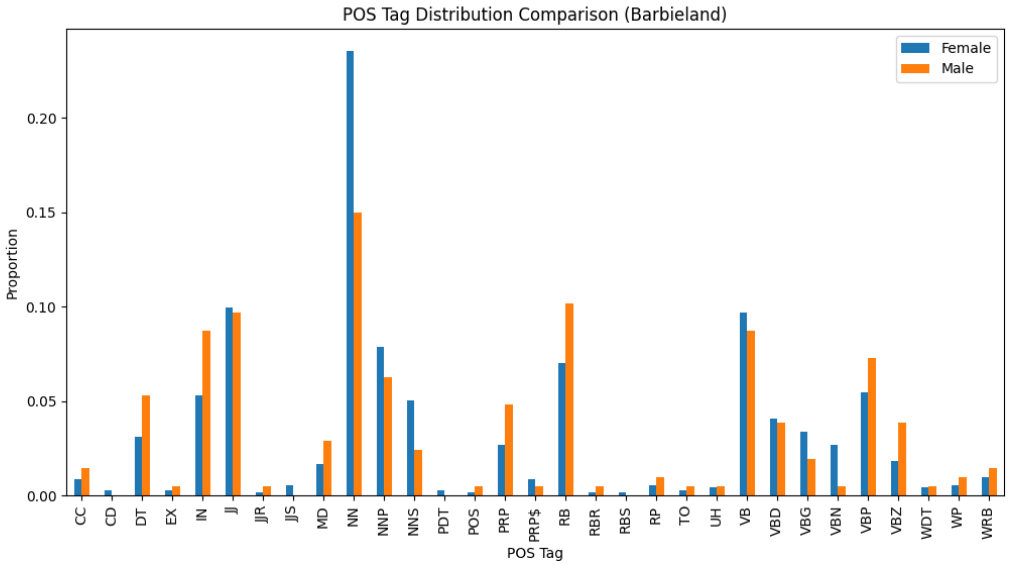


Figure 2: Distribution of POS in PreB

Gender differences in the affective norms of JJs+RBs were most prominent in scenario-specific analyses with significant effects limited to the PreB scene. In PreB as shown in Figure 4, females exhibited significantly higher valence scores for adjectives and adverbs compared to males. Their top 5 words *amazing, beautiful, fun, happy, perfect* – aligned with PreB's joyful, carefree narrative context, where female characters occupied central roles without external constraints. Females also showed elevated dominance scores in JJs+RBs usage, driven by terms such as *powerful, brave, free, perfect, and wonderful*. These lexical choices reflect PreB's thematic focus on female leadership and societal dominance. No statistically significant gender differences in JJs+RBs affective norms were observed in RW and PostB scenes ($p > .05$).

Gender differences in noun usage (NNs) across affective norms were more consistent across scenes, with males generally showing higher dominance scores (Figure 5 and 6). Their top 5 terms contributing to this pattern - *impeccable, best, professional, liberated, rich* – conveyed authority and control. No statistically significant differences were observed in noun valence, arousal, or dominance in PreB ($p > .05$), which aligns with the emphasis of the scene on collective female empowerment over individual hierarchy. Females' noun usage showed higher arousal scores in RW, driven by terms with negative emotional connotations, such as *destroy, anxiety, crisis, death, fascist*. This lexical pattern reflects Barbie's self-reflexive turmoil, blending postfeminist empowerment with underlying existential dread. In contrast, males' nouns in RW, such as *imagination, power, dreams, friends, understanding*, highlighted self-perceptions as leaders and significant figures, resulting in higher dominance scores. Females' nouns had higher valence and arousal scores, while males maintained higher dominance scores in noun usage in PostB.

Females consistently demonstrated higher valence and arousal scores across scenes, reflecting a linguistic emphasis on emotional rapport and expressive intensity – traits associated with feminine communication styles (Holmes and Stubbe, 2003). Males, in contrast, maintained higher dominance scores in noun usage across contexts, except in PreB, where females' elevated dominance in JJs+RBs mirrored Barbie Land's narrative of female societal leadership. These findings underscore how affective language use is co-constructed by gender and socio-narrative context.

3.3 Classification of 'Barbie Tone (BT)' and 'Ken tone (KT)'

GPT-3.5 Turbo was employed to classify all 361 JJs+ RBs into either 'BT' or 'KT' across the three experimental prompts. In Table 2, a significant majority of JJs+RBs were classified as BT, far exceeding KT classification in the film-contextualized classification (237 vs. 124). In contrast, stereotypes-driven classification favored KT over BT (150 vs. 211). Theoretically grounded classification leaned toward BT over KT.

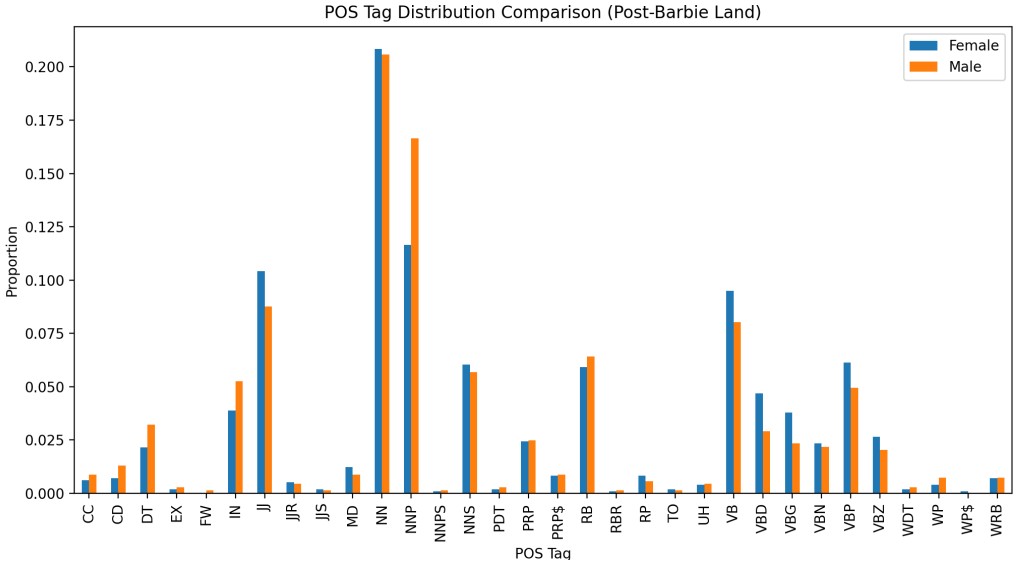


Figure 3: Distribution of POS in PostB

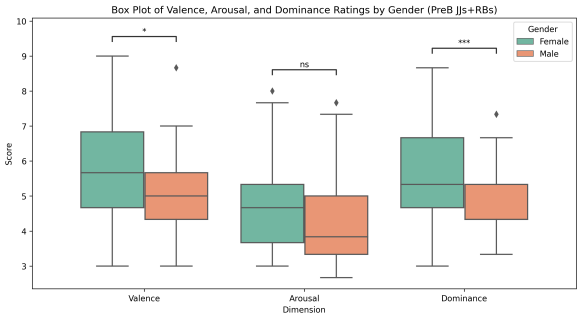


Figure 4: Affective norms value of JJs+RBs in PreB

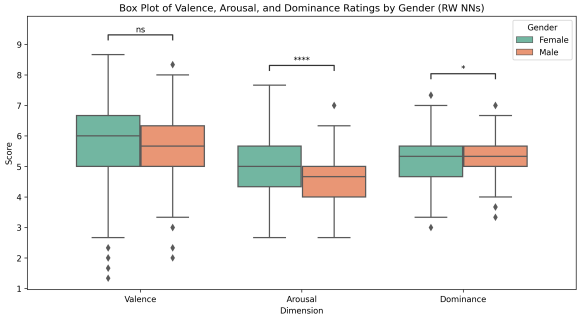


Figure 5: Affective norms value of NN in RW

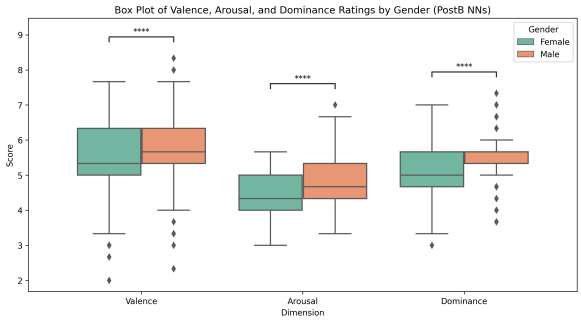


Figure 6: Affective norms value of NN in PostB

A highly significant difference between the film-contextualized (FC) and stereotypes-driven (SD) classification (χ^2=41.1887, $p < .001$), reflecting divergent classification patterns when contextualized film language versus general gender stereotypes. Consistent with Byrnes's (2025) concept of feminist ambivalence in *Barbie*, our quantitative result reveals no statistically significant difference between the FC and theory-guided (TG) classifications after Bonferroni correction for multiple comparisons (χ^2=3.9705, $p = 0.0463 > Bonferroni$ α=0.00167). This statistical alignment suggests the film's linguistic patterns organically embody the theoretical tension between feminine and masculine features. A significant discrepancy between stereotype-driven and theory-guided classifications (χ^2=19.2855, $p < .001$), indicating that general gender stereotypes diverge from theoretically grounded gendered linguistic features. Notable lexical discrepancies emerged for terms such as *com-*

Type	IA	FC	SD	TG
BT	361	237	150	210
KT	0	124	211	151
Total	361	361	361	361

Table 2: Classification of JJs+RBs as "Barbie tone" or "Ken tone" under different prompts

Note: Intrinsic Association (IA) refers to the model's classification based solely on its pre-existing knowledge, without any contextual prompt pr instructional framing.

Type	IA	FC	SD	TG
BT	840	561	335	451
KT	3	282	508	392
Total	843	843	843	843

Table 3: Classification of NNs as "Barbie tone" or "Ken tone" under different prompts

Note: Intrinsic Association (IA) refers to the model's classification based solely on its pre-existing knowledge, without any contextual prompt pr instructional framing.

plicated, exactly, pregnant. These were frequently categorized as KT under stereotypes-driven classification but aligned with BT in both the movie context and theory-guided classification. This divergence highlights a disconnect between culturally dominant stereotypes and actual linguistic features in the film, as well as theoretically defined gendered pragmatics (e.g., *expressive vs. assertive language*, Holmes and Stubbe, 2003).

GPT-3.5 Turbo demonstrated a strong overall bias toward classifying NNs as 'Barbie tone', labeling 840 out of 843 target nouns with this category and assigning only 3 as 'Ken tone' (Table 3). However, prompt-specific analyses revealed substantial variation in classification patterns. When comparing results between FC and SD prompts, a clear majority of nouns were classified as 'Barbie tone' in FC, the SD prompt triggered a sharp shift towards 'Ken tone' classifications. This difference was statistically significant (χ^2=120.5834, $p < .001$). Notably, stereotype-driven 'Ken tone' nouns include *CEO, airplanes, assistant, and autonomy* – terms linked to traditional masculine domains. A similar contrast emerged when comparing FC and TG prompt results, significantly more NNs were classified as 'Barbie tone' in the movie context (561 vs. 451), while the TG condition produced a higher proportion of 'Ken tone' classifications (282 vs. 392, χ^2=29.3677, $p < .001$). The comparison between SD and TG prompt results revealed that stereotype-driven classifications yielded far more 'Ken tone' labels (508 vs. 392), whereas the theory-guided classification resulted in more 'Barbie tone' (335 vs. 451, χ^2=31.5201, $p < .001$).

The analysis revealed no significant difference between the movie context and theory-guided classification for JJs and RBs (χ^2=41.1887, $p < .001$), indicating strong consistency between the film's linguistic portrayal and theoretically defined gendered features. This consistency indicates that the *Barbie* movie's use of descriptive language closely mirrors established linguistic framework (e.g., (Holmes and Stubbe, 2003)). Across both word categories, stereotype-based classifications differed significantly from both the movie context and theory-guided classifications, reflecting a clear mismatch between broad societal stereotypes and either the actual language of the film or theoretically grounded gendered linguistic traits. In particular, for NNs, even the movie context and Holmes and Stubbe's (2003) classification differed significantly, indicating nuanced distinctions in how gendered tones are operationalized in film versus linguistic theory. The tendency is consistent with Byrnes's (2025) concept of feminist ambivalence in *Barbie* that the film's linguistic patterns organically embody the theoretical tension between feminine and masculine features.

4 Conclusion

This study contributes to gendered language research by leveraging the 2023 film *Barbie* as a unique analytical lens to examine the interplay between feminist discourse and linguistic gendered performance. Consistent with Byrnes's (2025) interpretation of feminist ambivalence in this film, we demonstrate how the film's linguistic patterns organically embody contradictions between material culture and patriarchal stereotypes. Through POS distribution, affective norms and gendered-linked style classifications, we uncovered nuanced patterns that illuminate the contextual contingency of gendered language. Our findings revealed that gendered linguistic expressions are profoundly shaped by narrative context. Most notably, male characters employed more proper nouns (NNP), while female characters favored past tense verbs (VBD), consistent with prior research regarding gendered lexical preferences (e.g., Argamon et al., 2003; Jasmani et al., 2011). The results not only confirm the presence of gendered language in *Barbie* but also underscore that linguistic variations between gen-

ders are not static but dynamically co-constructed with socio-narrative environments.

Affective norms analyses further reinforced links between language and gendered communication styles. Females consistently exhibited higher arousal and valence scores across scenarios, aligning with Tannen's (1990) conceptualization of 'rapport talk', a speech style stereotypically associated with femininity. In contrast, males' elevated dominance scores resonated with 'report talk', reflecting a focus on authority and control. In particular, these patterns persist even as the film overtly promoted feminist ideals, indicating that *Barbie* retains and reinforces stereotypical gender norms through these linguistic patterns. This observation was further corroborated by the analysis of gendered tone (Barbie tone and Ken tone), which showed stronger alignment with theoretically grounded gendered features in Holmes and Stubbe (2003) than with broad social stereotypes, suggesting the films' portrayal of gendered language is nuanced, if not entirely subversive. These findings carry implications for understanding gendered communication in media and society. By demonstrating how narrative context shapes linguistic expressions of gender, our study highlights the limitations of decontextualized analyses of gender differences. The film's dual role in promoting feminism while retaining stereotypical linguistic patterns also underscores the complexity of media's role in reproducing or transforming gender forms. This study, while offering novel insights into gendered linguistic patterns and affective language in *Barbie*, is subject to several limitations that warrant consideration. First, regarding the scope of our data, the analysis is restricted to a single film and its scripted dialogue. This design, may limit the generalizability of the findings to broader media representation of gender and cannot capture the spontaneous nature of real-world gendered discourse. Second, concerning our computational methodology, our approach is primarily applicable to the English language only. Our experiments with LLMs aim to study how these models classify gendered words, as a general observation of the performance of LLMs. While our approach offers a robust method to augment and enrich the affective norms of words relevant to our research interests - particularly for cases where human responses are not yet available as a gold standard - it comes with limitations. We believe this methodology can be considered only for studying the English language. Applying it to other languages should be approached with caution, as most LLMs have been predominantly and, at times, biasedly trained on English-language data. Despite these limitations, our methodology offers a viable framework for investigating gendered language in narrative settings. Future research will extend this work by analyzing a broader corpus of media texts (e.g., *films, talk shows*) and comparing cross-cultural representations of gendered speech. By bridging computational linguistics and gender studies, this approach advances our ability to unpack the subtle, context-dependent ways gender is constructed and communicated through language.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Janet M. Bing and Victoria L. Bergvall. 1998. The question of questions: beyond binary thinking. In Jennifer Coates, editor, *Language and Gender - A Reader*, pages 495–510. Blackwell Publishing Ltd.

Diana Binnenpoorte, Christophe Van Bael, Els den Os, and Lou Boves. 2005. Gender in everyday speech and language: a corpus-based study. In *INTERSPEECH*, pages 2213–2216.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology.

Alicia Byrnes. 2025. Surface and depth: ambivalence as postfeminist ideal in barbie. *Feminist Media Studies*, 25(3):767–773.

Deborah Cameron. 1998. Performing gender identity: young men's talk and the construction of heterosexual masculinity. In Jennifer Coates, editor, *Language and Gender - A Reader*, pages 270–284. Blackwell Publishing Ltd.

Agneta Fischer. 2000. *Gender and emotion: Social psychological perspectives*. Cambridge University Press.

Rosalind Gill. 2007. Postfeminist media culture: Elements of a sensibility. *European journal of cultural studies*, 10(2):147–166.

Janet Holmes and Maria Stubbe. 2003. "feminine" workplaces: stereotype and reality. In Janet Holmes and Miriam Meyerhoff, editors, *The Handbook of Language and Gender*, pages 573–595. Blackwell Publishing Ltd.

Mohd Faeiz Ikram Mohd Jasmani, Mohamad Subakir Mohd Yasin, Bahiyah Abdul Hamid, Yuen Chee Keong, Zarina Othman, and Azhar Jaludin. 2011. Verbs and gender: The hidden agenda of a multicultural society. *3L: The Southeast Asian Journal of English Language Studies*, 17(Special Issue):61–73.

Mary Ritchie Key. 1972. Linguistic behavior of male and female. *Linguistics*, 10(88):15–31.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change*, 2(2):205–254.

Robin Lakoff. 1975. *Language and Woman's Place*. Harper & Row, New York.

Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2017. Differences in type-token ratio and part-of-speech frequencies in male and female russian written texts. In *Proceedings of the Workshop on Stylistic Variation*, pages 69–73.

Xin Luo and Chu-Ren Huang. 2024. Analyzing the gendered power dynamics in addressing practices: A corpus-based approach. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 1259–1267, Tokyo, Japan. Tokyo University of Foreign Studies.

Caelan Marrville. 2017. Gender and dominance in action: World view and emotional affect in language processing and use.

Julie R. McMillan, A. Kay Clifton, Diane McGrath, and Wanda S Gale. 1977. Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex roles*, 3(6):545–559.

Matthias R. Mehl and James W. Pennebaker. 2003. The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857–870.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.

Anthony Mulac, James J. Bradac, and Pamela Gibbons. 2001. Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, 27(1):121–152.

Nabila Myisha, Dinda Sabila, Angelia Brigita Maharani, Akira Hilal Ramadhan, and Mirza Fathima Jauhar Kamalia. 2023. Decoding the perpetuation of patriarchal culture in the barbie movie. *Cultural Narratives*, 1(2):71–82.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. William Morrow, New York.

Peter Trudgill. 1972. Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in society*, 1(2):179–195.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

A corpus-assisted metaphor analysis in portraying businesswomen: Diachronic changes in Hong Kong English news

Yifan Li[1], Yanlin Li[2], Kathleen Ahrens[2]
Department of English and Communication
& Research Centre for Professional Communication in English
The Hong Kong Polytechnic University, Hong Kong
[1]yifaan.li@connect.polyu.hk, [2]{yanlin.yl.li, kathleen.ahrens}@polyu.edu.hk

Abstract

Previous studies have explored the use of metaphors in businesswoman-related news in mainstream English media. The study presents a metaphor analysis of Hong Kong English news during the two decades before the handover of Hong Kong to China in 1997. To compare the diachronic changes during this period, our study constructed a corpus comprising news from the South China Morning Post. The corpus comprises two sub-corpora, covering the periods from 1978 to 1987 and 1988 to 1997, respectively. Overall, the two sub-corpora demonstrate a similar pattern across source domains. The most frequent source domains include COMPETITION, JOURNEY, MACHINE, BUILDING, and PHYSICAL OBJECT. This study identifies a mix of traditional masculine and feminine traits in the metaphors used. Drawing on the actions and the social realities that businesswomen encountered, the study provides insights into the reflective value of metaphorical expressions in news discourse.

1 Introduction

Hong Kong thrived as one of the newly industrialized regions in the 20th century. In the 1970s, the Hong Kong economy experienced a 10-year growth rate of 8.9% per annum. The opening-up policy of China further boosted Hong Kong's economic growth, with a 10-year trend of 8-9% growth since 1978. In the early 1990s, the 10-year growth rate stabilized at around 6.5% (Financial Services and the Treasury Bureau, 2025).

The growth in the labor force was a significant factor in the economic success, contributing approximately 2.2% of the total increase of 6.3% from 1980 to 1996. In 1996, the total labor force was 1.5 times that of 20 years ago, while the female labor participation rate increased to 1.8 times. During the 20 years, the proportion of managers and administrators in the female labor force also rose from 0.6% to 7.1%, indicating significant progress towards gender equality in higher positions (Census and Statistics Department, as cited in Chu, 2004).

However, the limited opportunities for women in career development have been a long-standing issue in pre-handover Hong Kong. As traditional Confucian values have influenced the society, women have faced various forms of discrimination, including economic, legal, and social. In terms of political participation, due to the centralized administration style and the influence of "old-boys" networks, women remained underrepresented until the 1960s (Lim, 2015). Although the gender gap has narrowed since 1976, the persistent imbalanced gender distribution in professional and management roles still highlights the challenges faced by women (Chu, 2004).

The growing number of female business leaders and the ongoing issue of gender equality have drawn researchers' increasing attention to investigating the representation of these leaders in news media. Metaphors, as a substantial form of figurative language, are deeply rooted in humans' cognitive system and powerfully shape our way of conceptualizing the world (Lakoff and Johnson, 2003). As individual and social resources jointly influenced metaphors, metaphor analysis may provide insights into the underpinning ideological motivations behind discourse (Charteris-Black, 2004).

This study aims to 1) analyze the metaphors used to depict businesswomen in Hong Kong English news, 2) identify the metaphorical patterns in the representation of businesswomen in Hong Kong English news, and 3) examine the function and semantic changes of metaphors related to the social reality faced by businesswomen before the handover to China in 1997.

2 Literature review

2.1 Corpus-assisted metaphor analysis in political discourse

Metaphor analysis in political discourse has been studied in a wide range of contexts, including public speeches and debates. Recent studies have shown that metaphors often serve as rhetorical devices to achieve specific political objectives, which may vary depending on factors such as gender and personal preferences. For instance, Charteris-Black's (2009) research on metaphors in the UK parliament debate verified the emotive or persuasive function of metaphors in political discourse. Likewise, metaphors in news discourse may simultaneously increase linguistic diversity while influencing readers' perceptions of the information.

One of the most prevalent metaphorical source domains in metaphor research is the WAR source domain. Research on WAR metaphors in politicians' speeches provided detailed analyses of the contextual usage, which reflect the association with an individual politician's positioning. For example, Charteris-Black (2005) revealed that Margaret Thatcher used more WAR metaphors and primarily did so to attack her opponents. In contrast to Thatcher, the WAR metaphors Hillary Clinton used mainly positioned her as a protector or defender rather than a fighter (Ahrens, 2019). Thus, a detailed investigation of the metaphorical language used by an individual female politician may provide valuable insight into the self-description employed in political discourse. In addition, news discourse may complement these studies and offer insight into the image of female professionals through the use of metaphorical language.

2.2 Corpus-assisted metaphor analysis in mainstream English media

Previous studies have explored the use of metaphor to describe businesswomen in mainstream English media. Koller (2004a), for instance, compared metaphors for businesswomen and businessmen in corpora of mainstream English business news (e.g., Forbes and Financial Times). Her study argued that the prevalence of the WAR metaphors in the businesswomen corpus could be a result of imposing masculine labels on women within the male-dominated business world. In contrast, the use of CAREGIVER, DOCTOR, and GARDENER metaphors highlighted the traditional femininity that is often attributed to women. This hybridity in the feminine and masculine tendencies of metaphors may reflect the evolving roles of businesswomen in the male-dominated business sector. More specifically, this implies mixed societal attitudes towards businesswomen. Although they seem to integrate into their careers in a masculine way, the societal expectation of their feminine traits remains deeply rooted.

More recently, Li et al. (2024) further illustrated that the application of WAR metaphors in one of the mainstream English business news outlets, "Bloomberg Businessweek," where metaphors were leveraged to describe businesswomen's experiences addressing different issues amidst changing societal circumstances over time. The study showed that business media used metaphors to capture the challenges faced by businesswomen, reflecting a growing awareness of the public about discussing businesswomen-related issues and shaping public perception of advocating for gender equality in the business sector.

2.3 Corpus-assisted metaphor analysis in the Sinosphere

Metaphors are not only cognitively motivated but also carry strong connotations with the cultural context.

The use of metaphor in Asian cultures has been studied across different regions and within specific areas. Targeting the English newspapers in Mainland China, Hong Kong, and Taiwan, Ahrens and Zeng (2021) compared the source domain variation in the discussion of democracy directly before and after the 2016 US presidential election. They found that the three regions differed in the overall frequency of metaphors and distribution across source domains. Beijing editorials showed the highest frequency of metaphorical expressions. The preferred source domain was associated with the political-cultural context: Hong Kong typically used the BUILDING source domain to conceptualize the election process.

In contrast, Taipei used more JOURNEY metaphors to discuss the future goal of democracy. This research validated the significance of metaphor research in distinguishing the English varieties. In the Hong Kong context, Ahrens et al. (2021) revealed how the use of the BUILDING source domain was tailored to different goals of political leaders in policy addresses. During the sovereignty transition period, Hong Kong Governors and Hong Kong Chief Executives employed

their metaphorical language to achieve a better understanding of the audience on critical issues relevant to Hong Kong's development.

The above research inspires further examination of the metaphorical patterns in Hong Kong English in relation to social changes over time. To track the metaphor changes in Hong Kong's economic development and increasing female labor participation, it is worthwhile to study the pre-handover period.

2.4 Research gap and research questions

Previous studies have demonstrated the dominant role of WAR metaphors in business media and explored the evolving roles of gendered metaphors, including both masculine-oriented and feminine-oriented metaphors. However, most studies have focused on mainstream English media. As an outer circle region according to the Three Circles Model of World Englishes, Hong Kong has received limited attention in metaphor analysis of English news (Kachru, 1990). Additionally, few studies have explored the nuanced differences across various metaphorical source domains and the changes in metaphorical language within Hong Kong English news over time.

To address the research gaps, this study proposes the following research questions:

RQ1. How were businesswomen portrayed in Hong Kong English news before the handover?

RQ2. What were the proportions and general patterns of metaphorical expressions in businesswoman-related news? Did they share similar or distinct functions?

RQ3. What were the diachronic changes in metaphor usage during the two decades in the pre-handover period?

3 Methodology

3.1 Corpus creation

In this study, South China Morning Post (SCMP), the leading English newspaper in Hong Kong, was selected as the corpus data source. The data collection consists of two sub-corpora, each containing the entirety of 20 news articles: Corpus 1, covering the period 1978-1987 (18,358 words), and Corpus 2, covering the period 1988-1997 (12,041 words). Table 1 below shows the distribution of news articles by the published year. Compared with other corpus linguistic studies (Ahrens, 2019; Ahrens and Zeng, 2021; Koller, 2004a; Li et al., 2024), the corpus size was relatively small. We have planned to expand the data after the handover time for further comparison and diachronic analysis.

The data were collected through keyword searching in the ProQuest Historical Newspaper Database and a manual review for news related to businesswomen. The keywords input in the searching bar were: "businesswoman" OR "businesswomen" OR "female entrepreneurs" OR "female executive" OR "female CEO" OR "female business leader" OR "woman entrepreneur" OR "women entrepreneur."

By combining general terms (e.g., businesswomen) with more specific references to women (e.g., female executive), we aim to include more relevant articles in the corpus. However, some mentions about women professionals that were not specified by gender markers may be missing, which inspires future optimization of the search method.

Corpus	Year	No. of Articles	Word count
1978-1987	1980	3	3,304
	1981	2	1,254
	1982	4	3,976
	1983	1	835
	1984	4	3,199
	1985	3	3,445
	1987	3	2,345
Sub-total	**1978-1987**	**20**	**18,358**
1988-1997	1988	4	1,729
	1989	2	1,254
	1990	3	1,555
	1991	2	1,027
	1992	1	533
	1994	1	1,098
	1995	3	2,828
	1996	3	1,797
	1997	1	220
Sub-total	**1988-1997**	**20**	**12,041**
Total	**1978-1997**	**40**	**30,399**

Table 1: SCMP corpus 1978-1997

3.2 Metaphor analysis

According to Charteris-Black (2004), the critical metaphor analysis (CMA) generally involves three stages: metaphor identification, interpretation, and explanation. In the first stage, metaphors are identified through a careful examination of keywords and context. The second stage involves analyzing the relationship between metaphors and discourse construction. The third stage addresses the social actors and the ideological and rhetorical motivations behind the metaphor usage. Following this framework, this study adopts a corpus-assisted approach, guided by Steen et al.'s (2010b) Metaphor Iden-

tification Procedure Vrije Universiteit (MIPVU), to systematically identify metaphors. First, the metaphor-related words were recorded by analyzing the texts. Then, we examined the contexts to identify some potential non-literal meanings and cross-domain mappings. Words originating from other semantic domains that contributed to their contextual meaning were classified as metaphors. In this study, we only included direct metaphors that were more salient, excluding implicit metaphors in the forms of substitution and ellipsis.

To verify the source domains of the identified metaphors, we employed the method proposed by Ahrens and Jiang (2020), consulting resources such as SUMO (Suggested Upper Merged Ontology), WordNet, and online dictionaries (e.g., Longman Dictionary). Based on the identified metaphors and their source domains, we further analyzed the relationship between metaphors and discourse construction. Finally, social factors, as well as the ideological and rhetorical motivations behind metaphor usage, were analyzed to explain how metaphors shape social perceptions.

This study utilized the corpus analysis software AntConc to search for all occurrences in the corpus (Anthony, 2024). After importing the news into AntConc, we conducted a secondary search for occurrences containing the following keywords: businesswoman, businesswomen, entrepreneur(s), executive(s), she, woman (noun), and women (noun). To ensure the relevance of the unmarked terms (e.g., entrepreneur, executive), we manually reviewed the cases by clicking the keywords and using the "file view" interface to examine the context.

During the metaphor analysis process, metaphors were annotated as direct or indirect, depending on whether the subject in the sentence referred to a businesswoman or businesswomen.

4 Findings & Discussion

4.1 The general pattern across source domains

4.1.1 The overview of the use of metaphors

The occurrence of metaphors in the two corpora is similar. The log-likelihood test results indicate that the differences in metaphor usage frequency between the two corpora are not statistically significant (LL = 0.04), suggesting a similar pattern in the two examined decades. The normalized frequency of metaphors is 338 per 100,000 words in Corpus 1, compared to 324 per 100,000 words in Corpus 2.

During the first decade (1978-1987), the South China Morning Post used more metaphors that directly referred to businesswomen than in the second decade (1988-1997). In Corpus 1, metaphors occurred 62 times, of which 49 were direct metaphors. Similarly, in Corpus 2, metaphors occurred 39 times, of which 32 were direct metaphors. Our statistical analysis was based on these direct metaphors.

We used a threshold cumulative percentage of up to 85% to determine the frequent source domains (see Table 2). The following figure shows the normalized ratio of the most frequently used source domains in the two corpora.

Although metaphors are common in communication, they constitute only 16.4% of words in written news, 7.7% in conversations, and 18.5% in academic texts (Burgers, 2016; Steen et al., 2010a). As a result, some source domains with lower raw frequencies were included in the analysis to reflect the overall scarcity of metaphors and their important role while ensuring consistency across the top five most frequent source domains.

	Corpus 1		Corpus 2	
Source domains	**Raw**	**Cum.%**	**Raw**	**Cum.%**
JOURNEY	15	30.61	9	28.13
COMPETITION	10	51.02	7	50.01
BUILDING	9	69.39	5	65.64
PHYSICAL OBJECT	7	83.68	2	71.89
MACHINE	6	95.92	5	87.52

Table 2: The cumulative frequency of the frequently used source domains

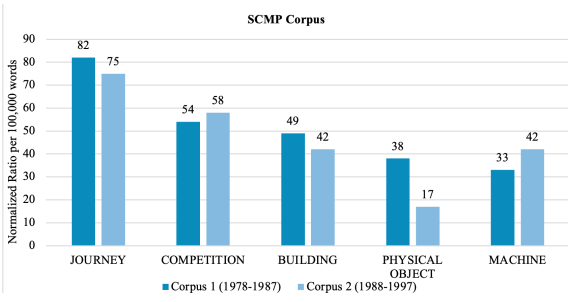


Figure 1: Frequently used metaphorical source domains in SCMP news

4.1.2 The distribution of source domains

While the frequency of metaphor occurrences in both corpora was similar, the distribution of source domains varied between the two time periods. JOURNEY and COMPETITION were the top two frequently used source domains for both decades, followed by BUILDING, PHYSICAL OBJECT, and MACHINE. Regarding the variety of metaphors, the two decades shared seven common source domains: the five most frequent source domains, along with PLANT and ANIMAL. Two unique source domains, PERSON and MONEY, were identified in the second decade.

4.1.3 The gendered orientation of source domains

The perceived gendered orientation of source domains has been discussed through a range of studies. Zeng et al. (2020) employed a qualitative approach to examine the interplay between metaphor, gender, and other political factors, using Hong Kong as a case study. They analyzed the metaphor usage of male and female politicians in relation to the gendered orientation of source domains. The results of the metaphor analysis demonstrated that the use of specific source domains was consistent with their gendered classification and aimed to construct a desirable image of political leadership. For example, the PHYSICAL OBJECT source domain was categorized as neutral in gender orientation. For instance, Female Secretaries for Justice tended to use the PHYSICAL OBJECT source domain in their speeches to enhance objectivity.

More recently, Ahrens et al. (2024) surveyed native English speakers to investigate the gender stereotypes of keywords associated with five prevalent source domains in metaphor research. Their findings showed that COMPETITION and BUILDING are perceived as more masculine, while JOURNEY and PLANT are perceived as more feminine.

The above studies have analyzed the complicated interplay between metaphor and gender. This study aims to enrich the existing research through the perspective of news discourse. Regarding the MACHINE source domain, Philip (2009) proposed that metaphors related to machinery fall into the masculine domain. We followed the classification of gender stereotypes on the most frequently occurring source domains to examine the patterns of metaphor usage in the discussion of businesswoman-related news.

4.1.4 Semantic changes of source domain usage

4.1.4.1 Similarities

In the two sub-corpora, the JOURNEY source domain demonstrates similar semantic functions and contextual usage. As discussed in section 4.2.1, the phased feature of the source domain is frequently employed to describe the actions taken by female entrepreneurs at various stages of business development. Additionally, the JOURNEY metaphors are also extended to suggest the ongoing nature of entrepreneurship and career progression.

Furthermore, the BUSINESS IS A JOURNEY metaphors often mark a significant position or change in status. Example 1 illustrates the disadvantaged status of businesswomen in terms of public social recognition. In Example 2, the phrase "took off" denotes the remarkable moment of the businesswoman in the business arena. Interconnecting with the COMPETITION and ANIMAL source domains, the media depicts the businesswoman as successful and adventurous in her business operations.

Example 1: Whenever the pollsters quiz the German public as to which woman has made the greater public impact, the names are always from the same circles - they are either in political life, or they are artists and athletes, or possibly wives of politicians. Far behind, in last place, if at all, follow businesswomen (SCMP, 1980).

Example 2: At 42, she is the hottest name in Britain's cosmetics industry, the winning filly no punter in his right mind would have backed when she took off in Brighton nine years ago (SCMP, 1985).

4.1.4.2 Differences

In this study, the COMPETITION source domain is an overarching category that includes WAR and SPORT/GAME. The former is often associated with words aligned with the WAR frame, such as "fight," "battle," and "combat." The latter typically involves less aggressive words, such as challenge. The COMPETITION source domain, with a focus on SPORT/GAME, only occurred in the first decade.

Regarding the discursive effect, the SPORT/GAME frame attenuates the aggressiveness of business activities by portraying

them as a competition between female and male entrepreneurs (see Example 3). On the contrary, the WAR frame tends to foreground the exclusivity of business competition (see Example 4) or address the welfare issue of businesswomen (see Example 5).

Example 3: Besides managing restaurants and boutiques, which can be considered a woman's natural domain, they have become tenacious in challenging their male counterparts in industrial fields such as construction, machinery, furniture manufacturing for export, soap and detergent making, and even in computer science (SCMP, 1984).

Example 4: Jennifer d'Abo made history this month when she became the first woman to launch a takeover battle on the London stock market (SCMP, 1985).

Example 5: Now, English businesswoman Diana Newhofer has come up with a way to fight back. Combatting loneliness, discomfort, and danger while travelling alone, she says, requires teamwork (SCMP, 1995).

4.2 The functions of source domains

4.2.1 Describe the efforts and achievements of businesswomen

Our study found that BUILDING and JOURNEY source domains primarily emphasize businesswomen's efforts in developing management skills and advancing to higher positions. According to Lakoff and Johnson (2003), a metaphorical concept could entail a coherent system of metaphorical expressions. For instance, under the TIME IS MONEY conceptual framework, some expressions correspond to money (spend, invest), some to limited resources (use, use up), and others to valuable commodities (have, give). In this case, two metaphors are under the two subcategories of BUSINESS IS BUILDING: CAREER IS A LADDER and MANAGEMENT ROLE IS THE TOP OF THE LADDER.

Example 6: Women working their way up the corporate ladder suffer far more than their male colleagues from stress, nightmares, and depression, and are four times as likely to seek psychological help, a new survey concludes (SCMP, 1981).

Example 7: At 29, she had made it to the top of the computer consultancy firm she worked for as general manager (SCMP, 1987).

Examples 6 and 7 above adopt BUILDING as the masculine-oriented source domain to display the process of striving for career advancement as businesswomen. As examined in Zeng et al.'s (2021) study, the BUILDING source domain primarily specifies a building construction or signifies a stage change. In this BUILDING source domain, career development is metaphorically conceptualized as climbing a corporate ladder, with the top representing executive roles. This metaphor illustrates the hierarchical corporate structure in the business world and the ongoing process of advancing to higher positions.

Likewise, the BUSINESS IS JOURNEY metaphor also serves to describe the continuous nature of career development. Aligned with Zeng et al.'s (2021) findings, JOURNEY metaphors often feature the ongoing process toward a goal and the progress made toward achieving it. In Example 8, the feminine-oriented JOURNEY metaphor conveyed an implicit positive attitude toward Tina Ti's career transition from actress to renowned businesswoman. By conceptualizing the career as a JOURNEY, her career advancement is conceptualized as the goal. Meanwhile, the metaphorical keyword "come a long way" highlights her unremitting efforts in her professional pursuits.

Example 8: Tina Ti, who survived as a hostage in yesterday's fatal Victoria Peak robbery, has come a long way since the 1960s when she was a screen sex bomb in Hong Kong (SCMP, 1992).

4.2.2 Discuss the challenges businesswomen faced in their careers

PHYSICAL OBJECT and COMPETITION are two source domains that often refer to the barriers that businesswomen encounter. The metaphorical expressions reinforce the hardships that businesswomen face when navigating challenges and career opportunities. In Example 9, the invisible male domination in the executive council is conceptualized as a physical bastion that hinders women from entering the executive board.

Example 9: Only two women have managed to break through the male bastion of the Executive Council, while seven of the 46 Legislative Councillors are women (SCMP, 1987).

The WAR frame is dominant in the occurrences within the COMPETITION conceptual metaphor. Corresponding to Li et al.'s (2024) research, the

business media depicted businesswomen as fighters for overcoming societal and cognitive-level challenges using WAR metaphors. As Example 10 suggests, businesswomen are expected to carefully balance their work and family responsibilities. By framing the struggle as a WAR with the metaphor keyword "conflict," the news media portrays the work-life balance as hard to reconcile. Furthermore, the gender prejudice against female business leaders indicates the disadvantaged perception they face in the business world. In Example 11, the speaker, Dr. Lber-Schade, used the BUSINESS IS WAR metaphor to highlight the urgent need for businesswomen to break those misconceptions and prejudices. By referring to businesswomen as active participants, she emphasized the significance of women's initiative in addressing the cognitive-level challenges. These two examples demonstrate the application of metaphors as persuasive and emotive devices.

Example 10: Even the most committed businesswoman, however, can feel the conflict of work and home if she has two young daughters, as Emme has (SCMP, 1980).

Example 11: "Women executives in West Germany, as in most other Western countries, still have to fight against lingering misconceptions and prejudices," she complains (SCMP, 1980).

4.2.3 Attach traditional masculine or feminine traits to businesswomen

4.2.3.1 Traditional masculine traits

Metaphor analyses of the COMPETITION source domain suggest that it tends to focus on the aggressive feature of business practice and the adventurous leadership style (Koller, 2004a,b; Li et al., 2024). This study further verifies that the depiction of businesswomen with traditional masculine traits, including being ambitious and competent in business, is often achieved through the COMPETITION source domain. The following examples feature explicit metaphor keywords in the WAR frame. The keyword "vengeance" successfully conveys Mrs. Chow's ambition in the business competition and her desire to outperform her competitors (see Example 12). The news media also used the BUSINESS IS WAR metaphor strategically regarding the economic and social environment (see Example 13). In contrast to strict social constraints, female professionals in Korea are portrayed as self-conscious and independent in adapting to economic changes.

Example 12: As general manager of ATV, she has embarked on her job with a vengeance, hiring new staff and expanding the fare that the station has to offer (SCMP, 1989).

Example 13: Women are learning to survive in the swirling currents of Korea's rapid economic development, although the society itself is still weighed down by Confucian doctrines and imbued with the tacit sense of male supremacy (SCMP, 1984).

4.2.3.2 Traditional feminine traits

Echoing Koller (2004a), our study also finds that some unique metaphors appeared in the news about businesswomen. In the corpus, ANIMAL, PLANT, and PERSON are three source domains foregrounding prototypical femininity. The metaphor in Example 14, WOMEN ARE RESCUERS, highlights the feminine nature of the business activity while attenuating masculine traits.

Example 14: Marisa Bellisario, who became Italy's best-known woman business executive when she rescued the giant Italian telecommunications company from the brink of collapse, died on Thursday (SCMP, 1988).

Example 15: Dispensing with make-up and dressed plainly in a white shirt and black pants, the businesswoman said she reaped much more than money from her work (SCMP, 1995).

The PLANT source domain, which is perceived as more feminine, describes the traditional feminine trait less visibly. By conceptualizing WOMEN as GARDENERS, the business activity is defined as planting and sowing (see Example 15). Similarly, it is focused on the actions of businesswomen but downplays the fierce business competition.

4.3 Additional findings

Direct metaphors typically focus on reporting the achievements of an influential businesswoman or a group of businesswomen, and indirect metaphors often reflect the social realities that businesswomen face. The two indirect metaphors below use the PHYSICAL OBJECT and PLANT source domains. In Example 16, the description of Hong Kong was aligned with its essential status after the opening-up of China in 1978. As a bridge between the global

business and China, Hong Kong attracted business professionals with its unique economic system and friendly environment.

Example 16: Hong Kong is an important springboard into China for our products (SCMP, 1982).

Example 17: She added: "Hong Kong is a breeding ground for entrepreneurs because of its lack of government intervention [. . .] (SCMP, 1996)."

However, businesswomen's rights were not consistent with the demand of economic growth, triggering advocacy voices for gender equality. Some indirect metaphors are juxtaposed with contrastive devices to intensify the challenges faced by businesswomen (see Examples 18, 19). Using the BUSINESS IS A JOURNEY metaphor, the speaker indicates the necessity of reforming the maternity protection system in Hong Kong (see Example 18).

Example 18: During a study comparing maternity leave rights in Asian countries 10 years ago, Ms. Thaler found that Hong Kong was the only place in Asia without maternity protection. "Even Burma was ahead of us," she says (SCMP, 1987).

Example 19: They often do have to face a heavier burden of family responsibilities, too, and express negative emotions, compared with male colleagues (SCMP, 1989).

From the examples, indirect metaphors complement the direct metaphors in the business environment and common challenges faced by businesswomen. They form a nearly complementary relationship, supporting each other to present a more comprehensive metaphorical portrayal of female entrepreneurs.

5 Conclusion

In general, the metaphors in Hong Kong English news present a positive semantic prosody towards businesswomen. The news highlights the achievements of female business leaders while acknowledging the challenges they face in advancing their careers. The news adopts a combination of traditional masculine and feminine traits through metaphors to shape the image of businesswomen. Particularly, women were depicted as ambitious and competent in business operations, which was often associated with traditional masculinity. While also emphasizing traditional femininity, such as kindness and compassion.

The two sub-corpora demonstrated similar distribution across source domains, including COMPETITION, JOURNEY, MACHINE, BUILDING, and PHYSICAL OBJECT. In the first corpus, there were more metaphors, while in the second corpus, metaphors showed greater diversity in terms of their source domains.

To conclude, this study contributes to the analysis of metaphors in business-related news. It also extends the research on media in Hong Kong, which is an important outer circle region in the Three Circles Model of World Englishes.

Acknowledgments

This research was supported by the Undergraduate Research and Innovation Scheme (URIS) at The Hong Kong Polytechnic University (Project Number: P0053583) and The Hong Kong Polytechnic University Research Fund (Project Number: P0045314).

References

Kathleen Ahrens. 2019. First Lady, Secretary of State and Presidential Candidate: A comparative study of the role-dependent use of metaphor in politics. In *Variation in political metaphor*, pages 13–34. John Benjamins Publishing Company.

Kathleen Ahrens and Menghan Jiang. 2020. Source domain verification using corpus-based tools. *Metaphor and Symbol*, 35(1):43–55.

Kathleen Ahrens, Menghan Jiang, and Winnie Huiheng Zeng. 2021. BUILDING metaphors in Hong Kong policy addresses. *Metaphor in Language and Culture across World Englishes*, pages 105–128.

Kathleen Ahrens and Winnie Huiheng Zeng. 2021. Expressing concepts metaphorically in English editorials in Sinosphere. *Exploring the ecologies of world Englishes in the twenty-first century: Language, society and culture. Edinburgh University Press, UK*, pages 170–192.

Kathleen Ahrens, Winnie Huiheng Zeng, Christian Burgers, and Chu-Ren Huang. 2024. Metaphor and gender: Are words associated with source domains perceived in a gendered way? *Linguistics Vanguard*, 10(1):711–720.

Laurence Anthony. 2024. Antconc (version 4.3.1) [computer software].

Christian Burgers. 2016. Conceptualizing change in communication through metaphor. *Journal of Communication*, 66(2):250–265.

Jonathan Charteris-Black. 2004. *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan.

Jonathan Charteris-Black. 2005. *Politicians and rhetoric: The persuasive power of metaphor*. Palgrave Macmillan.

Jonathan Charteris-Black. 2009. Metaphor and gender in British parliamentary debates. In *Politics, gender and conceptual metaphors*, pages 139–165. Springer.

Priscilla Pue Ho Chu. 2004. *The making of women entrepreneurs in Hong Kong*. Hong Kong University Press.

Financial Services and the Treasury Bureau. 2025. Executive summary. Government Report. Accessed: 2025-01-01.

Braj B Kachru. 1990. World Englishes and applied linguistics. *World Englishes*, 9(1):3–20.

Veronika Koller. 2004a. Businesswomen and war metaphors: 'Possessive, jealous and pugnacious'? *Journal of Sociolinguistics*, 8(1):3–22.

Veronika Koller. 2004b. *Metaphor and gender in business media discourse: A critical cognitive study*. Springer.

George Lakoff and Mark Johnson. 2003. *Metaphors we live by*. University of Chicago press.

Yanlin Li, Jing Chen, Kathleen Ahrens, and Chu-Ren Huang. 2024. The evolving use of WAR metaphors in businesswomen-focused media discourse. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 1377–1386.

Adelyn Lim. 2015. *Transnational feminism and women's movements in post-1997 Hong Kong: Solidarity beyond the state*. Hong Kong University Press.

Gill Philip. 2009. Non una donna in politica, ma una donna politica: Women's political language in an Italian context. In *Politics, gender and conceptual metaphors*, pages 83–111. Springer.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010a. Metaphor in usage.

Gerard J Steen, Aletta G Dorst, Tina Krennmayr, Anna A Kaal, and J Berenike Herrmann. 2010b. A method for linguistic metaphor identification.

Huiheng Zeng, Dennis Tay, and Kathleen Ahrens. 2020. A multifactorial analysis of metaphors in political discourse: Gendered influence in Hong Kong political speeches. *Metaphor and the Social World*, 10(1):141–168.

Winnie Huiheng Zeng, Christian Burgers, and Kathleen Ahrens. 2021. Framing metaphor use over time: 'Free Economy' metaphors in Hong Kong political discourse (1997–2017). *Lingua*, 252:102955.

Analysis of the Correlation Between Theory of Mind and Dialogue Ability to Identify Essential ToM for Dialogue Systems

Haruhisa Iseno[1,2], Atsumoto Ohashi[1,2], Tetsuji Ogawa[3], Shinnosuke Takamichi[4], Ryuichiro Higashinaka[1,2]

[1]Graduate School of Informatics, Nagoya University, [2]NII LLMC,
[3]Department of Communications and Computer Engineering, Waseda University,
[4]Department of Information and Computer Science, Keio University

{iseno.haruhisa.h4@s.mail, ohashi.atsumoto.c0@s.mail,higashinaka@i}.nagoya-u.ac.jp

ogawa.tetsuji@waseda.jp, shinnosuke_takamichi@keio.jp

Abstract

In large language models (LLMs), improvements in theory of mind (ToM), which is the ability to infer others' mental states, are expected to enhance dialogue performance. However, the quantitative verification of this relationship remains insufficient. Therefore, this study evaluates the performance of seven high-performing LLMs across three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and six dialogue tasks to verify the correlation between ToM and dialogue performances. Our findings revealed a fundamental correlation between ToM and dialogue performance, though significant differences emerged depending on the ToM aspects examined. Specifically, we observed high correlations with dialogue performance for both ToM evaluated in conversational formats and ToM assessed with questions directly asking about beliefs. Additionally, ToM in situations involving conflicting beliefs between agents strongly correlates with dialogue performance. Furthermore, stable correlations were observed between first-order ToM and dialogue capabilities. These findings provide crucial guidelines for developing dialogue systems with human-like dialogue capabilities.

1 Introduction

Dialogue systems based on large language models (LLMs) have recently demonstrated remarkable performance improvements across diverse dialogue tasks (OpenAI et al., 2023; Yi et al., 2024). To achieve more human-like advanced dialogue capabilities, it is essential to enhance not only language processing but also the ability to understand and reason about the mental states of others, i.e., theory of mind (ToM).

Numerous benchmarks have been proposed to evaluate ToM in LLMs, as improvements in dialogue capabilities through enhanced ToM are believed to be vital (Le et al., 2019; Gandhi et al., 2023; Wu et al., 2023; Kim et al., 2023; Chen et al., 2024; Shinoda et al., 2025). These benchmarks evaluate ToM by presenting LLMs with stories or dialogue histories and conducting question-answering tasks regarding the mental states of the people involved, such as their beliefs and intentions. Despite such efforts being undertaken, the relationship between LLMs' ToM benchmark performance and dialogue performance has not been quantitatively verified, and it remains unclear whether performance improvements in ToM benchmarks lead to improvements in dialogue performance.

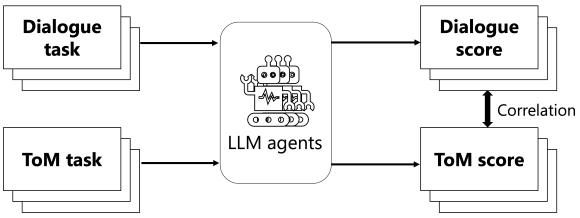


Figure 1: Experimental framework for analyzing correlations between dialogue and ToM task performances in LLMs.

Therefore, this study quantitatively examines the extent to which existing ToM benchmarks accurately capture the ToM required for dialogue. Specifically, we evaluated the performance of seven state-of-the-art LLMs on three different ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and six dialogue tasks (Taboo, Wordle, Drawing, Reference Game, Private & Shared, and MutualFriends), and systematically investigated the correlation between ToM and dialogue performance.

The contributions of this study are as follows.

- We developed a framework for analyzing the relationship between LLMs' ToM and dialogue performance, and verified their relation-

ship through correlation analysis of multiple ToM benchmarks and dialogue tasks (Fig. 1).

- We found strong correlations between ToM performance, particularly when evaluated in conversational formats or through questions directly probing beliefs, and overall dialogue performance. Moreover, ToM in scenarios where others hold beliefs that differ from one's own showed a strong correlation with dialogue performance.

- We observed stable correlations between first-order ToM and dialogue capabilities; however, correlations decreased markedly for second- and higher-order ToM, suggesting that the current dialogue tasks may be insufficient for capturing the relevance of higher-order ToM abilities.

2 Related Work

In this section, we review the ToM benchmarks and recent evaluations of dialogue system performance.

2.1 ToM Benchmarks

A variety of benchmarks for evaluating ToM from diverse perspectives have been proposed (Le et al., 2019; Ma et al.; Gandhi et al., 2023; Wu et al., 2023; Chen et al., 2024; Xu et al., 2024). Most of these benchmarks measure the accuracy for problems regarding characters' mental states, such as beliefs and intentions, based on story contexts, as exemplified by the Sally-Anne task (Baron-Cohen et al., 1985). For example, ToMBench (Chen et al., 2024) comprehensively analyzes ToM using 20 types of diverse story-format problems.

ToM benchmarks using conversations as contexts have also been proposed. Benchmarks such as FANToM (Kim et al., 2023), NegotiationToM (Chan et al., 2024), and ToMATO (Shinoda et al., 2025) adopt ToM evaluation with conversations as the context and attempt to construct ToM evaluation environments that approximate the actual interaction settings.

Most benchmarks evaluate first-order ToM (estimating a person's mental state) and second-order ToM (estimating a person's mental state about another person's mental state). An exception is HiToM (Wu et al., 2023), which adopts a design that systematically evaluates higher-order ToM (up to the fourth order) in addition to the conventional first- and second-order ToM, enabling the measurement of more complex ToM.

Based on these benchmarks, ongoing discussions have focused on whether LLMs exhibit ToM. Kosinski (2023) reported that GPT-3.5 demonstrated a performance equivalent to children aged 7–10 years on ToM tasks, suggesting the possibility that ToM spontaneously emerged in LLMs. Conversely, Ullman (2023) and Shapira et al. (2024) showed that the accuracy rate of LLMs decreased significantly with only minor modifications to ToM benchmarks, arguing that LLMs have not developed ToM but rather solved ToM problems by relying on superficial pattern matching. In the current study, we assume that there is some relationship between LLMs' ToM and dialogue capabilities and investigate which aspects of ToM correlate and to what extent.

2.2 Evaluation of Dialogue System Performance

Dialogues are broadly classified as task-oriented and non-task-oriented (McTear, 2022). While the evaluation methods differ for each type, this study focuses on task-oriented dialogue, which allows for easier quantitative evaluation to conduct correlation analysis.

Benchmarks such as MultiWOZ (Budzianowski et al., 2018) and schema-guided dialogue (Rastogi et al., 2020) measure dialogue system performance by utilizing dialogue state tracking accuracy, response generation quality, and task success in specific tasks, such as restaurant reservations and hotel searches. Numerous collaborative dialogue tasks requiring conversational grounding for task completion have also been proposed (He et al., 2017; Udagawa and Aizawa, 2019; Kim et al., 2019; Bara et al., 2021). These tasks measure the ability to establish common ground (Clark, 1996) with interlocutors through multi-turn dialogues.

Frameworks for automatic evaluation of the dialogue capabilities of LLMs using dialogue between LLMs have recently been developed. Chalamalasetti et al. (2023) proposed Clembench, a framework that evaluates the dialogue capabilities of LLMs through LLM-to-LLM interactions in a game format, enabling a comprehensive automatic evaluation by quantitatively measuring LLM performance across multiple dialogue tasks.

In the current study, we focus on dialogue tasks that require conversational grounding, where ToM is strongly involved, and analyze the relationship

between dialogue performance and ToM. We use Clembench to evaluate the dialogue performance.

3 Approach

We conduct a correlation analysis between LLMs' ToM and dialogue performance to verify their relationship. Correlation analysis has frequently been utilized to examine whether evaluation metrics can appropriately measure the intended targets. For example, in machine translation, Papineni et al. (2002) demonstrated a strong correlation between BLEU scores and human evaluations. The confirmation of this correlation led to the understanding that improvements in BLEU scores directly lead to the generation of better-quality translations for humans.

Although BLEU has sometimes been used for dialogue evaluation, Liu et al. (2016) demonstrated that the correlation between human dialogue evaluation and BLEU-based dialogue evaluation is not significantly high. Therefore, other evaluation measures have increasingly been used for dialogue evaluation (Zhang et al., 2019; Mehri and Eskenazi, 2020).

Since evaluation metrics must appropriately measure what they are intended to, and ToM is regarded as a fundamental ability underpinning human social interaction (Baron-Cohen et al., 1985; Frith, 1994), it is important to examine whether the performance measured by ToM benchmarks is related to actual dialogue capabilities. Therefore, this study analyzes the correlation between ToM and dialogue capabilities. As shown in Fig. 1, we use m ToM benchmarks and l dialogue tasks on n state-of-the-art LLMs and calculate the correlation coefficients between LLMs' ToM benchmark accuracy rates and dialogue task scores. We verify whether correlations exist between ToM and dialogue capabilities by conducting a correlation analysis individually for each aspect of ToM and determine which aspects exhibit stronger correlations.

The following sections describe the selection of the ToM benchmarks and dialogue tasks addressed in this study.

3.1 Selection of ToM Benchmarks

Although evaluating the overall ToM performance is important, evaluating the performance across various aspects of ToM is necessary to comprehensively assess the correlations with dialogue performance. Therefore, this study selects benchmarks by focusing on the following aspects that can be particularly relevant to dialogue performance.

The first aspect is the context format. ToM benchmarks include tasks that estimate characters' beliefs from narrative-format contexts, such as the Sally-Anne task (Wimmer and Perner, 1983; Baron-Cohen et al., 1985), and tasks that infer characters' mental states from conversational texts. Even within narrative formats, there are settings that include dialogue between characters within the story and other settings that do not. To analyze the impact of these differences in context format on correlations with dialogue capabilities, we select benchmarks to comprehensively cover cases where contexts are in narrative and dialogue formats, and cases where narratives include dialogue and those that do not.

The second aspect is the mental state targeted for inference. In ToM tasks, ToM that infers belief states different from those of the reasoner (e.g., a situation where one knows that the cookies are in box B, but the other person who does not know this believes they are in box A) are important (Quesque and Rossetti, 2020; Shinoda et al., 2025), as in false belief tasks (Wimmer and Perner, 1983). However, it is unclear which ability—estimating mental states different from one's own or estimating the same mental states—is more strongly related to actual dialogue capabilities. Therefore, to investigate this relationship, we select benchmarks that include both types of inference.

The third aspect is the question format. ToM benchmarks include diverse question formats that directly ask about beliefs, such as "Where does A think X is?", and formats that require identifying knowledge holders, such as "Who knows where X is?" Although these questions necessitate different response content, correctly answering them requires the same ToM: "Where does a person think X is?" To investigate whether differences in question format affect correlations with dialogue capabilities even when the required ToM is the same, we select benchmarks that include different question formats.

The final aspect is the order of beliefs. First-order beliefs are a person's own beliefs, such as "A thinks X." Second-order beliefs are beliefs about others' beliefs, such as "B thinks that A thinks Y." Furthermore, some ToM tasks measure ToM for even higher-order beliefs, such as third- and fourth-order beliefs. To analyze the correlations between ToM for each of these first- to fourth-order beliefs

and dialogue capabilities, we select benchmarks that include ToM for beliefs of multiple orders.

3.2 Selection of Dialogue Tasks

To verify the correlation between ToM and dialogue performance, we select task-oriented dialogue tasks in which task achievement can be quantified with clear evaluation metrics and conversational grounding is required for task completion; such grounding is believed to be strongly related to ToM.

4 Experiments

We conducted experiments to investigate the correlation between ToM and dialogue performance. First, we selected ToM and dialogue tasks to be used in the correlation analysis. We then used seven state-of-the-art LLMs (GPT4.1, Gemini2.5-Flash, Claude4-Sonnet, Grok4, Llama3.3-70B, Qwen3-32B, and Mistral-Small) to answer ToM tasks and perform dialogue tasks. Subsequently, we examined the correlation between ToM and dialogue performance.

4.1 ToM Tasks

We selected three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) that comprehensively encompass the differences in context format, mental states targeted for inference, question format, and order of beliefs, enabling a correlation analysis from each perspective. Examples of these benchmarks are provided in Appendix A, and their details are as follows.

ToMBench A benchmark that estimates characters' belief states using stories as context. A distinctive feature of ToMBench is its comprehensive evaluation of ToM in 20 diverse story contexts. These stories include various tasks, ranging from those based on the classic Sally-Anne task to mental state estimations in more complex social situations. Each story is structured with questions about the characters' mental states, enabling the measurement of whether LLMs can accurately estimate the characters' beliefs, desires, intentions, and other mental states based on narrative contexts. In this dataset, most problems involve first-order ToM, while one of the 20 stories, the false-belief task, involves ToM for both first- and second-order beliefs.

FANToM A benchmark that estimates specific speakers' belief states using multiparty conversations as context. A distinctive feature of this benchmark is that dynamic information asymmetry arises when characters leave and rejoin the dialogue. Since conversations continue even when speakers are absent, the known information differs among characters, resulting in a structure in which each character develops a different belief state. FANToM has three subtasks: (1) BeliefQA, which are tasks that directly ask about characters' belief states; (2) InfoAccessibilityQA (InfoQA), which are tasks that enumerate people who possess specific information; and (3) AnswerabilityQA (AnsQA), which are tasks that enumerate people who can correctly answer BeliefQA questions. BeliefQA includes ToM problems for first- and second-order beliefs and is classified into two conditions: accessible and inaccessible. In the accessible condition, questions are asked about belief states in which information is shared among characters, whereas in the inaccessible condition, questions are asked about belief states in which information is not shared.

Hi-ToM A benchmark that evaluates higher-order ToM. Its distinctive feature is requiring ToM up to the fourth order. It adopts narratives that extend the Sally-Anne task as context, with settings in which multiple characters enter and exit rooms while moving objects. Each story includes five questions that gradually increase in complexity, from "Where is X?" (0th-order) to "Where does A think B thinks C thinks D thinks X is?" (4th-order). Additionally, there are two types of settings: those that include communication between characters in the story (Tell condition) and those that do not (No_Tell condition).

We used the accuracy rates of the LLMs on these three ToM benchmarks as indicators of LLMs' ToM performance.

To analyze the effects of the differences in context formats, we utilized ToMBench and Hi-ToM as representative narrative-format benchmarks and FANToM as a representative dialogue-format benchmark. Furthermore, we analyzed the changes in the correlations caused by the presence or absence of communication elements in narrative-format benchmarks by comparing correlations under the two experimental conditions of Hi-ToM's Tell and No_Tell conditions.

To analyze the effects of the differences in mental states targeted for inference, we used FANToM's accessible and inaccessible conditions. The

Task	Task description	Evaluation metrics
Taboo	A vocabulary explanation task in which one of two players explains a target word without using specific forbidden words, while the other player guesses the target word from the explanation.	A score comprising the accuracy rate of finding target words and the number of dialogue turns until success.
Wordle	A deduction task in which players identify a five-letter English word within six attempts. After each attempt, feedback is provided indicating whether correct letters are in the correct positions.	A score comprising the accuracy rate of identifying five-letter English words and the number of dialogue turns until success.
Drawing	A task in which one of two players describes a virtual image composed of a 5×5 character grid using only language, while the other player reconstructs the original character string.	F1 score between the reconstructed character string and the original character string.
Reference Game	A reference resolution task in which two players identify one common image from among three virtual images composed of 5×5 character grids.	Accuracy rate of selecting the correct image.
Private & Shared	A task in which a questioner and answerer share information through dialogue, and the answerer estimates which information the questioner knows and does not know at each point in the dialogue.	An integrated score combining information sharing success rate and accuracy rate of estimating the partner's belief state.
MutualFriends	A task in which two speakers are each given different friend lists and find mutual friends through dialogue.	Accuracy rate of correctly identifying common friends.

Table 1: Task descriptions and evaluation metrics of dialogue tasks used in this study.

accessible condition requires the inference of mental states identical to that of the reasoner, whereas the inaccessible condition requires the inference of mental states that are different from those of the reasoner. This enables the analysis of changes in correlation caused by the differences in targeted mental states (inference of mental states identical to one's own versus inference of mental states different from one's own).

To analyze the effects of the differences in question formats, we used FANToM's three subtasks (BeliefQA, InfoAccessibilityQA, and AnswerabilityQA). As these questions ask for the same ToM inference through different question formats, we analyzed the effects of the differences in the question formats on the correlations.

To analyze the effects of the differences in order of beliefs, we utilized Hi-ToM's first- to fourth-order ToM tasks and first- and second-order belief estimation tasks from ToMBench and FANToM. We analyzed the effects of the differences in the order on the correlations by comparing the correlations for each order.

4.2 Dialogue Tasks

To evaluate the dialogue capabilities of LLMs, we implemented five types of text games conducted through dialogue (Taboo, Wordle, Drawing, Reference Game, Private & Shared). These five tasks are also used by Clembench (Chalamalasetti et al., 2023). In addition, we used the MutualFriends (He et al., 2017) task as the sixth dialogue task. All of these tasks have quantitative evaluation metrics and are tasks in which ToM is believed to be important for task completion. The description and evaluation metrics of each dialogue task are provided in Table 1.

All tasks other than Wordle were conducted through dialogue between the same LLMs. In contrast, since Wordle can proceed with only simple feedback from the user, dialogue tasks were conducted through dialogue between LLMs and a rule-based user simulator. Subsequently, based on these dialogues, the LLM performance in each dialogue task was scored on a 0–100 scale to indicate the dialogue performance. For MutualFriends, the scores were calculated using only dialogue success rates, and for all other tasks, the scoring method defined by Clembench (Chalamalasetti et al., 2023) was applied. Furthermore, the average score of the six tasks was calculated as "Average" and utilized to indicate the overall dialogue capabilities of the LLMs.

	ToMBench	FANToM	Hi-ToM
Drawing	0.52	**0.68**	0.15
Private & Shared	0.68	**0.76**	0.73
Reference Game	0.65	**0.91**	0.50
Taboo	0.69	**0.91**	0.61
Wordle	0.66	**0.89**	0.26
MutualFriends	0.37	**0.54**	0.31
Average	0.66	**0.85**	0.45

Table 2: Pearson correlation coefficients between ToM tasks and each dialogue task. Bold values indicate the strongest correlations for each dialogue task.

	No_Tell	Tell
Drawing	0.15	0.12
Private & Shared	0.72	0.61
Reference Game	0.43	0.52
Taboo	0.58	0.56
Wordle	0.29	0.16
MutualFriends	0.27	0.31
Average	0.43	0.39

Table 3: Pearson correlation coefficients under settings where narrative tasks include interactions (Tell) and settings where they do not (No_Tell).

4.3 Experiment Procedure

To execute the ToM tasks, we presented the LLMs with story or dialogue texts as context and asked them to answer questions about the characters' mental states in a multiple-choice format. For Hi-ToM, we used the existing question-answering prompts included in the dataset, while for ToMBench and FANToM, we designed new prompts for this study (see Appendix B).

To execute the dialogue tasks, we controlled the LLMs using the existing prompts provided by the benchmark when the five tasks included by Clembench (Taboo, Wordle, Drawing, Reference Game, Private & Shared) were performed. We used the prompts designed for this study when the MutualFriends task was performed (see Appendix C).

We quantitatively evaluated the ToM and dialogue performance of each model using the above-mentioned methods and calculated the Pearson correlation coefficients between them. Given our sample size of seven state-of-the-art LLMs, correlation coefficients of ≥ 0.670 indicate a significant trend ($p < 0.1$), ≥ 0.755 indicate statistical significance ($p < 0.05$), and ≥ 0.875 indicate high statistical significance ($p < 0.01$). However, with such a limited sample size, individual correlation coefficients may lack statistical robustness. Therefore, rather than relying solely on the statistical significance of individual correlations, we prioritized identifying consistent overall trends that emerged across multiple tasks and conditions, using the correlation coefficients as a reference for interpreting the strength and direction of observed relationships.

4.4 Results

This section presents the experimental results of the correlations between ToM and dialogue performance based on the four perspectives mentioned in Section 3.1: (1) context format, (2) mental states targeted for inference, (3) question format, and (4) order.

4.4.1 Effect of Context Type

Table 2 lists the correlation coefficients between the overall accuracy rates of the three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and the success rates of each dialogue task. The results show a clear trend in the correlations with dialogue tasks. In almost all dialogue tasks, FANToM, which uses conversations as context, exhibits higher correlations than the narrative-format ToMBench and Hi-ToM. For the Private & Shared task, the correlation coefficients are nearly equivalent for the three ToM benchmarks. This is likely because this task is originally designed to perform question-answering tasks with content similar to ToM tasks, resulting in a high structural similarity with ToM benchmarks.

Table 3 presents the results of a comparative analysis of the correlations with dialogue tasks in Hi-ToM tasks under settings where communication occurs between characters (Tell condition) and where communication does not occur (No_Tell condition). The results indicate no significant changes in the correlation caused by the presence or absence of communication elements. The results reveal fundamental limitations of narrative-format contexts. Specifically, even when introducing a small number of conversational elements within stories, the same improvement in the correlation with dialogue capabilities as when using conversations as the context is not observed. We therefore consider the inclusion of characters' utterances in stories insufficient to bring about essential improvements in dialogue capability.

4.4.2 Effect of Mental State Type

We analyzed the extent to which ToM toward others who hold belief states different from their own correlates with dialogue performance. Table 4 presents the results of classifying the BeliefQA task within

	Accessible	Inaccessible
Drawing	−0.74	**0.71**
Private & Shared	−0.85	**0.96**
Reference Game	−0.64	**0.82**
Taboo	−0.72	**0.88**
Wordle	−0.77	**0.82**
MutualFriends	−0.84	**0.75**
Average	−0.85	**0.91**

Table 4: Pearson correlation coefficients between each of the accessible and inaccessible conditions in FANToM and dialogue tasks. Bold values indicate higher correlations for each dialogue task.

Model	Accessible	Inaccessible
GPT4.1	80.26	72.10
Gemini2.5-Flash	79.71	78.05
Claude4-Sonnet	62.34	80.16
Grok4	61.61	80.06
Llama3.3-70B	77.15	62.54
Mistral-Small	92.14	28.90
Qwen3-32B	93.24	20.24

Table 5: Task accuracy rates for each LLM on FANToM.

the FANToM benchmark into accessible and inaccessible conditions and comparing the correlations with dialogue tasks. The results show that the inaccessible condition exhibits strong positive correlations across all dialogue tasks, whereas the accessible condition consistently exhibits strong negative correlations. This result indicates that the ability to estimate belief states that differ from one's own is important for high dialogue capabilities. The ability to correctly grasp situations in which others hold beliefs that are different from one's own and predict their actions and reactions on the basis of those beliefs is thus considered a crucial capability that determines success in an actual dialogue.

For a more detailed analysis of the causes of the negative correlations observed in the accessible condition, we individually compared each LLM performance under accessible/inaccessible conditions. Table 5 lists the accuracy rates of each LLM on the FANToM benchmark under accessible/inaccessible conditions. As we can see, Mistral and Qwen3 exhibit extremely high accuracy rates in the accessible condition, but low accuracy rates in the inaccessible condition. Models that scored highly in the accessible condition tend to answer assuming that all information is shared without considering others' perspectives and might have been conducting inference based on the incorrect assumption that "all information is equally accessible to all people." The negative correlation is also caused by Grok4 and Claude4, which performed well in the dialogue tasks but exhibited relatively low performance in the 60-point range in the accessible condition.

	BeliefQ	AnsQ	InfoQ
Drawing	<u>0.65</u>	0.41	**0.67**
Private & Shared	**0.94**	<u>0.45</u>	0.44
Reference Game	**0.83**	0.78	<u>0.81</u>
Taboo	**0.88**	<u>0.75</u>	<u>0.75</u>
Wordle	<u>0.78</u>	0.72	**0.86**
MutualFriends	**0.67**	0.26	<u>0.36</u>
Average	**0.87**	0.59	<u>0.72</u>

Table 6: Pearson correlation coefficients for each FANToM subtask. Bold values indicate the strongest correlations for each dialogue task, and underlined values represent the second strongest correlations.

4.4.3 Effect of Question Format

For the correlations with dialogue capabilities owing to differences in question format, we analyzed the relationship between FANToM's three subtasks (BeliefQ, AnsQ, and InfoQ) and dialogue task performance. As presented in Table 6, BeliefQ consistently exhibits higher correlations than the other two subtasks. These results clearly indicate that although all three subtasks of FANToM require the same ToM inference, the correlations with dialogue capabilities differ significantly depending on the question format. Although formats that ask for direct belief estimation, such as BeliefQ, exhibit high correlations with dialogue tasks, formats that ask for applications of inference results, such as AnsQ and InfoQ, exhibit low correlations despite dealing with the same ToM inference. This result indicates that, even for problems requiring the same ToM inference, the obtained evaluation varies depending on the question format. In particular, the results clearly show that problem formats that perform direct estimation of beliefs are more reliable indicators of dialogue capabilities.

4.4.4 Effect of Reasoning Order

To examine how the order of inference in ToM affects dialogue performance, we conducted a correlation analysis between performance by order of beliefs on three ToM benchmarks (Hi-ToM, ToMBench, FANToM) and dialogue task performance.

Tables 7 and 8 present the correlation coefficients between ToM up to the fourth order for each

	0th	1st	2nd	3rd	4th
Drawing	0.57	**0.83**	−0.33	−0.44	−0.41
Private & Shared	**0.99**	0.74	0.28	0.12	0.29
Reference Game	**0.80**	0.62	0.02	−0.03	0.15
Taboo	**0.86**	0.57	0.16	0.08	0.33
Wordle	**0.70**	0.66	−0.27	−0.38	−0.11
MutualFriends	**0.75**	0.72	−0.15	−0.28	−0.24
Average	**0.84**	0.80	−0.08	−0.20	−0.05

Table 7: Pearson correlation coefficients between ToM by order in Hi-ToM and dialogue tasks. Bold values indicate the strongest correlations for each dialogue task, and underlined values represent the second strongest correlations.

	ToMBench		FANToM	
	1st	2nd	1st	2nd
Drawing	**0.63**	−0.18	**0.73**	0.48
Private & Shared	**0.73**	0.18	**0.94**	0.88
Reference Game	**0.80**	−0.06	**0.83**	0.77
Taboo	**0.84**	0.15	**0.88**	0.83
Wordle	**0.81**	−0.00	**0.83**	0.66
MutualFriends	**0.44**	−0.07	**0.72**	0.54
Average	**0.78**	−0.02	**0.91**	0.75

Table 8: Pearson correlation coefficients between ToM by order in ToMBench and FANToM and dialogue tasks. Bold values indicate the strongest correlations for each dialogue task.

benchmark and dialogue task performance. The results show that the correlations with dialogue capabilities differ between lower-order ToM (0th- and 1st-order) and higher-order ToM. Across all benchmarks, the first-order ToM exhibits stable positive correlations with dialogue tasks. In particular, the high correlation with first-order belief estimation in FANToM indicates that first-order ToM evaluation in a conversational format is strongly related to dialogue capabilities. In contrast, second-order and higher-order ToM exhibit markedly low correlations, with negative or no correlations observed in many cases. In Hi-ToM, correlations decrease significantly from second-order ToM onward. In ToMBench and FANToM, second-order ToM also exhibits lower correlations compared to first-order ToM.

These results do not mean that second- and higher-order ToM are unnecessary in dialogue. Although both are important cognitive abilities in complex interactions, within the scope of the collaborative tasks addressed in this study, the recognition of world states (0th-order ToM) and estimation of others' beliefs and intentions (first-order ToM) likely played more important roles. In an actual human dialogue, different orders of ToM inference are dynamically utilized depending on the situation. To achieve fundamental improvements in the LLMs' ToM in the future, the introduction of more advanced tasks that require higher-order ToM inference (Wang et al., 2019; Kano et al., 2024) will be necessary.

5 Conclusion

This study conducted a comprehensive correlation analysis of the relationship between ToM and dialogue performance in LLMs using three ToM benchmarks and six dialogue tasks. We confirmed a fundamental correlation between ToM and dialogue performance, though significant differences emerged depending on the ToM aspects examined. Specifically, we observed high correlations with dialogue performance for both ToM evaluated in conversational formats and ToM assessed with questions directly asking about beliefs. Additionally, ToM in situations involving conflicting beliefs between agents strongly correlates with dialogue performance. Our findings indicate that dialogue tasks requiring higher-order ToM inferences are crucial for a more comprehensive evaluation of the dialogue capabilities of LLMs. This study provides the first systematic empirical analysis of the relationship between ToM and dialogue performance, with the results serving as valuable guidelines for developing dialogue systems with human-like dialogue capabilities.

This study has several limitations. First, the LLMs evaluated are limited to seven types, and it is unclear whether similar trends would be obtained when more diverse models and architectures are included. Second, Pearson correlation has several assumptions, and it is possible that the current experimental setting may not fully satisfy these assumptions. Therefore, it will be necessary to analyze the correlations in more detail by using other indicators (e.g., Spearman correlation) in the future. Third, the findings regarding the correlation between second- or higher-order ToM and dialogue ability are limited. The dialogue tasks utilized in this study are not considered to include tasks where second- or higher-order belief estimation abilities directly contribute to task achievement, and this constraint in task design is likely one of the reasons for the weak correlations observed in higher-order ToM. In the future, we plan to introduce dialogue tasks that require second- or higher-order belief

estimation. Additionally, in the Hi-ToM used in this study, the accuracy rate of tasks by humans has not been measured, and human performance on higher-order ToM reasoning tasks remains unclear. In the future, by measuring the accuracy rate of humans, we will be able to more accurately position the relationship between ToM for second- or higher-order beliefs and dialogue ability. Fourth, the dialogue tasks used in this study are limited to the five game-style tasks included in Clembench and the MutualFriends task, and it remains unclear whether the findings would generalize to tasks covering broader domains (such as MultiWOZ) or to non-task-oriented dialogues (such as casual conversation). Finally, it will also be necessary to determine how dialogue models can be systematically improved using the empirical insights gained from this study.

Acknowledgments

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology.

References

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Uta Frith. 1994. Autism and theory of mind in everyday life. *Social Development*, 3(2):108–124.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *In Proceedings of the 37th International Conference on Neural Information Processing Systems*, 36:13518–13529.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing, pages 5872–5877.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 15–26.

Michael McTear. 2022. *Conversational AI: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in LLM-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

A Example Problems in ToM benchmarks

Table 9 presents example problems of the ToM benchmarks used in the study.

BenchMark	Context	QA
ToMBench (Chen et al., 2024)	Li Lei and Han Meimei are wandering in the living room, they see the cabinet, box and handbag, they find a sweatshirt in the cabinet, Han Meimei leaves the living room, Li Lei moves the sweatshirt to the handbag.	Question: Where is the sweatshirt now? (A) Briefcase (B) Box (C) Cabinet **(D) Handbag** Question: After Han Meimei returns to the living room, where does Li Lei think Han Meimei looks for the sweatshirt? (A) Box (B) Wardrobe (C) Handbag **(D) Cabinet**
FANTOM (Kim et al., 2023)	Kailey: Hey guys, I'll go grab a coffee. Sally: See you, Kailey! Hey Linda, did you get a dog? Linda: Yeah, I got a golden retriever. She's so adorable. ... Kailey: I'm back, what are you guys discussing now? Sally: Linda was just telling us that her dog can do special moves! Linda: Yeah, she can stand on her feet and do a dance move to music!	BeliefQA: What breed would Kailey think Linda's dog is? (A) Kailey believes Linda has a golden retriever. **(B) Kailey does not know the breed.** AnswerabilityQA: Who knows the correct answer to "What breed would Kailey think Linda's dog is"? **Linda, David, Sally** InfoAccessibilityQA: Who knows about "Linda has a golden retriever"? **Linda, David, Sally**
HI TOM (Wu et al., 2023)	William, Jack, Charlotte, Noah and Hannah entered the hall. Noah saw a monkey. The carrot is in the red_basket William exited the hall. ... Jack exited the hall. Charlotte exited the hall. Noah moved the carrot to the green_envelope. Noah exited the hall. Hannah moved the carrot to the red_basket. Hannah exited the hall. William, Jack, Charlotte, Noah and Hannah entered the waiting_room. Charlotte publicly claimed that carrot is in the green_envelope. Hannah privately told Charlotte that the carrot is in the blue_container.	Question-order0: Where is the carrot really? (A) green_envelope, **(B) red_basket**, ⋯ Question-order1: Where does William really think the carrot is? **(A) green_envelope**, (B) red_basket, ⋯ Question-order2: Where does Hannah think William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯ Question-order3: Where does Jack think Hannah thinks William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯ Question-order4: Where does Charlotte think Jack thinks Hannah thinks William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯

Table 9: Examples from the three ToM benchmarks addressed in this study. FANToM has three subtasks: BeliefQA, which directly estimates beliefs; AnswerabilityQA, which asks about the answerability of the questions; and InfoAccessibilityQA, which asks about people who know the information. In Hi-ToM, questions corresponding to 0th- to 4th-order ToM inference are set from Question-order 0 to 4. Bold portions in the QA items indicate the correct answers for each question.

B ToM Task Prompts

This section presents the prompts used by LLMs to solve ToMBench and FANToM. The prompt for solving ToMBench is as follows. `{context}` contains the context that serves as the basis for inference, `{question}` contains questions about characters' mental states, and `{a}`, `{b}`, `{c}`, and `{d}` are the answer choices.

```
Please read the passage and the question I will ask. Choose the correct answer from options A, B, C, and D.
{context}
{question}
A: {a}
B: {b}
C: {c}
D: {d}
Please answer with the letter of the option that you think is correct and do not output anything other than a single letter.
```

The following are the prompts used to solve FANToM, which is used for BeliefQ, InfoQ, and AnsQ. `{context}` contains the dialogue text that serves as the basis for inference, and `{BeliefQ}`, `{InfoQ}`, and `{AnsQ}` contain question texts defined for each task by the dataset. Additionally, `{factQ}` and `{factA}` contain the facts asked in BeliefQ, and `{candidates}` lists the names of the characters.

```
{context}
Question: {BeliefQ}
{ans_a}
{ans_b}
Please choose either a or b as the correct answer. Output only a or b.
```

```
{context}
Information: {factQ} {factA}
Question: {InfoQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

```
{context}
Target: {factQ}
Question: {AnsQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

C Dialogue Task Prompt

The prompts used in the MutualFriends task are as follows. Among these, `{subject}` and `{friends}` contain a list of friends given to the player, and `{history}` contains the dialogue history.

```
You are a smart cooperative agent named Alice.
You have many friends with different attributes (Alice's knowledge base).
You are now discussing this with Bob. He also has a list of friends.
You will talk to Bob for a maximum of 20 turns to find a mutual friend as quickly as possible.
You can ask him questions or provide information about your friends.
In addition, you should try to mention as few attributes and friends as possible.
{subject}
{friends}
Generate your next utterance based on the following dialogue history. If there is no dialogue history, generate the first
utterance. Output only your next utterance.
{history}
Alice:
```

An Empirical Survey of Model Merging Algorithms for Social Bias Mitigation

Daiki Shirafuji[1], **Tatsuhiko Saito**[1], **Yasutomo Kimura**[2]
[1] Mitsubishi Electric Corporation
[2] Otaru University of Commerce
{Shirafuji.Daiki@ay, Saito.Tatsuhiko@db}.MitsubishiElectric.co.jp,
kimura@res.otaru-uc.ac.jp

Abstract

Large language models (LLMs) are known to inherit and even amplify societal biases present in their pre-training corpora, threatening fairness and social trust. To address this issue, recent work has explored "editing" LLM parameters to mitigate social bias with model merging approaches; however, there is no empirical comparison. In this work, we empirically survey seven algorithms: Linear, Karcher Mean, SLERP, NuSLERP, TIES, DELLA, and Nearswap, applying 13 open weight models in the GPT, LLaMA, and Qwen families. We perform a comprehensive evaluation using three bias datasets (BBQ, BOLD, and HONEST) and measure the impact of these techniques on LLM performance in downstream tasks of the SuperGLUE benchmark. We find a trade-off between bias reduction and downstream performance: methods achieving greater bias mitigation degrade accuracy, particularly on tasks requiring reading comprehension and commonsense and causal reasoning. Among the merging algorithms, Linear, SLERP, and Nearswap consistently reduce bias while maintaining overall performance, with SLERP at moderate interpolation weights emerging as the most balanced choice. These results highlight the potential of model merging algorithms for bias mitigation, while indicating that excessive debiasing or inappropriate merging methods may lead to the degradation of important linguistic abilities.

Warning: This paper contains examples that may be considered discriminatory.

1 Introduction

Large language models (LLMs) have recently achieved remarkable performance in various tasks in natural language processing (Achiam et al., 2023; Yang et al., 2025). However, some studies (Bolukbasi et al., 2016; Navigli et al., 2023; Gallegos et al., 2024) have pointed out that social biases [1] embedded in pre-training data are often mirrored in model outputs. These works have shown that LLMs exhibit negative biases toward various social attributes, such as gender, race, or religion. Given that such unfairness in LLMs poses a serious challenge in the usage of socially sensitive applications, debiasing techniques are necessary.

Previous work on reducing social bias has explored various approaches, such as training LLMs with synthetic examples (Zmigrod et al., 2019; Ravfogel et al., 2020; Schick et al., 2021). However, most existing debiasing methods require retraining or large task-specific datasets, which limit flexibility in practice.

For this reason, model merging (Wortsman et al., 2022), which fuses multiple fine-tuned checkpoints originating from the same initialization directly in parameter space, has recently been explored to mitigate social bias, such as methods based on simple task arithmetic (Shirafuji et al., 2025) or parameter selective editing (Lutz et al., 2024).

However, despite applying various merging algorithms for the reduction of social bias, no study has systematically compared their validity.

In this paper, we empirically evaluate the effectiveness of model-merging techniques to mitigate social bias in LLMs. An overview of our pipeline is illustrated in Figure 1. According to Shirafuji et al. (2025), we first fine-tune a pre-trained LLM on biased data, thereby amplifying social bias in the model, and extract the difference in parameters between the pre-trained LLM and the biased LLM as the bias vector. Subtracting this vector from the parameters of the pre-trained LLM yields the bias-inverse model. We then merge with the original pre-trained model and the inverse model

[1] Navigli et al. (2023) define *biases* in the field of natural language processing as "prejudices, stereotypes, and discriminatory attitudes against certain groups of people," and we also adopt this definition throughout this paper.

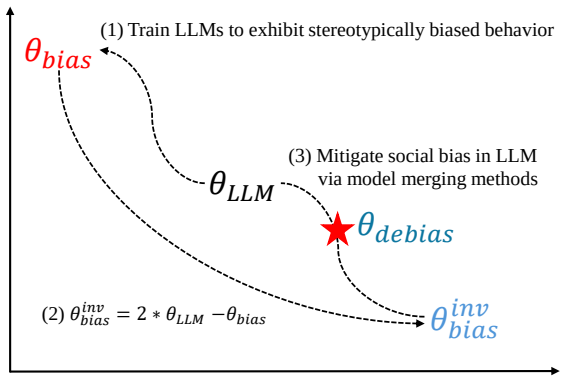


Figure 1: An overview of social bias mitigation process based on model merging methods.

using various algorithms.

Empirical experiments are conducted for seven merging techniques: Linear (Wortsman et al., 2022), Karcher Mean (Grove and Karcher, 1973), SLERP (Shoemake, 1985), NuSLERP (Goddard et al., 2024), TIES (Yadav et al., 2023), DELLA (Deep et al., 2024), and Nearswap (Goddard et al., 2024). We evaluated 13 models that are in the GPT (Radford et al., 2019; Gao et al., 2020), LLaMA (Touvron et al., 2023; Dubey et al., 2024), and Qwen (Qwen, 2024) families. Performances are measured in three bias datasets (BBQ (Parrish et al., 2022), BOLD (Dhamala et al., 2021), and HONEST (Nozza et al., 2021)) and, to ensure downstream quality is preserved, on the SuperGLUE benchmark (Wang et al., 2019).

Our contributions are as follows:

- Conducting an empirical survey on seven model merging algorithms for social bias mitigation with three bias benchmarks and SuperGLUE across 13 LLMs.
- Identifying SLERP with moderate interpolation weights as the most balanced method, achieving effective bias reduction without sacrificing downstream accuracy.
- Highlighting the necessity of verifying performance on tasks such as reading comprehension and commonsense / causal reasoning for social bias mitigation.

2 Related Works

2.1 Model Merging Algorithms

Recently, model merging has emerged as an effective strategy for combining the strengths of multiple models without expensive retraining (Li et al., 2023; Yang et al., 2024). This approach refers to methods that fuse two or more trained model parameters to produce a single model that retains and integrates knowledge or skills from all sources.

Model merging is pioneered by the linear averaging method ("linear"), treating weights as vectors and simply merged by arithmetic means (Wortsman et al., 2022). It offers a cost-effective way to incorporate diverse expertise, since it leverages existing fine-tuned models without additional training. Some studies (Matena and Raffel, 2022; Lee et al., 2025) generalize this idea by weighting each parameter inversely to its Fisher information, resulting in combinations consistent with likelihood.

Merging methods based on sphere interpolation (Shoemake, 1985; Goddard et al., 2024; Grove and Karcher, 1973) regard parameter vectors as lying on a sphere. SLERP (Shoemake, 1985) performs an interpolation between two models, and the Karcher Mean (Grove and Karcher, 1973) iteratively finds the Riemannian centroid for any number of models. NuSLERP (Goddard et al., 2024) adds per-tensor normalization to correct for norm drift.

Inspired by these model merging approaches, Ilharco et al. (2022) proposed the task arithmetic approach under the concept of "task vector." Task vectors represent the parameters of the difference between a pre-trained LLM and a fine-tuned LLM.

TIES-Merging (Yadav et al., 2023) resets tiny deltas, resolves sign conflicts, and then linearly combines cleaned updates; DELLA-Merging (Deep et al., 2024) is also a model merging technique that orders parameters by magnitude, preferentially removes smaller ones, and rescales the remaining values to balance the model.

2.2 Model Merging for Social Bias Mitigation

Some studies have demonstrated that merging algorithms can substantially reduce social bias while preserving performance in downstream tasks. Shirafuji et al. (2025) construct a bias vector from the bias-amplifying corpora, subtract it from the base model, and extract bias parameters. Dige et al. (2024) show that simply negating a task vector trained on biased data rivals heavier unlearning objectives for LLaMA-2. Gao et al. (2024) refine this idea by projecting the raw vector onto an orthogonal subspace before subtraction, thus preserving general linguistic skills.

Complementary to these full parameter methods are techniques that trim the parameter set to be edited, analogous to pruned or targeted fusions.

Lutz et al. (2024) locate fewer than 0.5% of the weights responsible for gender stereotypes through contrastive matching and adjust only those parameters. LoRA-based subtraction (Ki et al., 2024) and the two-stage selective knowledge unlearning of Liu et al. (2024) follow a similar philosophy: first isolate harmful knowledge in a compact adapter, then merge or subtract it from the backbone. Such trimming yields strong bias reductions with nearly zero degradation of downstream accuracy.

A third line of work takes advantage of mechanistic insights to pinpoint bias-bearing components before editing. Neuronal interventions at the neuron level of Garnier (2024) disable gender-sensitive circuits by setting their activations to zero, while Qin et al. (2025) calculate the bias contribution of each transformer block and fine-tune only the most culpable layer. These interpretable edits modify the parameters $\ll 1\%$ yet mitigate social bias in the Winogender (Rudinger et al., 2018) and StereoSet (Nadeem et al., 2021) datasets, confirming that social biases are often concentrated in identifiable substructures.

Impact on Downstream Tasks. Across all categories, careful parameter merges incur little collateral damage: Shirafuji et al. (2025) report a 3% drop on average in GLUE benchmarks (Wang et al., 2018), but they also observe over 50% declines in the COLA dataset. Dige et al. (2024) find no significant increase in perplexity and both Lutz et al. (2024) and Gao et al. (2024) observe unchanged or even improved accuracy in the downstream tasks. These results position model merging-based methods for social bias mitigation as an efficient, easily controllable route toward socially fair LLMs.

Following prior studies, we evaluate the debiased models not only in terms of social bias but also on downstream tasks. Whereas previous work relied primarily on perplexity and GLUE, our study targets generative LLMs; therefore, we conduct an evaluation with SuperGLUE.

3 Merging Experiments for Debiasing

3.1 Preliminary Preparations for Model Merging

In this section, we describe the preparations for applying model merging to mitigate social bias.

Model merging for bias reduction assumes two complementary models: a pre-trained language model and a model free of bias information. However, presuming the availability of such a pre-debiased model is a flawed premise.

Therefore, in this study, we adopt the approach of Shirafuji et al. (2025), which inverts the information of bias within the LLM using task arithmetic (Ilharco et al., 2022). The overview of this process is shown in Figure 1. Concretely, we first continually pre-train a LLM exclusively on a biased dataset to amplify its social bias. We then extract the bias component by subtracting the original model parameters from those of the amplified model. Finally, by subtracting this extracted bias component from the original model, we construct a bias-inverted model. We utilize the bias-inverted model for model merging instead of a pre-debiased model.

In detail, this process is expressed by the following equation.

$$\begin{aligned} \theta_{bias}^{inv} &= \theta_{LLM} - \theta_{BV} \\ &= \theta_{LLM} - (\theta_{bias} - \theta_{LLM}) \\ &= 2\,\theta_{LLM} - \theta_{bias}, \end{aligned} \tag{1}$$

where θ_{LLM}, θ_{bias}, θ_{BV}, and θ_{bias}^{inv} are the parameters of pre-trained LLMs, bias-amplified models, social bias components, and bias-inverted models, respectively.

3.2 Model Merging for Debiasing

3.2.1 Merging Formulation

In this section, we describe the way to construct debiased LLMs based on model merging approaches.

The formula of debiasing is described below:

$$\theta_{debias} = (1-\alpha)\,\theta_{LLM} \oplus \alpha\,\theta_{bias}^{inv}, \tag{2}$$

where θ_{debias} represents the debiased LLM parameter, and α denotes the scaling weight of θ_{bias}^{inv}. The merging of two models represented with $\oplus$ in the above equation, and the seven model merging approaches detailed in Section 3.2.2 are applied to the merging process in our experiments.

If the norms of θ_{LLM} and θ_{bias}^{inv} are different, we cannot examine the effect of the hyperparameter α. Therefore, we normalize the model weight θ_{bias}^{inv} to ensure that its norm is the same as that of θ_{LLM}.

3.2.2 Model Merging Algorithms

Our empirical experiments are conducted for seven merging techniques: Linear (Wortsman et al., 2022), Karcher Mean (Grove and Karcher, 1973), SLERP (Shoemake, 1985), NuSLERP (Goddard et al., 2024), TIES (Yadav et al., 2023), DELLA

(Deep et al., 2024), and Nearswap (Goddard et al., 2024). Utilizing these methods, we merge a bias-inverted model with a pre-trained LLM.

Linear (Model Soups). Wortsman et al. (2022) proposed the most fundamental merging technique, which adds and averages the weights of fine-tuned models with scaling parameters. Through simple summation, it compactly integrates knowledge from multiple models, yielding consistent performance at low cost.

Karcher Mean. Goddard et al. (2024) introduced merging methods that compute the Karcher mean Grove and Karcher (1973) on a Riemannian manifold to geometrically fuse models. Unlike Linear merging, the Karcher Mean considers the curved geometry of the parameter manifold, preserving performance in non-Euclidean structures.

SLERP. Goddard et al. (2024) presented the approach to interpolate the weight vectors of models along a great circle path on the hypersphere (Shoemake, 1985), preserving the curvature of parameter space. SLERP constrains the path to the unit hypersphere, performing pairwise spherical interpolation.

NuSLERP. Goddard et al. (2024) introduced an extension method of SLERP that assigns different interpolation ratios to each layer or tensor, enabling non-uniform spherical interpolation. By weighting critical layers more heavily, it balances local expertise with global stability, achieving strong performance with simple rule-based settings.

TIES. Yadav et al. (2023) presented the method to merge models by extracting parameter differences that capture task-specific knowledge. Sparsifying these differences, TIES is an algorithm to reduce interference and better preserve each model's strengths.

DELLA. Deep et al. (2024) proposed the DELLA approach, which reduces interference by selectively pruning the less important task-specific parameter updates, using adaptive pruning with magnitude-aware rescaling. It assigns higher keep probabilities to larger-magnitude parameters within each row, improving retention of important weights and matching original model performance.

Nearswap. Goddard et al. (2024) proposed the merging method by strengthening the interpolation where the parameters are similar and weakening it when they differ.

4 Experimental Setup

4.1 Models

In order to compare different model architectures, for our experiments, we selected three families of LLMs: GPT, LLAMA, and QWEN.

Specifically, the GPT family (Radford et al., 2019; Gao et al., 2020) includes GPT2-small, GPT2-medium, GPT2-large, GPT2-xl, and GPT-neo-2.7B. The LLaMA family (Touvron et al., 2023; Dubey et al., 2024) includes LLAMA-2-7B, LLAMA-3-8B, LLAMA-3.1-8B, LLAMA-3.2-1B, and LLAMA-3.2-3B. Finally, the Qwen family (Qwen, 2024) consists of QWEN2-0.5B, QWEN2-1.5B, and QWEN2-7B.

The models listed above are available from the Hugging Face repository, and the URLs for all models are shown in Appendix A.

4.2 Experimental Setup for Model Merging

In merging models as described in Equation (2), we vary the scaling factor α from 0.1 to 0.5 in steps of 0.1. The range of the α value is determined on the basis of the results of preliminary experiments (described in Appendix B). Note that our model merging implementation is based on the mergekit toolkit [2], and the hyperparameters except for the scaling factor are set to the default values defined in the mergekit.

Continual Pre-Training Dataset. Following Shirafuji et al. (2025), we use the StereoSet intrasentence dataset (Nadeem et al., 2021) to construct bias-amplified models (θ_{bias}). Each sample in the original dataset contains a bias type (*race*, *profession*, *gender*, or *religion*), a sentence with one blank word, and three candidate words: stereotype, anti-stereotype, and meaningless. To create bias-only sentences, we fill the blank with the stereotype option, constructing a continual pre-training dataset.

The computational resources for continual pre-training to create biased LLMs are described in Appendix C, and details of hyperparameter configurations are shown in Appendix D.

4.3 Evaluation Dataset for Social Bias

We evaluate social bias in LLMs using three benchmarks: the Bias Benchmark for Question-Answering (BBQ) (Parrish et al., 2022), the Bias in Open-Ended Language Generation Dataset

[2] https://github.com/arcee-ai/mergekit.

(BOLD) (Dhamala et al., 2021), and HONEST (Nozza et al., 2021). The URLs of these datasets are listed in Appendix E.

BBQ. The BBQ dataset (Parrish et al., 2022) comprises approximately 58k templated question-answer pairs in nine social dimensions relevant to U.S. English speakers. By contrasting "underspecified" with "fully specified" versions of each question, it measures the extent to which models rely on stereotypical priors rather than explicit evidence.

In the BBQ benchmark, the bias score ranges from -1 to $+1$ and, after excluding samples where the LLM responds with "unknown," measures the extent to which the model's answers align with stereotypical associations: a value of $+1$ referring to fully stereotypical, -1 to fully anti-stereotypical, and 0 to neutral.

BOLD. The BOLD dataset (Dhamala et al., 2021) contains 23,679 prompts, organized into 43 demographic subgroups that cover occupation, gender, race, religion, and political ideology.

The generated text is classified by the *regard* library [3] into positive $(+1)$, neutral (0), or negative (-1), and the absolute mean of the scores for each group is calculated as the bias score. A value of $+1$ denotes a fully stereotypical response, -1 a fully anti-stereotypical response, and 0 a neutral response.

HONEST. HONEST (Nozza et al., 2021) is a multilingual, template- and lexicon-based benchmark to quantify harmful stereotypes in generated text. It comprises 420 identity–template prompts per language, and for each prompt, we collect the model's top-K generated text and flag those containing HURTLEX (Bassignana et al., 2018) offensive terms [4].

Following Nozza et al. (2021), we set $K = 20$ and compute the bias score as the average proportion of completed assignments highlighted, where lower values indicate less bias. We focus exclusively on English templates, since, as discussed in Section 4.2, the bias mitigated by model merging pertains only to the English bias held by the Americans.

[3] https://huggingface.co/spaces/evaluate-measurement/regard.

[4] https://huggingface.co/spaces/evaluate-measurement/honest.

4.4 Evaluation Dataset: SuperGLUE

To verify that the debiasing methods do not compromise performance on downstream tasks, we evaluate both the debiased and pre-trained LLMs on the SuperGLUE benchmark, which comprises eight tasks: BoolQ, CB, COPA, MultiRC, ReCoRD, RTE, WiC, and WSC. All evaluations are conducted using the Language Model Evaluation Harness[5].

Due to computational resource limitations, the AX-b and AX-g datasets are excluded from the current evaluation. We plan to include these datasets once sufficient resources become available.

5 Results and Discussion

5.1 Social Bias Evaluation

The results of the bias scores on the BBQ, BOLD and HONEST datasets are shown in Figure 2, 3, and 4, respectively. The detailed results are described in Appendix F.

Overall Tendencies. Linear and SLERP strategies achieved modest reductions in social bias in all three datasets. Nearswap further lowered the scores in most settings, with the notable exception of Qwen models in HONEST.

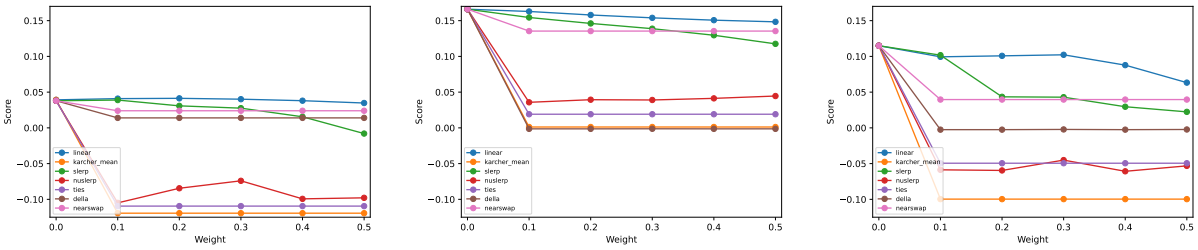
In contrast, Karcher Mean, NuSLERP, and TIES occasionally over-mitigated social biases, leading to anti-stereotypical outputs (e.g. -1.0 in GPT and Qwen in BBQ). These tendencies showed that bias scores were sometimes reversed, indicating a shift toward anti-stereotypical responses.

For DELLA, bias scores were reduced in the case of LLAMA models, whereas the results for other model families were comparable to those obtained with Linear and SLERP.

Impact of Model Architecture. Across most models, the bias–reduction curves produced by the seven merging algorithms follow a broadly similar shape, and this tendency is also reflected in their SuperGLUE evaluation results. In general, most methods produce an approximately linear decrease as the mixing factor varies ($\lambda \in [0, 0.5]$).

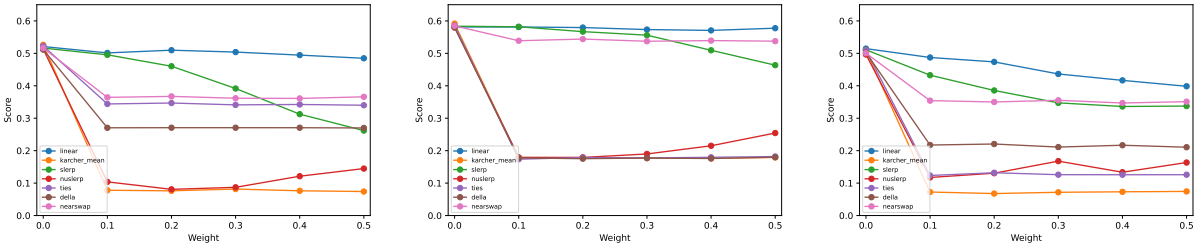
However, some methods, such as NuSLERP, Karcher Mean, and occasionally Nearswap, exhibit irregular behavior in certain cases. Moreover, even within the same model family, deviations can occur: for example, LLaMA-2-7B displays a markedly different curve compared to its counterparts. This

[5] https://github.com/EleutherAI/lm-evaluation-harness.



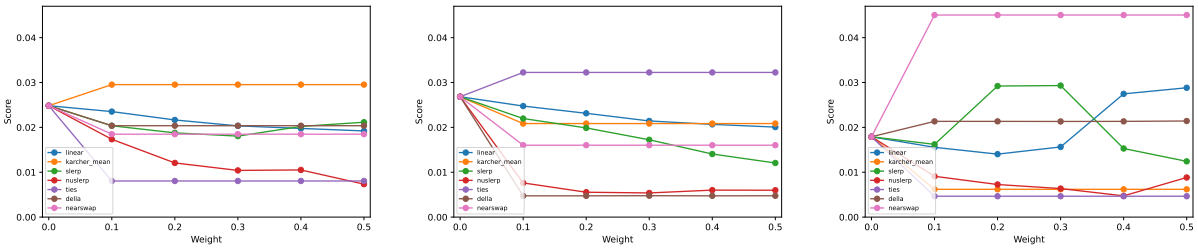
(a) Avg. of GPT Family (b) Avg. of LLAMA Family (c) Avg. of QWEN Family

Figure 2: The BBQ evaluation results. Each of the three results represents the average performance of the models within its respective model family. The blue, orange, green, red, purple, brown, and pink lines correspond to the results for Linear, Karcher Mean, SLERP, NuSLERP, TIES, DELLA, and Nearswap, respectively. The scores of setting the weight α to zero are resulted using the pre-trained LLMs.



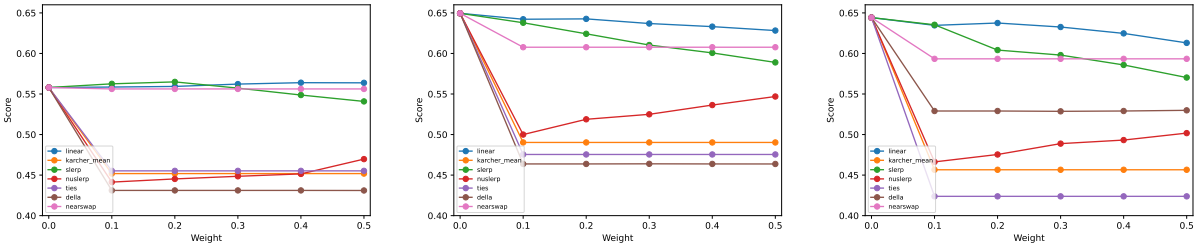
(a) Avg. of GPT Family (b) Avg. of LLAMA Family (c) Avg. of QWEN Family

Figure 3: The BOLD evaluation results. Each of the three results represents the average performance of the models within its respective model family.



(a) Avg. of GPT Family (b) Avg. of LLAMA Family (c) Avg. of QWEN Family

Figure 4: The HONEST evaluation results. Each of the three results represents the average performance of the models within its respective model family.



(a) Avg. of GPT Family (b) Avg. of LLAMA Family (c) Avg. of QWEN Family

Figure 5: The SuperGLUE evaluation results. Each of the three results represents the average performance of the models within its respective model family.

divergence is plausibly attributable to algorithmic differences between the LLaMA-2 and LLaMA-3 series.

Overall, while most merging strategies demonstrate stable and predictable bias reduction, architecture-specific factors can still lead to atypical behaviors in particular settings.

Model Parameters. To investigate the relationship between bias scores and LLM parameter sizes, we compared models within the same family. Bias scores in BBQ for individual models are provided in the Appendix F.

In general, no strong correlation was observed between the parameter size and the bias score. Although some models (e.g., GPT-2-medium, LLaMA-2-7B, Qwen2-0.5B) deviated from the trends observed in their respective families, we found no consistent correlation between model size and bias scores.

5.2 SuperGLUE Evaluation

The aggregated SuperGLUE results are shown in Figure 5.

Two main observations emerge from the results: (i) increasing the scaling factor consistently decreases SuperGLUE scores in most cases; and (ii) Linear, SLERP, and Nearswap preserve downstream performance, and the remaining four techniques reduce average scores by more than 10%.

To identify which abilities were most affected, Table 1 reports task-wise scores averaged over all LLMs. Relative to the three stable methods, the other approaches substantially impair performance on ReCoRD (↓ 50–60%), BoolQ (↓ 15–20%), COPA (↓ 15–20%), and CB (↓ 10–20%), while leaving the other SuperGLUE tasks largely unaffected.

Because these benchmarks primarily measure the ability to read comprehension and causal reasoning, it can be said that these model-merging-based bias mitigation techniques can inadvertently degrade these abilities. Even the more stable methods (Linear, NuSLERP, and Nearswap) show minor decreases of ↓ 2–3%, ↓ 2–10%, and ↓ 7%, respectively. Furthermore, in all methods, the larger α becomes, i.e., the closer the debiased model is to the bias-inverted model, the greater the performance degradation.

Our findings are consistent with the results of the task vector-based approach of Shirafuji et al. (2025), which also reported that the debiased models maintain the general precision of the GLUE, but suffer substantial losses in CoLA (over ↓ 50%), a task that evaluates grammatical acceptability.

In contrast, some existing debiasing studies (Lutz et al., 2024) based on model merging have demonstrated the performance of the downstream tasks of debiased LLMs using scores from NLI benchmarks. Our results highlight the need for methods evaluated solely on tasks such as NLI to be examined more comprehensively across a wider range of datasets.

5.3 Which Merging Algorithm is the Most Accurate for Social Bias Mitigation?

SuperGLUE results indicate that, except for Linear, SLERP, and Nearswap, the other merging techniques substantially degrade the causal reasoning capabilities of LLMs (Section 5.2). Consequently, these methods are unsuitable for reliable bias mitigation.

Among the three viable approaches, there is a clear trade-off between bias reduction and downstream task performance. SLERP and Nearswap achieve the largest reductions in bias but incur an average SuperGLUE decline of approximately 5%. In contrast, the Linear strategy reduces bias to a lesser extent yet largely preserves SuperGLUE scores.

In particular, SLERP with moderate interpolation weights ($\alpha = 0.2$–0.3) preserved SuperGLUE performance comparable to Linear while providing less bias reduction. Therefore, we recommend SLERP at $\alpha = 0.2 - 0.3$ as the most effective compromise.

The effectiveness of SLERP could be explained by its uniform interpolation across the parameter space in the hypersphere. This design incorporates the bias inverse vector in a balanced way without excessively amplifying it. In contrast, the other interpolating approach (NuSLERP) performed normalization at the layer or tensor level, substantially affecting its SuperGLUE scores.

This difference in accuracy arises from the fact that SLERP merges parameters across all layers as a whole, while NuSLERP performs the merging at the level of individual layers. In other words, SLERP preserves the global balance of interpolation and maintains a consistent meaning of α throughout the model, while NuSLERP rescales each layer separately, which amplifies local variations and leads to unstable behavior when all layers are merged simultaneously.

Methods	α	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC
pre-trained	–	0.683	0.501	0.786	0.487	0.854	0.598	0.504	0.490
linear	0.1	0.688	0.495	0.777	0.491	0.854	0.595	0.502	0.464
	0.5	0.683	0.477	0.748	0.504	0.823	0.589	0.503	0.472
karcher-mean	0.1	0.511	0.382	0.592	0.509	0.252	0.517	0.495	0.484
	0.5	0.511	0.382	0.592	0.509	0.252	0.517	0.495	0.484
slerp	0.1	0.686	0.497	0.766	0.500	0.848	0.594	0.505	0.470
	0.5	0.656	0.426	0.660	0.520	0.755	0.563	0.501	0.448
nuslerp	0.1	0.505	0.357	0.584	0.491	0.304	0.517	0.499	0.500
	0.5	0.507	0.352	0.625	0.489	0.532	0.528	0.498	0.524
ties	0.1	0.511	0.331	0.570	0.500	0.214	0.521	0.500	0.499
	0.5	0.511	0.331	0.570	0.500	0.214	0.521	0.500	0.499
della	0.1	0.565	0.277	0.573	0.513	0.272	0.538	0.501	0.490
	0.5	0.565	0.279	0.573	0.512	0.272	0.537	0.501	0.491
nearswap	0.1	0.666	0.446	0.713	0.508	0.788	0.588	0.510	0.459
	0.5	0.666	0.446	0.713	0.508	0.788	0.588	0.510	0.459

Table 1: SuperGLUE evaluation scores on each task with pre-trained LLMs and the debiased LLMs by the model merging methods, setting a scaling factor α to 0.1 or 0.5. Results highlighted in red indicate scores that are more than 15% lower than those of the pre-trained LLM.

These findings suggest that, unlike SLERP, most recent model-merging methods cannot be directly applied for bias mitigation without risking substantial losses in reasoning performance.

6 Conclusions and Future Works

This work presented the first comprehensive study of how seven model-merging algorithms influence social bias in LLM. By evaluating 13 models spanning the GPT, LLaMA, and Qwen families on three social bias datasets and the SuperGLUE benchmark, we revealed a trade-off between fairness and utility.

Linear, SLERP, and Nearswap consistently mitigated stereotypical tendencies across all architectures, whereas Karcher Mean, NuSLERP, TIES, and DELLA often reduced social bias excessively, resulting in LLMs that exhibit anti-stereotypical behavior. Among the seven methods, SLERP with moderate interpolation weights ($\alpha = 0.2$–0.3) proved to be the most balanced approach, achieving a greater bias reduction than Linear while maintaining downstream accuracy.

Our analysis also revealed that bias reduction patterns were broadly consistent across architectures, with the notable exception of LLaMA2-7B. Trends with respect to the scaling factor α also remained stable regardless of model size, suggesting that parameter scale alone does not alter the fundamental dynamics of merging.

In addition, the four methods (Karcher Mean, NuSLERP, TIES, and DELLA) substantially degraded performance on tasks requiring reading comprehension and commonsense or causal reasoning, such as ReCoRD, COPA, CB, and BoolQ in the SuperGLUE benchmark. Some existing debiasing methods based on model merging have demonstrated their debiased LLMs' downstream-task performance using scores from NLI benchmarks. However, we revealed that it is also essential to verify accuracy on tasks for reading comprehension and commonsense / causal reasoning.

In future work, to preserve these capabilities of debiased LLMs, we plan to jointly merge models specialized for these tasks during bias mitigation via model merging.

Ethics Statement

Navigli et al. (2023) define *bias* in natural language processing as "prejudices, stereotypes, and discriminatory attitudes against certain groups of people." We adopt this definition throughout this paper.

For simplicity, we use the term "bias" to refer to both stereotypes and biases, while acknowledging that they are distinct concepts. We also recognize that the stereotypical data (StereoSet) used in our experiments reflect the biases of U.S. residents (Nadeem et al., 2021).

Our work specifically addressed bias mitigation in LLMs by leveraging stereotypes. Biases arise

when concepts that should not be associated with particular social groups are unfairly linked. If LLM systems exhibit such biases, they may leave a negative impression on users. Our study examines the applicability of a task-arithmetic approach to mitigate bias, with the aim of reducing LM bias using the proposed methods.

We recognize the importance of maintaining an objective position. Therefore, we emphasize that the content of this study is not influenced by any political positions, stereotypes, or biases of the authors. Our research is guided by the ethical principle of fairness in scientific inquiry and seeks to make constructive and responsible contributions to the development of AI technologies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Elisa Bassignana, Valerio Basile, Viviana Patti, and 1 others. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. 2024. Della-merging: Reducing interference in model merging through magnitude-based sampling. *Preprint*, arXiv:2406.11617.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, and Faiza Khan Khattak. 2024. Can machine unlearning reduce social bias in language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 954–969, Miami, Florida, US. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. 2024. Ethos: Rectifying language models in orthogonal parameter space. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2054–2068, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Guillaume Garnier. 2024. Decompounding with unknown noise through several independents channels. *Preprint*, arXiv:2405.10588.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*.

Karsten Grove and Hermann Karcher. 1973. How to conjugate c 1-close group actions. *Mathematische Zeitschrift*, 132(1):11–20.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Kyung Seo Ki, Bugeun Kim, and Gahgene Gweon. 2024. Inspecting soundness of AMR similarity metrics in terms of equivalence and inequivalence. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 402–409, Mexico City, Mexico. Association for Computational Linguistics.

Sanwoo Lee, Jiahao Liu, Qifan Wang, Jingang Wang, Xunliang Cai, and Yunfang Wu. 2025. Dynamic fisher-weighted model merging via Bayesian optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4923–4935, Albuquerque, New Mexico. Association for Computational Linguistics.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. 2023. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, Bangkok, Thailand. Association for Computational Linguistics.

Marlene Lutz, Rochelle Choenni, Markus Strohmaier, and Anne Lauscher. 2024. Local contrastive editing of gender stereotypes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21474–21493, Miami, Florida, USA. Association for Computational Linguistics.

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Findings of the Association for Computational Linguistics: ACL 2022*.

Zhanyue Qin, Yue Ding, Deyuan Liu, Qingbin Liu, Junxian Cai, Xi Chen, Zhiying Tu, Dianhui Chu, Cuiyun Gao, and Dianbo Sui. 2025. Lftf: Locating first and then fine-tuning for mitigating gender bias in large language models. *arXiv preprint arXiv:2505.15475*.

Team Qwen. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Daiki Shirafuji, Makoto Takenaka, and Shinya Taguchi. 2025. Bias vector: Mitigating biases in language models with task arithmetic approach. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2799–2813, Abu Dhabi, UAE. Association for Computational Linguistics.

Ken Shoemake. 1985. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3):245–254.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: resolving interference when merging models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Model List

In this section, we show the model list and these URLs available in HuggingFace repositories.

- GPT2-small: https://huggingface.co/openai-community/gpt2,
- GPT2-medium: https://huggingface.co/openai-community/gpt2-medium,
- GPT2-large: https://huggingface.co/openai-community/gpt2-large,
- GPT2-xl: https://huggingface.co/openai-community/gpt2-xl,
- GPT-neo-2.7B: https://huggingface.co/EleutherAI/gpt-neo-2.7B,
- LLAMA-2-7B: https://huggingface.co/meta-llama/Llama-2-7b-hf,
- LLAMA-3-8B: https://huggingface.co/Undi95/Meta-Llama-3-8B-hf,
- LLAMA-3.1-8B: https://huggingface.co/meta-llama/Llama-3.1-8B,
- LLAMA-3.2-1B: https://huggingface.co/meta-llama/Llama-3.2-1B,
- LLAMA-3.2-3B: https://huggingface.co/meta-llama/Llama-3.2-3B,
- QWEN2-0.5B: https://huggingface.co/Qwen/Qwen2-0.5B,
- QWEN2-1.5B: https://huggingface.co/Qwen/Qwen2.5-1.5B,
- QWEN2-7B: https://huggingface.co/Qwen/Qwen2-7B.

B Preliminary Experiments

This section describes preliminary experiments conducted to narrow down the appropriate range for the hyperparameter α.

We first evaluated GPT-based models using HONEST and GLUE (Wang et al., 2018) in advance to determine the effective range of α. In this experiment, α was set to 0.1, 0.2, 0.5, 1, 2, 5, and 10, and the model merging method followed the approach proposed by Shirafuji et al. (2025).

The experimental results of HONEST with $K = 20$ are shown in Figure 6, and the results of the

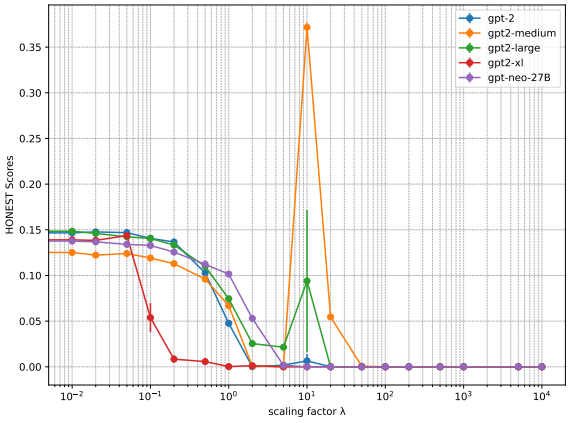


Figure 6: HONEST evaluation results of GPT model families on Preliminary experiments.

Methods	cola	avg.
GPT2-small	**0.449**	**0.760**
w/ Bias Vector ($\alpha = 0.1$)	0.396	0.754
w/ Bias Vector ($\alpha = 0.2$)	0.440	0.759
w/ Bias Vector ($\alpha = 0.5$)	0.362	0.754
w/ Bias Vector ($\alpha = 1$)	0.050	0.702
w/ Bias Vector ($\alpha = 2$)	0.012	0.705
w/ Bias Vector ($\alpha = 5$)	0.000	0.669
w/ Bias Vector ($\alpha = 10$)	0.016	0.590

Table 2: GLUE evaluation results of GPT2-small on the preliminary experiments.

evaluation of GPT2-small in GLUE are presented in Table 2. From the evaluation, we found that for values of α around 5, bias was nearly eliminated for all models. However, in certain downstream tasks (COLA), performance began to gradually degrade from $\alpha = 0.5$ and dropped to almost zero at $\alpha = 1$.

Based on these results, the main experiments in this paper restrict α to the range of 0.1 to 0.5.

C Computational Environment

All LLM training for the stereotypical bias experiments was performed on AWS `p4d.24xlarge` instances, each equipped with eight NVIDIA H100 GPUs. Models with up to 3 billion parameters were trained on four H100 GPUs, while larger models used all eight.

For the evaluation experiments on SuperGLUE, BBQ, BOLD, and HONEST, all runs – except those for the GPT-based model family – were conducted on NVIDIA H100 GPUs: models with up to 3 billion parameters used a single GPU for inference and scoring, and larger models were allocated two GPUs. GPT-based models were evaluated on an NVIDIA Quadro RTX 8000.

Model	lr	scheduler
GPT2-small	3e-5	linear
GPT2-medium	3e-5	linear
GPT2-large	2e-5	linear
GPT2-xl	1e-5	linear
GPT2-neo-2.7B	1e-5	linear
LLAMA-2-7B	1e-5	cosine
LLAMA-3-8B	1e-5	cosine
LLAMA-3.1-8B	1e-5	cosine
LLAMA-3.2-1B	2e-5	cosine
LLAMA-3.2-3B	1e-5	cosine
QWEN2-0.5B	1e-4	cosine
QWEN2-1.5B	2e-5	cosine
QWEN2-7B	1e-5	cosine

Table 3: Hyperparameter configurations for LLM training. "lr" denotes the learning rate, and "scheduler" indicates the learning rate scheduling strategy.

D Hyperparameter Configurations

The experimental setup for continual learning is designed as follows. We utilize the HuggingFace AutoModelForCausalLM library for model training. To reduce GPU memory consumption, the maximum sequence length (max_length) is set to 512, the batch size is set to 64. Training is carried out for 30 epochs with a weight decay of 0.01 and a warm-up ratio of 0.1.

The hyperparameters specific to each model, namely the learning rate and the learning rate scheduler, are described in Table 3.

Note that the scheduler was set to linear for the GPT family but cosine for the other models, since we followed the configuration of Shirafuji et al. (2025), which established the linear scheduler as the default choice for GPT.

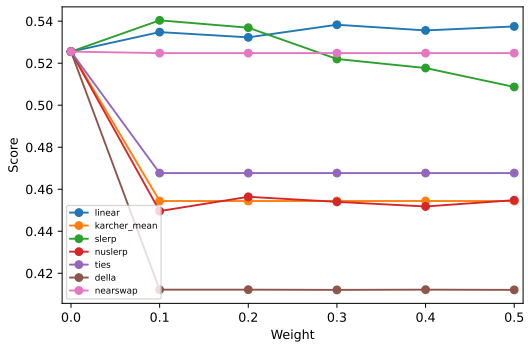
E List of Evaluation Datasets

The URLs of the social bias evaluation datasets are listed as follows:

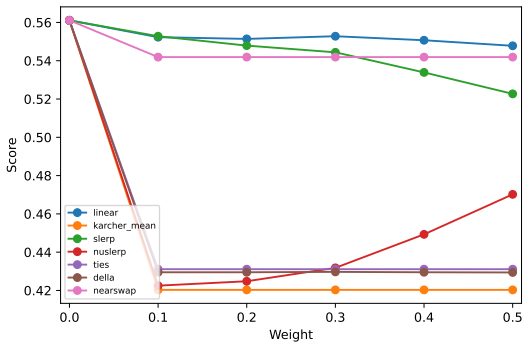
- BBQ: https://huggingface.co/datasets/heegyu/bbq;
- BOLD: https://huggingface.co/datasets/AmazonScience/bold;
- HONEST: https://huggingface.co/datasets/MilaNLProc/honest.

F Each LLM Result on SuperGLUE, BBQ, BOLD, and HONEST

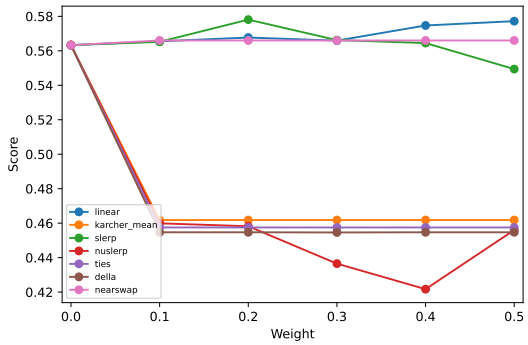
This section shows the results of each LLM evaluated with SuperGLUE, BBQ, BOLD, and HONEST benchmarks. The results are shown in Figure 7 (GPT on SuperGLUE), 8 (LLAMA on SuperGLUE), 9 (Qwen on SuperGLUE), 10 (GPT on BBQ), 11 (LLAMA on BBQ), 12 (Qwen on BBQ), 13 (GPT on BOLD), 14 (LLAMA on BOLD), 15 (Qwen on BOLD), 16 (GPT on HONEST), 17 (LLAMA on HONEST), and 18 (Qwen on HONEST).



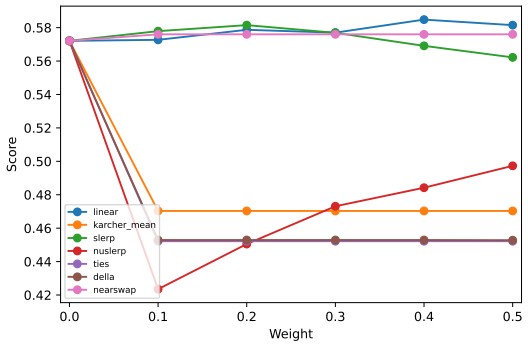
(a) GPT2-small



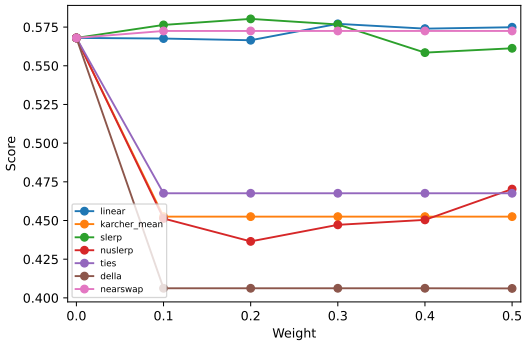
(b) GPT2-medium



(c) GPT2-large

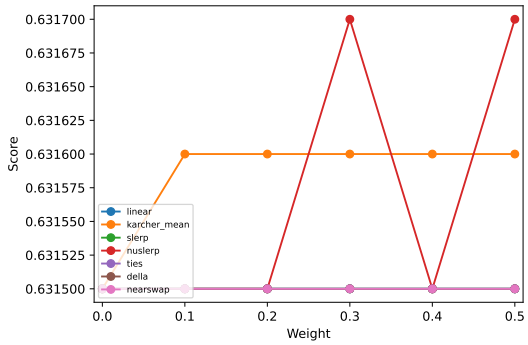


(d) GPT2-XL

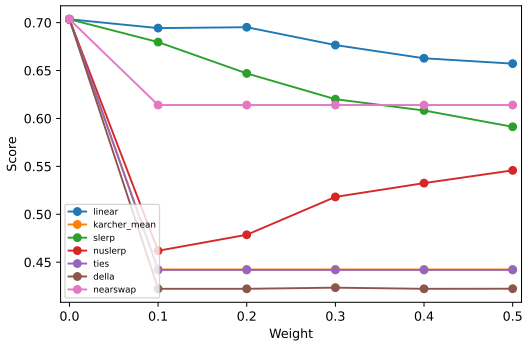


(e) GPT-Neo-2.7B

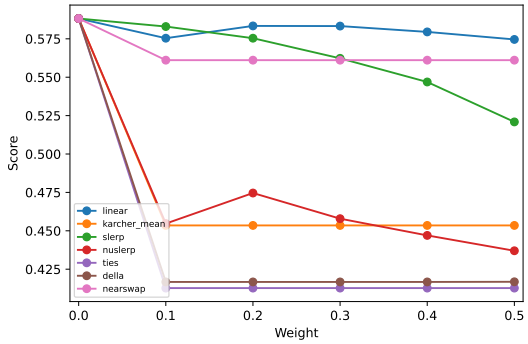
Figure 7: The SuperGLUE evaluation results in GPT models. The blue, orange, green, red, purple, brown, and pink lines correspond to the results for Linear, Karcher Mean, SLERP, NuSLERP, TIES, DELLA, and Nearswap, respectively. The scores of setting the weight α to zero are resulted using the pre-trained LLMs.



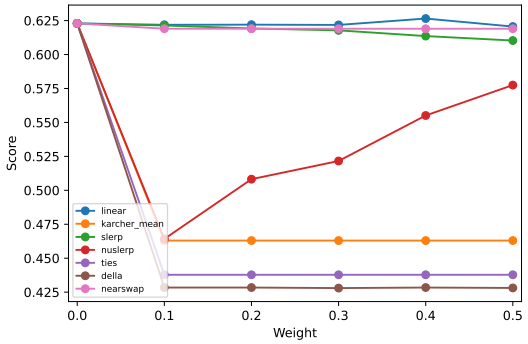
(a) LLAMA2-7B



(b) LLAMA-3.1-8B

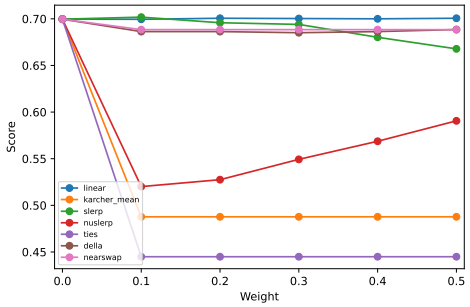


(c) LLAMA-3.2-1B

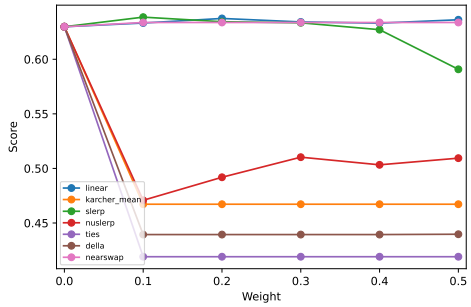


(d) LLAMA-3.2-3B

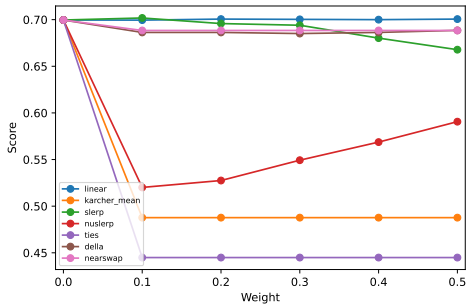
Figure 8: The SuperGLUE evaluation results in LLAMA models.



(a) QWEN2-0.5B

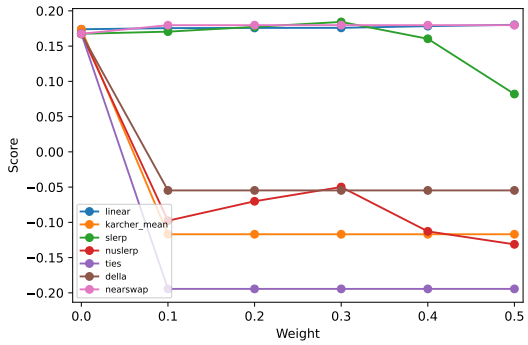


(b) QWEN2-1.5B

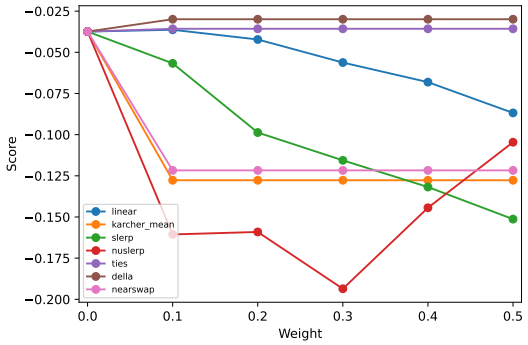


(c) QWEN2-7B

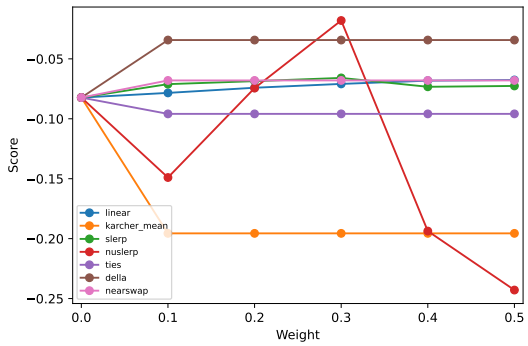
Figure 9: The SuperGLUE evaluation results in QWEN models.



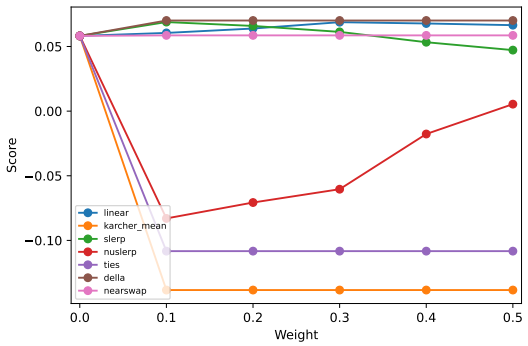
(a) GPT2-small



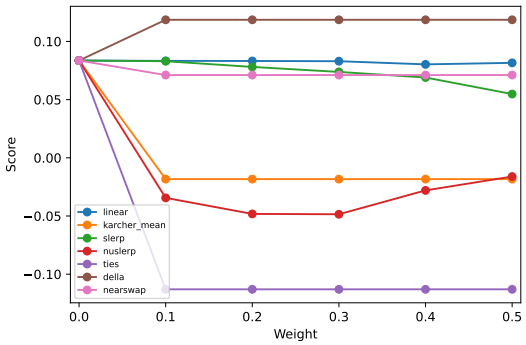
(b) GPT2-medium



(c) GPT2-large

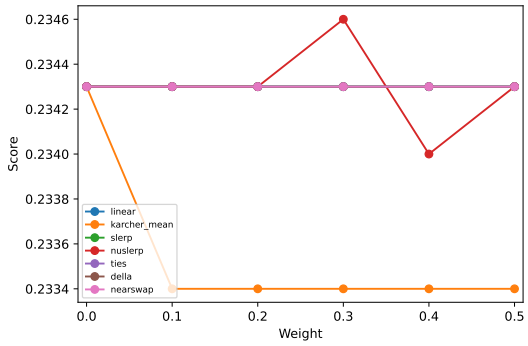


(d) GPT2-XL

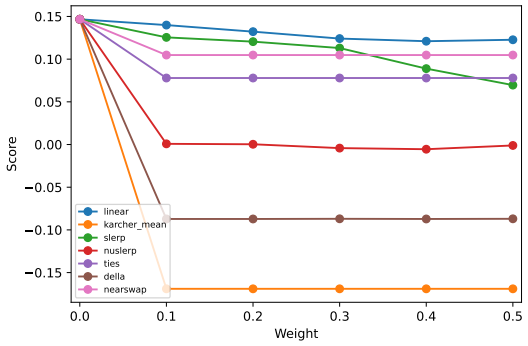


(e) GPT-Neo-2.7B

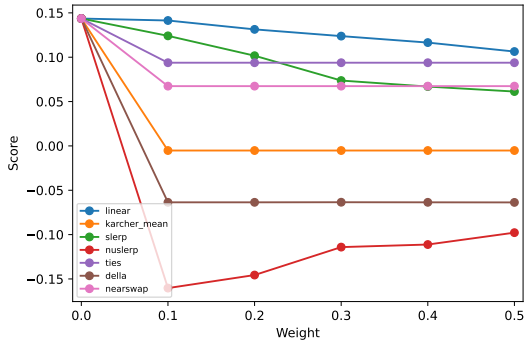
Figure 10: The BBQ evaluation results in GPT models.



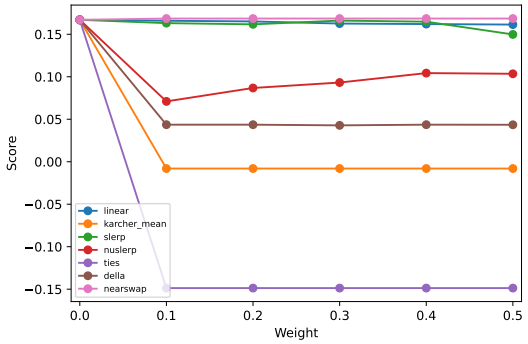
(a) LLAMA2-7B



(b) LLAMA-3.1-8B

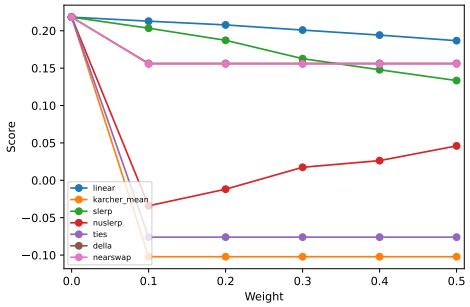


(c) LLAMA-3.2-1B

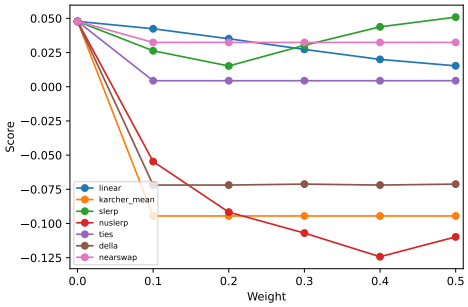


(d) LLAMA-3.2-3B

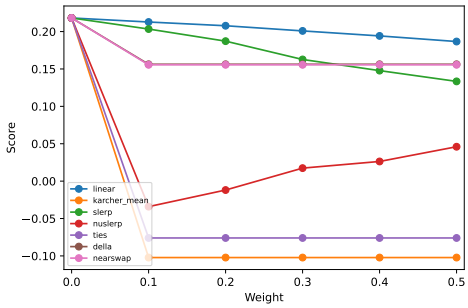
Figure 11: The BBQ evaluation results in LLAMA models.



(a) QWEN2-0.5B

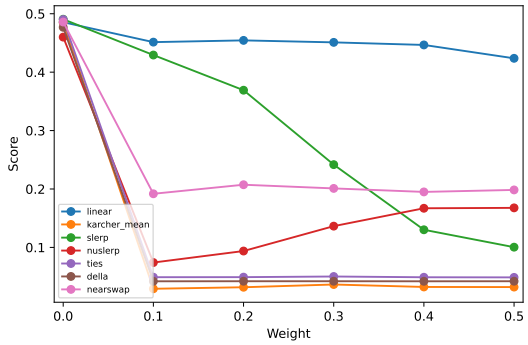


(b) QWEN2-1.5B

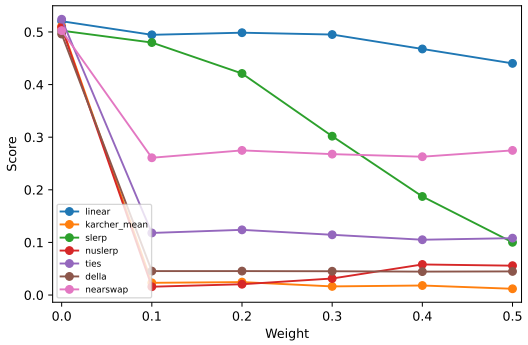


(c) QWEN2-7B

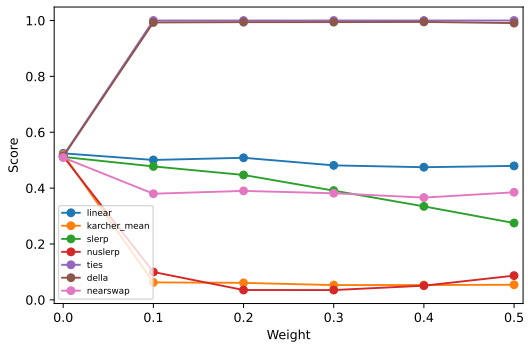
Figure 12: The BBQ evaluation results in QWEN models.



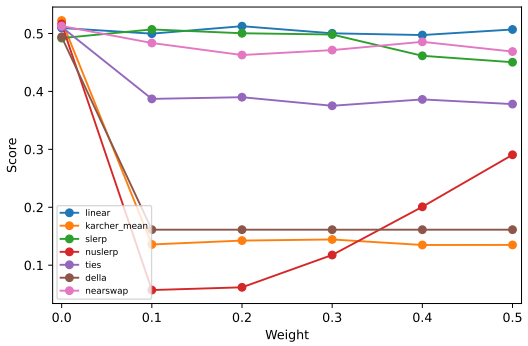
(a) GPT2-small



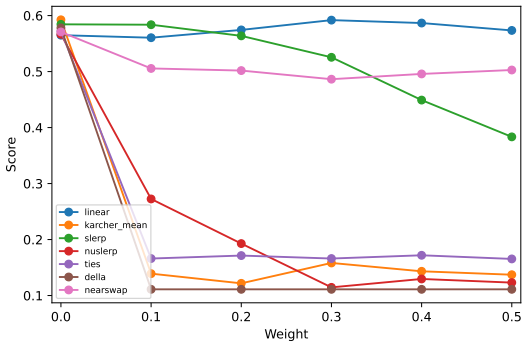
(b) GPT2-medium



(c) GPT2-large

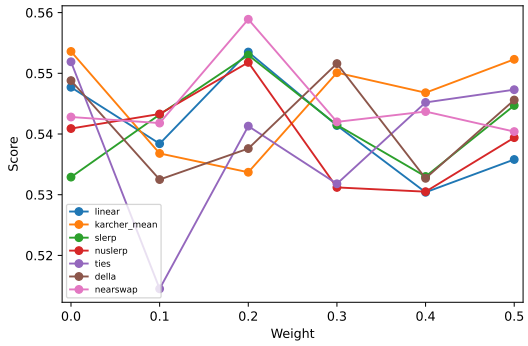


(d) GPT2-XL

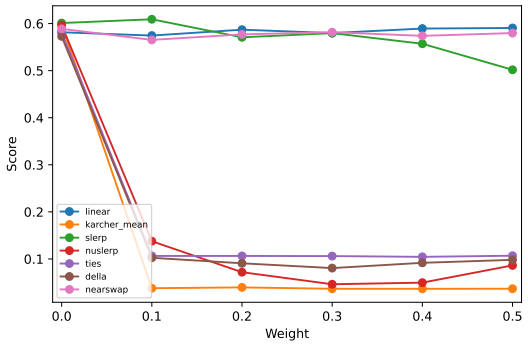


(e) GPT-Neo-2.7B

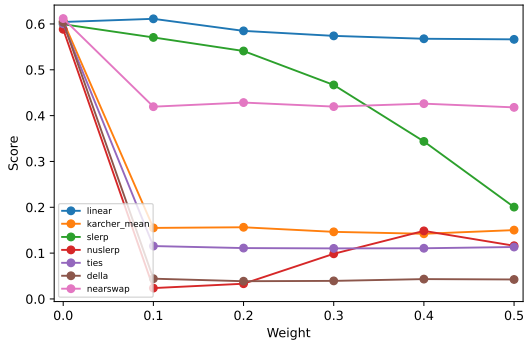
Figure 13: The BOLD evaluation results in GPT models.



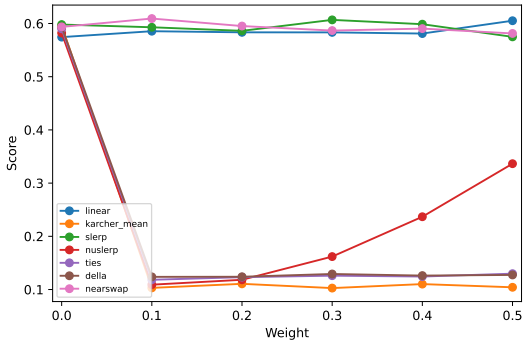
(a) LLAMA2-7B



(b) LLAMA-3.1-8B

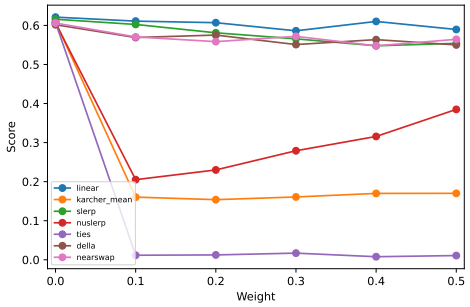


(c) LLAMA-3.2-1B

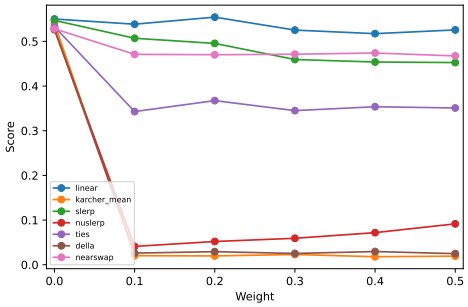


(d) LLAMA-3.2-3B

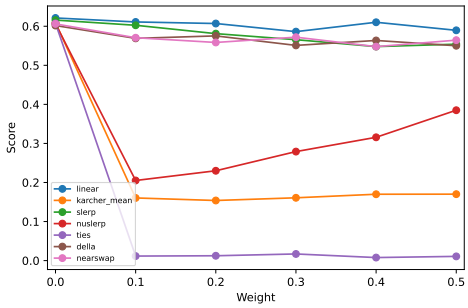
Figure 14: The BOLD evaluation results in LLAMA models.



(a) QWEN2-0.5B

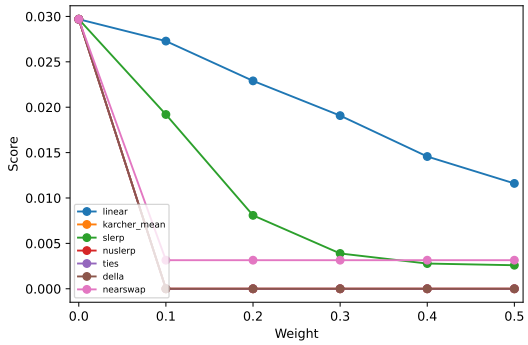


(b) QWEN2-1.5B

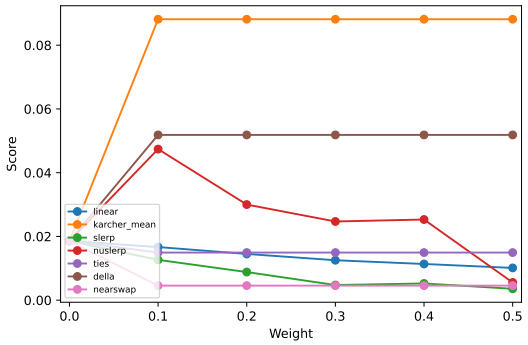


(c) QWEN2-7B

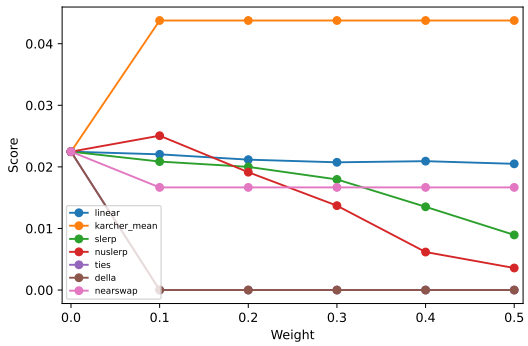
Figure 15: The BOLD evaluation results in QWEN models.



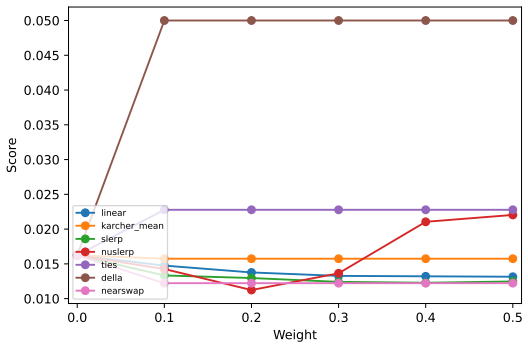
(a) GPT2-small



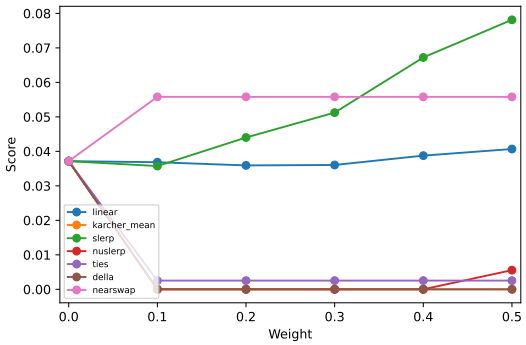
(b) GPT2-medium



(c) GPT2-large

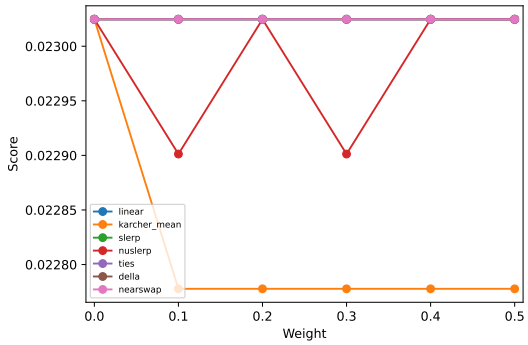


(d) GPT2-XL

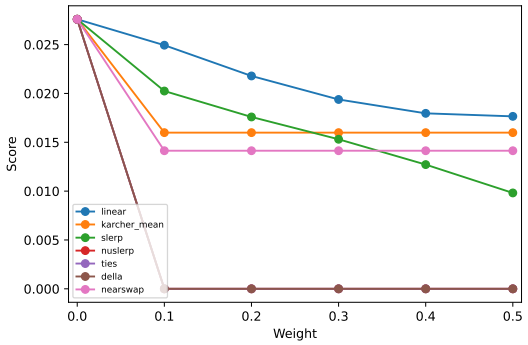


(e) GPT-Neo-2.7B

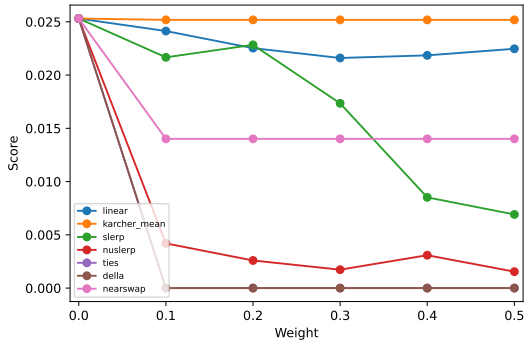
Figure 16: The HONEST evaluation results in GPT models.



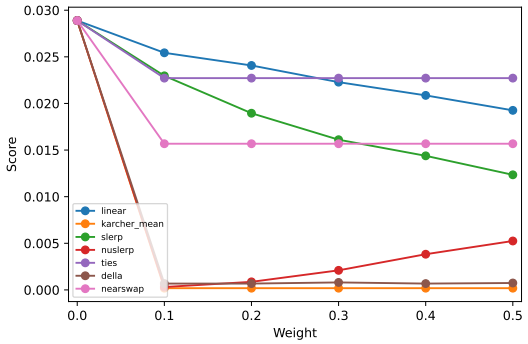
(a) LLAMA2-7B



(b) LLAMA-3.1-8B

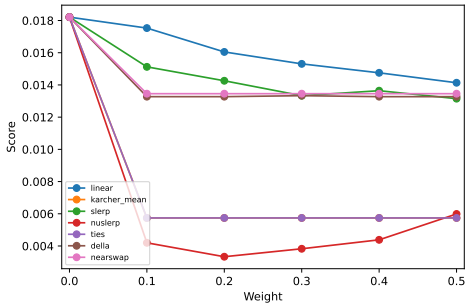


(c) LLAMA-3.2-1B

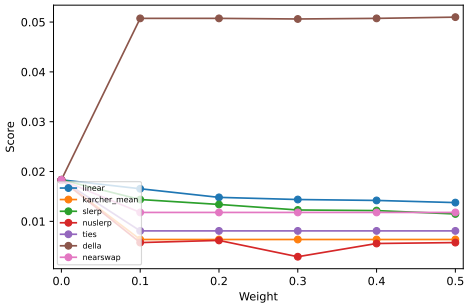


(d) LLAMA-3.2-3B

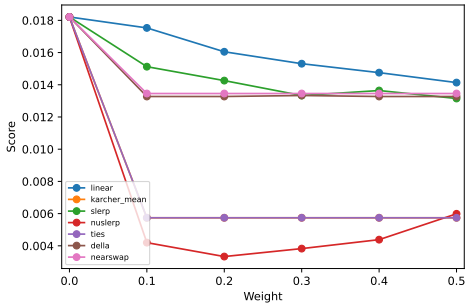
Figure 17: The HONEST evaluation results in LLAMA models.



(a) QWEN2-0.5B



(b) QWEN2-1.5B



(c) QWEN2-7B

Figure 18: The HONEST evaluation results in QWEN models.

The Propositional Idea Densities of Different Languages in Multi-Lingual Parallel Corpus

Yuka Kaise, Yuto Tsuchiya, and Masanori Oya
Graduate School of Global Japanese Studies, Meiji University
cu245002@meiji.ac.jp, cu245001@meiji.ac.jp
masanori_oya2019@meiji.ac.jp

Abstract

This study reports the propositional idea densities (PIDs) of different languages in parallel corpus in order to investigate whether these densities can function as language-independent measures of syntactic characteristics of sentences. The calculation is based on the Universal Dependencies annotations of dependency types in Parallel Universal Dependencies (PUD), a multi-lingual parallel corpus, and the results show a variety of PIDs across the languages in PUD, which reflect typological variations of information packaging across languages. Some issues of PID also have been pointed out for future research.

1 Introduction

This study reports the propositional idea densities (hereafter, PIDs) of different languages in parallel corpus in order to investigate whether PIDs can function as language-independent measures of syntactic characteristics of sentences. PIDs have been studied using the data of English as a measure of readability and as an indicator of future dementia, yet PID of other languages has not yet been conducted extensively. This study is the first attempt to investigate the PIDs across different languages based on the data of multi-lingual parallel corpus.

2 Previous studies on PID

PID is firmly grounded in well-established psychological theory. Within psycholinguistics, the proposition is considered the basic unit underlying text comprehension and memory (Kintsch & Keenan, 1973). A proposition may comprise diverse linguistic constituents—adjectives, adverbs, verbs, prepositions, and conjunctions—and PID is computed by dividing the number of propositions in a sentence or text by its total word count (Snowdon et al., 1996).

The PID construct has three principal functions: assessing textual readability, forecasting later dementia risk, and gauging sentence complexity in second-language acquisition (SLA) research. With respect to readability, Kintsch and Keenan (1973) showed that passages with lower PID scores are more readily recalled, underscoring the metric's relevance to ease of reading. Extending this work, Covington (2008) compared PID values across genres and observed that introductory and technical documents typically fall below 0.5, whereas research articles display a broader distribution. Such variability in research papers likely stems from their dual role: introducing novel concepts, like introductory texts, while simultaneously conveying detailed technical information, akin to technical documents.

PID has also been investigated as an indicator of future cognitive decline. Empirical evidence indicates that reduced PID in an individual's language output may presage the later emergence of dementia or Alzheimer's disease. In the longitudinal "Nun Study", Snowdon et al. (1996) analyzed autobiographical essays written in early adulthood and found that participants with lower PID scores were more prone to develop Alzheimer's disease five decades later. These results suggest that higher PID scores may signal a preserved capacity to handle syntactically complex structures and could therefore serve as an early marker of cognitive resilience. Similar results were observed in Kemper et al. (2001).

Given its theoretical grounding, PID has been adopted in SLA as an index of sentence complexity and, by extension, learner proficiency (Lopes & Pinto, 2022; Lunn et al., 2022, among others). Differences in learners' PID scores not only reveal variation in the syntactic complexity of their output but may also mirror underlying cognitive capacities for processing complex structures; nonetheless, using PID to predict future neurological decline lies

outside its intended scope.

Notwithstanding its promise, the cross-linguistic study of PID remains sparse. Most existing investigations rely on English data, and systematic analyses of PID in other languages are still lacking. Consequently, our understanding of how PID might operate in linguistic contexts beyond English is limited. Should translations conveying the same meaning exhibit comparable PID values across languages, it would imply that PID functions as a language-independent measure of sentence complexity. Conversely, substantial cross-linguistic divergence in PID for equivalent sentences would suggest that the metric's applicability may be confined to English.

3 This study

3.1 Research questions

This study aims to address the issue explained in the previous section, that is, the lack of cross-linguistic study of PID as a measure of sentence complexity. The research question of this study is as follows:

1. Do sentences of different languages with the same meaning share the same PID?
2. If their PIDs are varied across different languages, what are the cause(s) of the variations of PIDs?

If the answer to the question (1) is affirmative, then PID can be considered as a measure of sentence complexity which can be applied to a variety of languages. If it is negative, then we need to address the question (2) from the viewpoint of typological variations of languages.

3.2 Data

This investigation draws on the Parallel Universal Dependencies Treebanks (PUD) (de Marneffe et al. 2006, 2008; MacDonald et al. 2013; Petrov et al. 2013; Tsarfaty 2013; Zeman 2008; Zeman et al. 2017). Comprehensive documentation of the resource is provided on the CoNLL-2017 shared-task website, "Multilingual Parsing from Raw Text to Universal Dependencies" (http://universaldependencies.org/conll17/).

PUD encompasses 21 languages—Arabic, Czech, Mandarin Chinese, English, Finnish, French, German, Galician, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish—each represented by 1,000 sentences that are translations of an identical set of English source texts. The sentences were first morpho-syntactically annotated by Google and subsequently converted to the Universal Dependencies (UD) scheme in accordance with version 2 guidelines by members of the UD community in the CoNLL-U format.

For example, an example sentence "David has been writing several articles on Dependency Grammar and syntactic complexity." is annotated with Universal Dependencies as follows in Table 1 (some annotations have been deleted for the sake of simplicity):

1	David	David	NOUN	4	nsubj
2	has	have	AUX	4	aux
3	been	be	AUX	4	aux
4	writing	write	VERB	0	root
5	several	several	ADJ	6	amod
6	articles	article	NOUN	4	obj
7	on	on	ADP	9	case
8	Dependency	dependency	NOUN	9	compound
9	Grammar	grammar	NOUN	6	nmod
10	and	and	CCONJ	12	cc
11	syntactic	syntactic	NOUN	12	compound
12	complexity	complexity	NOUN	9	conj
13	.	.	PUNCT	4	punct

Table 1: The simplified UD annotation on "David has been writing several articles on Dependency Grammar and syntactic complexity."

Each column contains the following information from the lest to the right: (1) the order of the words in the sentence; (2) the words in the sentence; (3) the lemma (the dictionary form) of these words; (4) the parts of speech of the words; (5) the dependency head of each word; and (6) the dependency type. The first raw reads "The 1st word of this sentence is "David," which has the dictionary form "David," whose part of speech is NOUN; it depends on "writing," the 4th word of this sentence, and its dependency type is *nsubj* (nominal subject).

One of the characteristics of UD is that it focuses on the dependencies among content words and function words are all dependent on content words. For example, auxiliaries are dependent on the verbs which they add modal meanings, and prepositions are dependent on the nouns which follow them. In the example above, "has" and "been" are dependent on "writing" with the dependency type "aux," and "on" is dependent on "Grammar" with the dependency type "case." UD has chosen this annotation policy based on the insight that the meaning expressed by function words in certain languages (e.g., English) can be expressed not by

	ar	ch	cz	de	en	es	fi	fr	gl	hi	id
acl	34	20	112	18	193	116	223	150	429	338	246
acl:relcl	320	448	239	271	211	244	227	226	0	215	511
advcl	316	516	189	223	292	180	283	218	211	200	369
advmod	448	1225	661	1124	845	823	872	865	804	381	971
amod	1620	419	1817	1100	1348	1311	910	1394	1286	1412	585
ccomp	287	403	172	169	135	148	167	174	184	153	97
compound	386	1777	21	369	864	209	181	0	23	1277	35
conj	661	383	731	841	635	656	688	651	653	600	664
csubj	57	72	57	28	27	41	2	23	32	0	25
case	3047	1665	1857	2055	2511	3696	318	3208	3652	4076	1865
nummod	150	809	319	227	195	191	312	218	310	279	359
parataxis	24	3	23	68	97	105	108	107	90	94	114
root	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
xcomp	152	537	248	190	271	470	159	396	263	584	225
sum	8502	9277	7446	7683	8624	9190	5450	8630	8937	10609	7066
all	20747	21415	18463	21332	21126	23751	15813	24369	23309	23725	19858
PID	0.410	0.433	0.403	0.360	0.408	0.387	0.345	0.354	0.383	0.447	0.356

	is	it	ja	kr	pl	pt	ru	sv	th	tr
acl	171	208	1100	0	249	180	256	132	976	515
acl:relcl	303	241	0	1188	179	233	160	301	613	0
advcl	235	250	916	999	176	119	197	341	341	455
advmod	847	777	314	593	484	768	909	887	1190	574
amod	881	1395	84	208	1423	1328	1791	1253	654	1318
ccomp	124	137	75	68	85	119	131	122	275	173
compound	174	48	3061	2359	0	20	9	263	1927	519
conj	746	662	549	409	711	647	695	658	662	696
csubj	43	34	8	19	5	29	48	35	49	92
case	2132	3443	6496	404	1996	3604	2121	2225	2413	692
nummod	269	202	432	487	86	201	183	275	372	268
parataxis	85	99	0	0	0	102	195	134	5	15
root	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
xcomp	288	249	0	0	198	387	331	230	1070	127
sum	7298	8745	14035	7734	6592	8737	8026	7856	11547	6444
all	18835	23570	28788	16488	18384	23277	19355	19076	22289	16720
PID	0.387	0.371	0.488	0.469	0.359	0.375	0.415	0.412	0.518	0.385

Table 2: The frequencies of propositional dependency types and PIDs of the 21 languages in PUD. Abbreviations: ar: Arabic; ch: Chinese; cz: Czech; de: German; en: English; es: Spanish; fi: Finnish; fr: French; gl: Galician; hi: Hindi; id: Indonesian; is: Icelandic; it: Italian; ja: Japanese; kr: Korean; pl: Polish; pt: Portuguese; ru: Russian; sv: Swedish; th: Thai; and tr: Turkish.

independent words but by morphemes in content words in other languages (e.g., Russian). Focus on the dependencies among content words allows UD to capture cross-linguistic parallelism of the dependencies among them.

Because the dataset consists of semantically aligned translation pairs, cross-linguistic syntactic variation—including differences in dependency distances—can be analyzed while holding meaning constant.

3.3 Method

PIDs of the sentences in PUD are calculated based on the distinctions between propositions and non-propositions according to the type of dependency with which each word in a sentence depends on another in the same sentence. Propositions in this study are those words that depend on other words with the following 14 dependency types: *acl* for the verbs in adjectival participle clauses, *acl_relcl* for the verbs in relative clauses, *advcl* for the verbs in adverbial clauses, *advmod* for adverbs, *amod* for adjectives, *ccomp* for the verbs of clausal complements, *compound* for nominal compounds, *conj* for conjunctions, *csubj* for clausal subjects, *case* for prepositions, *nummod* for modifications by numericals, *parataxis* for paratactic phrases, *root* for the main verb of a clause, and *xcomp* for the verbs in external complements, which are those whose

subjects are either the subject or the object of the main clause. Words with these dependency types are expected to cover those defined as propositions according to Snowdon et al. (1996), which are adjectives, adverbs, verbs, and conjunctions. For each of the 21 languages in PUD, the number of these dependency types are calculated, then it is divided by the sum of the dependencies to obtain the PID of the language.

3.4 Results

Table 2 summarizes the result of calculating PIDs of the 21 languages in PUD. The mean of the PIDs of these 21 languages is .403, and their SD is .046. Approximately 40% of all tokens in PUD encode propositional content. The correlation between the total word counts and the PIDs of these languages is weak (r = 0.266). The top 3 PIDs are Thai, Japanese, and Korean, and the bottom 3 are Finnish, French, and Indonesian. The most cross-linguistically frequent dependency type is *case* (mean: 2546.48), which is followed by *amod* and *advmod*. The most cross-linguistically variable type is also *case* (SD: 1393.27), followed by *compound* and *amod*.

These frequent dependency types seem to have language-specific influence on the PID of a given language. For example, *case* is about 46% of the propositions in Japanese, while it is about 6% of those in Finnish; *compound* is about 21% in Japanese, while about 3% in Finnish, and 0% in French.

3.5 Discussions

PID reflects the density of propositional content encoded in a language. A higher PID of a language may indicate that it compresses more semantic content into more propositions (verbs, adjectives, adverbs, or conjuncts), which can show its structural compactness. Specifically, languages with frequent use of prepositions or postpositions (e.g., Thai and Japanese) show higher PIDs, while languages in which grammatical relations are expressed by inflections (e.g., Finnish, Turkish) show lower PIDs. This suggests that PIDs of different languages indirectly capture their typological variation. As such, PID allows us to measure how different languages distribute syntactic and semantic load. This makes PID a practical tool for comparing the surface density of propositional content across languages with different morphosyntactic strategies.

We need to point out some limitations on PIDs as a measure for syntactic characteristics of sentences: First, as the name PID indicates, it is a measure of density, not structural depth or complexity per se. This means that a high PID does not necessarily indicate a more complex syntax; rather, it may simply reflect fewer function words or more compact morphosyntax.

Second, PID should not be considered as a measure of sentence *complexity*, because it does not necessarily focus on the embeddedness of the syntactic structure, which is one of the factors of syntactic complexity. It is true that a sentence has a larger number of propositions if it contains many adverbial clauses and relative clauses (hence, more embedded) because these clauses contain verbs and possibly more adjectives, adverbs and prepositions, yet this also means it contains a larger total word count, hence its PID does not increase.

Third, the issue of annotation bias must be addressed across a variety of languages. PID is heavily influenced by the segmentation of sentence strings into words and annotation of them with parts of speech tags. The results that Japanese or Thai appear denser than others in PUD may be due to the annotation policy that postpositions or compound markers are counted as separate tokens, and we can expect that parallel corpora with different annotation policies may yield different results of PIDs for them.

4 Conclusion

This study reported the propositional idea densities (PIDs) of different languages in parallel corpus in order to investigate whether these densities can function as language-independent measures of syntactic characteristics of sentences. The calculation is based on the Universal Dependencies annotations of dependency types in Parallel Universal Dependencies, a multi-lingual parallel corpus, and the results show a variety of PIDs, which reflect typological variations of information packaging across languages. Three issues (PID not as a syntactic complexity measure, lack of consideration on the embeddedness, and annotation biases) have been raised for future research on propositional idea densities for characterizing syntactic properties of sentences in natural languages.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K04089.

References

Michael Covington. 2009. Idea Density — A Potentially Informative Characteristic of Retrieved Documents. *IEEE Southeastcon 2009*. https://ai1.ai.uga.edu/caspr/Covington-2009-Idea-Density-paper-SEC09-060.pdf

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of LREC*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Susan Kemper, Lydia Greiner, Jane Marquis, Katherine Prenovost and Tracy L Mitzner. 2001. Language decline across the life span: Findings from the nun study. *Psychology and Aging*, 16(2), 227–239. https://doi.org/10.1037/0882-7974.16.2.227

Walter Kintsch and Janice Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, 257–274.

Ângela Filipe Lopes and Maria da Graça Lisboa Castro Pinto. 2022. Assessing L2 Portuguese writing: idea density and sentence complexity. *Signo*, 47(88), 73–86.

Andrew M Lunn, Daniel Matthias Bürkle, Rebecca Ward, Alice P McCloskey, Adam Rathbone, Aaron Courtenay, Rachel Mullen, Andrea Manfrin. 2021. Spoken propositional idea density, a measure to help second language English speaking students: A multicentre cohort study. *Medical Teacher*, 44(3), 267–275. DOI: 10.1080/0142159X.2021.1985097

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. *Proceedings of ACL*.

Marije C Michel, Folkert Kuiken and Ineke Vedder. 2007. The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45, 241–259.

James R. Miller and Walter Kintsch. 1980. Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 335–354.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. *Proceedings of LREC*.

David A. Snowdon, Susan Kemper, James Mortimer, Lydia Greiner, David R. Wekstein, and William R. Markesbery. 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA*, 275, 528–532.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. *Proceedings of ACL*.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of LREC*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver.

Scaffolded AI Feedback for L2 Writing: Fostering Self-Correction in Japanese University Students

Atsushi Nakanishi
Osaka Institute of Technology
atsushi.nakanishi@oit.ac.jp

Abstract

This study introduces AI-assisted Translation Learning and Search (ATLaS), an AI-powered system designed to enhance Japanese university students' English writing through translation-based learning. Grounded in the constructivist learning theory and Vygotsky's Zone of Proximal Development, the system uses scaffolded feedback using Ferris's hierarchical error taxonomy, promoting metalinguistic awareness through staged intervention rather than direct correction. A case study of 26 Japanese university students demonstrated significant improvements, with holistic writing scores increasing from 2.3 to 3.3 points (43% improvement). The system identified an average of 6.7 corrections per student, with 67% of the flagged errors being independently corrected by the learners. The error analysis of 175 instances revealed distinctive patterns: elevated lexical errors driven by word-choice difficulties, in contrast to reduced morphological errors in traditional contexts. The results suggest that AI-assisted feedback systems effectively supported L2 writing development when incorporating appropriate pedagogical scaffolding.

1 Introduction

1.1 Automated written corrective feedback

Traditional L2 writing instruction struggles to provide detailed, individualized, and timely feedback due to large class sizes and limited teacher availability. Although teacher feedback remains valuable, it is often not sufficiently scalable to guide learners through the recursive processes of drafting, reflection, and revision. This pedagogical gap highlights the potential for technology-enhanced learning environments that supplement conventional instruction. Sophisticated AI, particularly large language models (LLMs), enables immediate, data-driven, and personalized feedback (Mizumoto, 2025; Woo et al., 2024).

Although students often perceive teacher feedback to be authoritative and reliable, it has inherent limitations. Teachers, frequently constrained by time and large class sizes, tend to focus on local issues, particularly grammar, and sometimes neglect higher-order concerns such as content and organization. This has generated interest in automated writing evaluation (AWE) systems that provide automated written corrective feedback (AWCF) to reduce teacher workload and offer students immediate support (Feng et al., 2025).

However, AWCF's pedagogical effectiveness remains debated, with complex implementation challenges. Some empirical studies have indicated that the AWCF does not necessarily improve writing quality. For example, Fan (2023) found no significant difference in writing quality between lower-proficiency EFL students who received combined Grammarly and teacher feedback and those who received teacher feedback alone. This lack of improvement can be attributed to several factors, including learners' low proficiency level, which hinders their ability to understand and process feedback, and their general unfamiliarity with AWE tools.

Beyond feedback accuracy and comprehensibility, a more critical pedagogical concern is the risk of learners developing over-reliance on automated systems. For instance, Karatay and Karatay (2024) highlighted that students may develop trust levels that lead them to accept automated suggestions uncritically, without engaging in thoughtful analysis and deliberation essential for skill development. This passive "correction" behavior can inhibit the growth of learners' autonomy and their ability to self-edit.

These limitations indicate that neither teachers nor automated feedback can provide a complete solution. Instead, a growing body of research has identified the importance of an integrated approach that leverages the strengths of both approaches.

A quasi-experimental study by Cheng and Zhang (2024) demonstrated the potential synergy of AWE-teacher integrated feedback. In their model, the AWE system first addressed local-level language errors, allowing teachers to focus their feedback on global issues such as content and organization. This integrated approach not only led to significantly greater improvements in all aspects of writing performance compared to teacher feedback alone but also promoted a deeper behavioral and cognitive engagement from students. By creating a writing-feedback-revision cycle, such a model helps learners move beyond the belief that writing is a one-time task and, thus, recognize the importance of revision.

This study suggests that the most promising path forward lies in a thoughtful human–machine partnership. The goal is not to replace the expertise of human instructors, but augment their capabilities, thus creating a more efficient and effective feedback ecosystem. The AI-assisted Translation Learning and Search (ATLaS) system proposed in this study is based on this principle. It is designed to address the shortcomings of conventional AWCF by providing a scaffolded learning environment that not only offers corrective feedback but also supports the deeper learning processes necessary for long-term writing development.

1.2 Error analysis

The systematic analysis of learner errors constitutes a fundamental component of second language acquisition research, facilitated by the development of large-scale learner corpora and assessment datasets, such as the ICNALE Global Rating Archives (Ishikawa, 2023). These resources enable a comprehensive investigation of the linguistic characteristics of learners' language. Recent advances in AI have further enhanced this field through automated error analysis tools, such as the Auto Error Analyzer (Mizumoto, 2025), which automates accuracy metric calculations in learner texts. These developments underscore the necessity for a systematic and theoretically grounded classification framework to effectively categorize learner errors.

For the development of the ATLaS system, a robust error taxonomy is essential to provide structured and comprehensible feedback to learners. As such, this study adopted the comprehensive error classification framework proposed by Ferris (2011), which is widely recognized for its application in analyzing and treating errors in L2 students' writing.

Error Category	Subcategory
Morphological	Verbs [Tense, Form, Subject-verb agreement], Nouns [Articles/determiners, Noun endings (plural/possessive)]
Lexical	Word choice, Word form, Informal usage, Idiom error, Pronoun error
Syntactic	Sentence structure, Run-ons, Fragments
Mechanical	Punctuation, Spelling
Miscellaneous	Unclassified errors

Table 1: Error classification framework.

This framework organizes errors into five primary domains integrated into the ATLaS error analysis engine, as detailed in Table 1.

By operationalizing this established taxonomy, ATLaS was designed to provide learners with feedback that is both accurate and pedagogically organized, enabling them to understand the specific nature of their errors and facilitate targeted improvement strategies.

1.3 Research objectives

This study introduces the ATLaS system. This system was designed not only to correct errors but also to promote deeper metalinguistic awareness and encourage self-regulated learning. By leveraging a powerful AI model, ATLaS provides users with detailed feedback on their Japanese-to-English translations and classifies errors into a hierarchical system of grammatical, lexical, structural, and stylistic categories.

The development of the system serves two primary purposes: (1) the enhancement of self-correction abilities in translation learning through a gradual feedback provision system, and (2) the verification of error-type specific learning support effectiveness. This study also reports on the implementation of ATLaS with 26 Japanese university students in a classroom setting, examining its impact on translation accuracy and learner engagement in error-correction processes.

2 System development

2.1 ATLaS design

ATLaS was developed in accordance with the constructivist learning theory and Vygotsky's Zone of Proximal Development (ZPD), wherein AI-generated feedback functions as a mediating instrument to facilitate the transition between students' existing translation competencies and their poten-

tial performance. Instead of offering instantaneous corrections, the system employs a staged intervention approach that fosters reflective self-correction processes. This methodology aligns with existing research on indirect corrective feedback, which enhances reflective thinking and problem-solving abilities.

ATLaS consists of two fundamental operational modes, each tailored to address distinct facets of translation learning: the Translation Feedback Mode and the Error Search Mode. The former offers structured AI-mediated correction assistance for individual translation assignments, while the latter allows learners to investigate aggregated error patterns and examples organized in accordance with established linguistic taxonomies. These complementary modes function synergistically to facilitate immediate learning requirements and foster long-term metalinguistic development. The system leverages GPT-4 as its core language processing engine, selected for its advanced multilingual capabilities and JSON-structured output support. The model operates with a temperature setting of 0.7 to balance creativity and consistency in feedback generation, with a maximum token limit of 4,000 to ensure comprehensive explanations while maintaining processing efficiency.

2.2 Translation Feedback Mode

The Translation Feedback Mode facilitates structured correction, helping learners progressively identify and address errors. This mode functions via a systematic workflow that includes text input, AI-assisted analysis, structured feedback provision, and revisions initiated by the learner.

Figure 1: Interface of the Translation Feedback Mode.

The system analyzes Japanese source texts and English translations using GPT-4 to identify up to 10 significant issues. Carefully engineered prompts establish the AI as a bilingual instructor prioritizing educational scaffolding over direct correction. The prompt engineering incorporates two key components: (1) a system prompt that defines the AI's instructional role and output format requirements, and (2) a user prompt that provides the specific translation task context. For example, the system prompt instructs the AI to "identify up to 10 important issues" and constrains error classification to predefined taxonomic categories, ensuring consistency with Ferris (2011) framework. The error type constraint is implemented through explicit enumeration: "The 'error_type' field MUST be exactly one of the predefined error types. Do not create new error type names."

A typical user prompt structure follows this template: "Please evaluate the following translation based on the instructions. Japanese Text: ```[source text]``` English Translation: ```[student translation]```" This format provides clear task boundaries while maintaining consistency across all system interactions.

The system generates structured JSON responses that consist of three components: marked_text, feedback_message, and correction_table. The structure of the correction_table enables comprehensive feedback delivery through the following components (Table 2).

This structured format enables the system to present feedback in a pedagogically organized manner, with each correction including contextual information, explicit error categorization, and scaffolded guidance questions.

Component	Description
japanese	Complete original Japanese sentence providing source context
original	User's complete original English sentence for comparison
correction	Proposed correct English sentence demonstrating target form
explanation	Detailed explanation in Japanese clarifying linguistic principles
error_type	Predefined error classification based on Ferris taxonomy
prompting_question	Guiding question in Japanese with error marker references

Table 2: Correction_table structure.

The system generates numbered markers within the original translation text, thereby establishing

clear visual connections between the identified issues and their corresponding feedback elements, as shown in Figure 2. This marking system enables learners to focus their attention on specific problematic segments while maintaining overall text coherence and understanding the translation.

The hint-generation mechanism produces culturally appropriate prompting questions in Japanese, which assist learners in self-correction without explicitly providing answers. These questions were designed to elicit relevant linguistic knowledge, while fostering metacognitive reflection on translation decisions. For example, verb tense errors may prompt questions regarding the temporal relationships between events, whereas article errors may pertain to patterns of noun countability and definiteness.

The correction submission interface illustrated in Figure 2 provides learners with an editable version of their original translation, facilitating direct text modification while maintaining the associations with the original error markers. This approach enables focused attention on the identified issues while concurrently allowing for a comprehensive revision of the entire translation. The system uses both the initial and revised translation versions for comparative analysis and learning assessment.

Upon correction submission, ATLaS provides comprehensive feedback, including corrected versions of each identified error, detailed explanations of the underlying linguistic principles, and error-type classifications. The explanations are provided in Japanese to ensure accessibility and comprehension by integrating examples and contrasts that clarify the rationale for the suggested corrections.

The feedback presentation shown in Figure 3 employs a structured table format that delineates the original text, corrections, and explanations, enabling systematic comparison and analysis. Error-type classification enables learners to recognize patterns in their translation challenges and develop targeted improvement strategies for future learning activities.

2.3 Error Search Mode

The Error Search Mode organizes accumulated error patterns using Ferris (2011) framework, enabling systematic error exploration for skill development. Users interact with the system via a hierarchical interface and select the main error categories from a dropdown menu that updates subcategory options contingent on the available data. Upon the

Figure 2: Feedback interface.

selection of specific error types, the system queries its database of correction instances and presents comprehensive concordance data.

The search results present the Japanese source text, original erroneous translations, corrected versions, and detailed explanations in a structured format (Figure 4). This systematic structure facilitates pattern recognition across multiple instances of similar errors while promoting both individual and collaborative learning through shared error analysis.

Each result entry provides comprehensive contextual information, enabling learners' comprehension of not only error identification but also the rationale for specific corrections within particular contexts. The concordance display presents multiple examples of similar error types, allowing learners to identify common patterns and the underlying

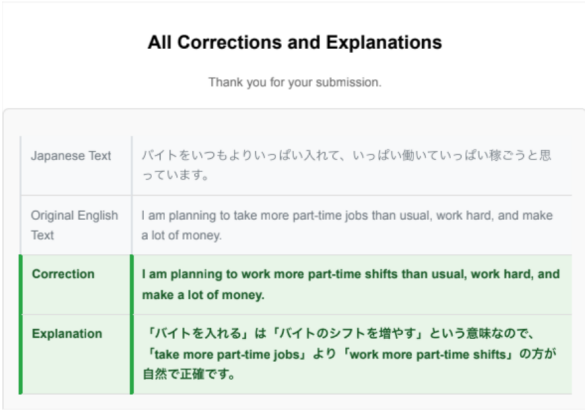


Figure 3: Final feedback interface.

Figure 4: Search results interface of "Tense."

linguistic principles.

This mode supports autonomous learning by enabling learners to search for personally encountered error types to reinforce their learning experience or explore unfamiliar categories to develop broader linguistic awareness. By providing access to the accumulated error patterns, the Error Search Mode reduces the dependence on immediate feedback and simultaneously promotes independent learning capabilities and metalinguistic awareness essential for long-term language development in line with the system's pedagogical goals.

3 Methodology

3.1 Research design

This study employed a single-group case study design to investigate the effectiveness of ATLaS in enhancing the English writing performance of Japanese university students. The design was deliberately selected to examine the learning processes and system interactions within an authentic classroom context, thereby facilitating the comprehensive documentation of improvement patterns and error-correction behaviors.

The study was conducted over two months (May–June 2025) with 26 Japanese university students enrolled in an "English Usage" course. A pre-post design was implemented to measure improvements in writing quality using holistic scoring rubrics, and a systematic error analysis was conducted on 175 correction instances categorized according to Ferris (2011) taxonomic framework.

Data collection focused on quantitative measures, including: (1) holistic writing scores using a five-point rubric, (2) correction frequency per student, and (3) error type distribution across morphological, lexical, syntactic, and mechanical categories. The qualitative analysis examined error patterns and improvement trajectories based on the systematic content analysis of the translation samples.

Ethical considerations were addressed through integrated informed consent procedures within the web application interface to ensure voluntary participation without academic coercion.

Several limitations should be noted. The single-group design limits causal inferences, as improvements may reflect the combined effects of AI feedback and concurrent instructor guidance rather than that of ATLaS alone. Additionally, the single writing topic constrains generalizability across different genres and discourse types.

3.2 Participants

The study employed a convenience sampling methodology by recruiting participants from an existing "English Usage" class at a private Japanese university. Initially, 28 third-year students who enrolled in the course were invited to participate. However, the final sample consisted of 26 students due to attrition: one student discontinued class attendance during the study period and another failed to submit the required assignments. An attrition rate of 7.1% was considered acceptable for educational research.

Participants' English proficiency levels were assessed using TOEIC Listening and Reading scores, which provide standardized measures of language ability. The proficiency distribution revealed considerable variation, with scores ranging from 165 to 635. The mean score was 377.5 points (SD = 125.84), whereas the median score was 367.5

points. This distribution indicates a predominantly intermediate-low to intermediate proficiency level, with some participants demonstrating more advanced abilities.

All participants were native Japanese speakers studying at a private university where instruction is conducted primarily in Japanese. The students were from the Faculty of Information Science, indicating that English was not their primary academic focus, but rather a supplementary skill requirement.

3.3 Implementation procedures

The implementation followed a four-session protocol advancing translation learning through AI feedback. Students worked on "Summer Vacation Plans" to maintain consistency while encouraging natural expression.

Session 1 entailed the development of original Japanese compositions (maximum 500 characters) alongside engaging in pre-editing activities to clarify ambiguous expressions and simplify complex structures. Session 2 focused on translating pre-edited texts into English using Microsoft Word, with dictionary access permitted; however, the use of AI translation tools prohibited the assessment of authentic linguistic competence.

Session 3 implemented a core intervention using the Translation Feedback Mode. Students submitted their initial translations and received AI-generated scaffolded feedback, which included error markers, contextual hints, and pedagogical explanations. Following the review of the feedback, the students revised their translations and resubmitted the corrected versions through the platform.

Session 4 used the Error Search Mode, enabling students to explore systematic error patterns in their work and in peer examples. This analytical phase developed students' metalinguistic awareness of common translation challenges while familiarizing them with a comprehensive error-classification framework.

Throughout the process, the students submitted four text versions: the original Japanese text, the pre-edited Japanese version, the initial English translation, and the ATLaS-revised English text. Instructor feedback complemented the system's local linguistic focus by addressing global issues, including content organization, coherence, and communicative effectiveness, thus creating an integrated feedback environment that addresses both surface-level accuracy and higher-order writing concerns.

3.4 Data collection and analysis

Data collection used multiple mechanisms to capture information about learning processes and system effectiveness. The primary data source consisted of automatically logged user interactions, including original Japanese texts, initial English translations, AI-generated correction feedback, and final revised versions. The quantitative metrics focused on measurable learning outcomes and system usage patterns. The key variables included (1) the number of corrections per student, (2) the distribution of error types according to Ferris (2011) taxonomy, and (3) holistic writing quality scores using a five-point rubric.

Writing quality was assessed using a standardized five-point holistic scoring rubric administered through OpenAI's GPT-4, evaluating both the initial and revised translations for grammatical accuracy, lexical appropriateness, syntactic complexity, and overall coherence. To determine if the change in writing quality was statistically significant, a paired-samples t-test was used to compare the holistic scores before and after the intervention.

A qualitative analysis was conducted through the systematic content analysis of the translation samples, focusing on error categorization and improvement patterns. A hierarchical error taxonomy based on Ferris (2011) framework categorized 175 correction instances into four primary domains. The error pattern analysis employed frequency distribution comparisons with established research and qualitative examinations of the characteristic difficulties faced by Japanese learners.

4 Results

4.1 Learning effectiveness

The statistical analysis demonstrated consistent improvement patterns across multiple writing quality dimensions. The mean number of corrections per student was 6.7 (SD = 2.4, range = 2–10), indicating that the system successfully identified meaningful improvement opportunities for learners across different proficiency levels. The distribution of correction frequencies showed that 46.2% of the participants received 4–6 corrections, which indicates an optimal cognitive load for learning effectiveness.

A quantitative assessment of learning effectiveness revealed substantial improvements across multiple writing quality dimensions. Comparative assessment of initial and revised translations using a five-point holistic scoring rubric administered

Error Type	Ferris (2011)	ATLaS %(n)
Morphological Errors		
Verbs		
Tense	10.9	5.1 (9)
Form	7.8	0.6 (1)
Subject-verb agreement	2.9	1.7 (3)
Nouns		
Articles/determiners	6.6	6.3 (11)
Noun endings	8.9	0.6 (1)
Lexical Errors		
Word choice	11.5	44.0 (77)
Word form	6.5	4.6 (8)
Informal usage	0.3	0.0 (0)
Idiom error	0.8	0.6 (1)
Pronoun error	2.9	0.6 (1)
Syntactic Errors		
Sentence structure	22.5	30.9 (54)
Run-ons	2.9	0.0 (0)
Fragments	1.8	1.1 (2)
Mechanical Errors		
Punctuation	6.8	0.6 (1)
Spelling	5.9	2.3 (4)
Miscellaneous	0.9	1.1 (2)

Table 3: Error distribution comparison.

through OpenAI's GPT-4 showed significant advancement from a mean initial score of 2.3 to a mean revised score of 3.3, representing a 43% improvement in the overall writing quality. A paired-samples t-test was conducted to verify the significance of this improvement. The results confirmed that the increase in scores was statistically significant, $t(25) = 6.43, p < .001$.

The analysis identified distinctive learning patterns based on initial proficiency levels. Advanced learners (initial scores of 4–5) demonstrated sophisticated refinements in stylistic choices and idiomatic expressions, whereas intermediate learners (scores of 2–3) showed substantial improvements in grammatical accuracy and sentence structure. Beginning learners (scores of 1–2) exhibited fundamental corrections in basic grammatical construction and vocabulary selection, although the improvement margins were modest.

4.2 Error pattern analysis

An analysis of 175 correction instances revealed distinct patterns that diverged from the established error distribution findings. The error distribution showed notable deviations from Ferris (2011) findings, reflecting the intersection of Japanese learners' characteristics, translation-based learning, and AI-mediated error detection (Table 3).

Most significantly, lexical errors dominated the ATLaS corrections (49.8%), substantially exceeding Ferris (2011) correction (22.0%). This was likely driven by word-choice difficulty (44.0% vs. 11.5%). For instance, the AI flagged subtle collocational errors typical for Japanese learners, such as correcting "I joined the exam" to "I took the exam." This suggests that AI systems are highly sensitive to semantic nuances that human instructors may overlook, based on leveraging vast linguistic databases to detect contextually inappropriate vocabulary.

Syntactic errors were the second-largest category (32.0%), approximating Ferris (2011) findings (27.2%). However, the AI captured different phenomena. For example, it corrected nuanced prepositional choices, such as changing "finish the work until tomorrow" to "by tomorrow," which affects grammatical precision more than immediate comprehensibility. This highlights AI's systematic identification of structural deviations, whereas instructors may prioritize communicative effectiveness.

Conversely, morphological errors showed a markedly lower frequency (14.3%) than in Ferris (2011) study (37.2%). This reduction may reflect the pattern-recognition capabilities of AI in identifying morphological consistency, in which source-text cues facilitate accurate grammatical choices. Mechanical errors also showed a substantial reduction, which was attributable to the digital writing environment and sophisticated AI checking of spelling.

These findings suggest that AI-mediated detection produces different error distributions compared to traditional human analysis, emphasizing the need for pedagogically informed AI training that balances systematic accuracy with communicative priorities in L2 writing instruction.

5 Conclusions

5.1 Summary

This study provides preliminary evidence that the ATLaS system may enhance Japanese university students' English writing performance through structured AI-mediated feedback. The scaffolded learning approach, grounded in the constructivist learning theory and Vygotsky's ZPD, appeared to promote metalinguistic awareness and increased learner autonomy among the 26 participants during the two-month intervention. Quantitatively, mean holistic scores increased from 2.3 to 3.3

(43% improvement), and a paired-samples t-test indicated this change was statistically significant, $t(25) = 6.43, p < .001$. However, because the study used a single-group pre–post design and instructor feedback was provided alongside ATLaS (see Methods 3.3), improvements cannot be unambiguously attributed to ATLaS alone. These results should therefore be interpreted as promising but preliminary, and future randomized controlled trials are required to isolate the specific effects of the system.

A systematic analysis of 175 correction instances revealed distinct error patterns. Lexical errors dominated the corrections (49.8%), substantially exceeding Ferris (2011) reported frequency of 22.0%, primarily driven by word-choice difficulties (44.0% vs. 11.5%). Syntactic errors constituted the second largest category (32.0%), whereas morphological errors showed a markedly lower frequency (14.3%) than in Ferris (2011) study (37.2%).

These findings suggest that AI-assisted feedback systems can contribute to L2 writing development when designed with appropriate pedagogical scaffolding. Nevertheless, given the single-group design and potential instructor–system interaction, further controlled research is needed to confirm causal mechanisms.

5.2 Further research directions

This study shows promising results for AI-assisted writing instruction, but several areas need further investigation. First, future research should include larger and more diverse participant groups to improve the generalizability of these findings. Randomized controlled trials would help isolate the specific effects of AI feedback and test the system with different writing tasks beyond personal narratives.

Second, longitudinal studies across multiple semesters are needed to determine whether the observed improvements persist over time. Such studies would reveal whether learners maintain self-correction skills after the intervention ends and how AI feedback affects long-term writing development.

Third, research should examine how cultural and institutional factors influence AI-assisted feedback effectiveness. Studies in different educational contexts would help us understand how various pedagogical approaches and technologies affect system adoption and success.

Finally, future versions of ATLaS should expand beyond Japanese-to-English translation to support multiple language pairs. This multilingual approach would make the system applicable to broader language learning contexts and enable comparative analyses across different first languages. Such expansion could lead to more inclusive instructional design and enhance the system's global relevance.

Acknowledgments

This research was supported by the Grant-in-Aid for Young Scientists (Grant Number: 23K12254).

References

Xiaolong Cheng and Lawrence Jun Zhang. 2024. Examining second language (L2) learners' engagement with AWE–teacher integrated feedback in a technology-empowered context. *The Asia-Pacific Education Researcher*, 33:1023–1035.

Ning Fan. 2023. Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *SAGE Open*, 13(2).

Haiying Feng, Kexin Li, and Lawrence Jun Zhang. 2025. What does AI bring to second language writing? a systematic review (2014–2024). *Language Learning & Technology*, 29(1):1–27.

Dana R. Ferris. 2011. *Treatment of Error in Second Language Student Writing*. University of Michigan Press, Ann Arbor, MI.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge, Abingdon, UK.

Yasin Karatay and Leyla Karatay. 2024. Automated writing evaluation use in second language classrooms: A research synthesis. *System*, 123:103332.

Atsushi Mizumoto. 2025. Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition*, pages 1–18.

David James Woo, Hengky Susanto, Chi Ho Yeung, Kai Guo, and April Ka Yeng Fung. 2024. Exploring AI-generated text in student writing: How does AI help? *Language Learning & Technology*, 28(2):183–209.

ScheduleMe: Multi-Agent Calendar Assistant

Oshadha Wijerathne[1], Amandi Nimasha[1], Dushan Fernando[1], Nisansa de Silva[1], and Srinath Perera[2]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{oshadha.20, amandi.20, dushan.20, NisansaDdS}@cse.mrt.ac.lk
[2]WSO2 LLC
srinath@wso2.com

Abstract

Recent advancements in LLMs have contributed to the rise of advanced conversational assistants that can assist with user needs through natural language conversation. This paper presents a *ScheduleMe*, a multi-agent calendar assistant for users to manage google calendar events in natural language. The system uses a graph-structured coordination mechanism where a central supervisory agent supervises specialized task agents, allowing modularity, conflicts resolution, and context-aware interactions to resolve ambiguities and evaluate user commands. This approach sets an example of how structured reasoning and agent cooperation might convince operators to increase the usability and flexibility of personal calendar assistant tools.

1 Introduction

The rapid advancements in natural language processing (NLP) and large language models (LLMs) have opened new opportunities for developing intelligent, user-friendly applications. Among these, conversational agents capable of understanding and acting upon human language inputs are becoming increasingly important for daily task management. Traditional scheduling systems often require rigid, form-based inputs and manual navigation, limiting user experience and efficiency. In contrast, leveraging LLMs enables the creation of systems that can interpret flexible, natural language instructions, offering a more intuitive and seamless interaction.

This research introduces a calendar management assistant built using the LangChain framework (Følstad and Skjuve, 2019) and OpenAI's GPT-4o mini model (OpenAI et al., 2024; Wang, 2025).The assistant enables users to manage their calendars through natural, conversational interactions.

The system adopts a modular, multi-agent architecture. Specialized agents handle specific operations such as scheduling, fetching, editing, and deleting events. A centralized supervisory chatbot coordinates these agents and manages the dialogue with the user. This separation improves modularity, reliability, and context-aware task execution.

The goal is to create an assistant that not only executes tasks accurately based on natural user requests, but also enhances transparency, reliability, and user satisfaction. Users will be able to manage their calendars simply by conversing with the chatbot, eliminating the need for complex interfaces or strict command formats. Through this research, our objective is to demonstrate how LLMs, when properly structured within a robust framework such as LangChain, can serve as powerful tools for building practical, real-world intelligent applications that align with human communication patterns.

2 Related Work

2.1 Limitations of Traditional Scheduling Dialogue Systems

Traditional task-oriented dialogue systems typically rely on intent classification and slot-filling methods (Surdeanu et al., 2011). These systems map user inputs to predefined actions and extract key details such as date, time, or participants.

Dialogue flows are often rigid and frame-based (Braggaar et al., 2024), collecting user input step-by-step. While effective in simple cases, this structure struggles when users provide information out of order, revise commands, or use unexpected phrasing. Systems such as *Calendar.help* (Cranshaw et al., 2017) have shown how these methods can be used in real-world scheduling applications, combining natural language understanding with backend tools such as webhook integrations. However, these systems often fail to handle ambiguous or incomplete inputs gracefully and are difficult to adapt to new use cases without retraining on labeled data. As a result, their user interactions can feel rigid and frustrating.

2.2 Advancing to LLM-Based Multi-Agent Systems

Earlier AI agents were designed using symbolic rules or simple learning methods were built to act independently, make decisions, and sometimes communicate with other agents (Farinetti and Canale, 2024; Alonso, 2002; Hazra et al., 2024). Although these early agents were effective in narrow tasks, they required a lot of manual programming and were not very flexible (Wang et al., 2024).

Recent advances in large language models (LLMs), such as GPT-4 (OpenAI et al., 2024; Qin et al., 2023), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023), have introduced new possibilities. These models can handle open-ended language tasks, reason about goals, and even use external tools when guided properly. Methods such as chain-of-thought prompting (Wei et al., 2022), in-context learning, and tool use through APIs have enabled LLM-based agents to solve more complex and varied problems (Warnakulasuriya et al., 2025).

Frameworks such as ReAct (Yao et al., 2023) and AutoGPT (Yang et al., 2023) have demonstrated how LLMs can be used as the 'brains' behind autonomous agents.However, many tasks such as scheduling or workflow management require multiple specialized agents to work together. A multi-agent setup allows for modular design, parallel task execution, and clearer delegation of responsibilities. Without structured coordination, these systems often struggle with effective communication, shared memory, and maintaining context, which limits their ability to handle complex or extended interactions.

2.3 Multi-Agent Approaches to Scheduling Assistants

Some recent systems split scheduling tasks into smaller parts, assigning different agents to handle event creation, editing, retrieval, and similar operations. *SmartCal* (Shen et al., 2024) improves tool use reliability and decision making through a self-aware supervisory framework, which coordinates agent actions, supports error recovery, and maintains context. Without such supervision, assistants often struggle with user corrections and overlapping goals during complex interactions.

Beyond orchestration, procedural and multi-step reasoning is crucial for advanced scheduling. ScriptWorld (Joshi et al., 2023) demonstrates how agents can learn and execute sequential tasks while maintaining state. Although it operates in simulation, the same principle applies to real-world scheduling, where agents must interpret goals, resolve conflicts, and complete tasks in the correct order.

To overcome the limits of static coordination, frameworks like LangGraph (Duan and Wang, 2024; Wang and Duan, 2024) provide graph-based, state-aware workflows. Nodes represent task-specific agents or tools, and edges define transitions based on system state, enabling conditional branching, iteration, and runtime adaptation. Integrating LangGraph with a central supervisory agent supports dynamic task execution, context maintenance, and natural conversational scheduling.

3 Methodology

3.1 Multi-Agent System Architecture

Having discussed prior work and current limitations in existing systems, we now present the architecture and design principles behind our proposed calendar assistant. Our system integrates large language model (LLM) based reasoning with a graph-driven orchestration framework using LangGraph, enabling dynamic coordination among agents.

The architecture is centered around a supervisory chatbot agent, which serves as the sole interface for user interaction and agent coordination. Upon receiving a user query, the supervisor agent interprets user intent and delegates tasks accordingly to one of the specialized functional agents. These include the scheduling agent, availability checking agent, event editing agent, and event deletion agent. All inter-agent communication is mediated through the supervisor agent. When a functional agent requires additional information to complete a task, it requests the chatbot to re-engage with the user to obtain the missing input.

Each agent follows the ReAct (Reasoning and Acting) paradigm, combining decision-making with the ability to invoke predefined tools. These tools are implemented as custom functions that interface with the Google Calendar API to perform specific actions. For instance, the availability checking tool queries calendar data for events within a given time range, while the scheduling tool creates new calendar events based on parameters such as title and time. Similarly, the editing and deletion tools update or remove events based on event ids and user-specified criteria. These tools

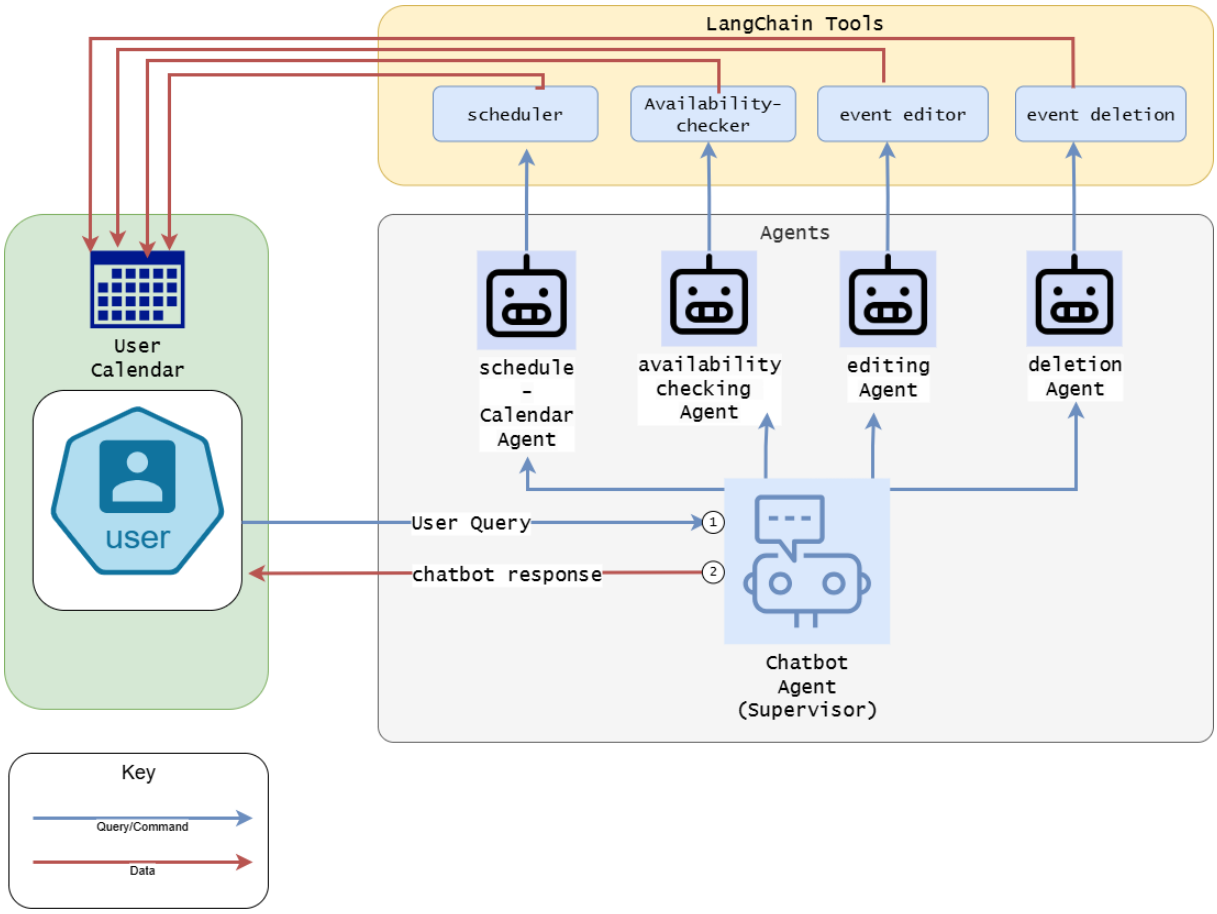


Figure 1: Multi-Agent System Architecture: All other agents are controlled by the *supervisor agent*, but we have opted not to draw the control and communication lines between the agents to reduce unnecessary clutter. When a command or a data item is relevant to all the entities in a parent entity, the relevant arrow terminates on the parent entity. Otherwise, it terminates on the specific relevant child entity. The numbers on the agents at arrow terminals indicate the order in which each action may happen in a typical execution.

abstract the underlying API calls, allowing agents to focus on high-level decision logic.

LangGraph is employed to structure the agent coordination process as a directed graph. In this configuration, each node represents an agent, and edges define the flow of execution based on the outcome of reasoning steps or user input. The graph structure enforces that communication paths flow through the supervisor agent, ensuring a controlled and interpretable interaction model. This setup allows the system to flexibly handle user queries in a stateful and modular manner. An overview of the system's architecture is depicted in Fig. 1, highlighting the agents and their interactions within the LangGraph execution framework.

3.2 Implementation Details

With the system architecture defined, we now describe how the AI Calendar Assistant is implemented in practice, detailing the technologies and components involved. The AI Calendar Assistant is implemented as a graph-based multi-agent system, where each node corresponds to an agent and each edge represents a possible transition in the task flow. The architecture is constructed using LangGraph's `StateGraph` module, which supports dynamic, stateful execution paths. The central supervisory chatbot agent initializes each interaction by processing user input and extracting intent, parameters, and task directives. These outputs determine the subsequent traversal of the graph and the activation of the appropriate functional agent.

The system integrates OpenAI's GPT-4o mini model via LangChain to perform natural language understanding and dialogue management. GPT-4 is known for high accuracy in complex reasoning but is resource-intensive and costly (Gunathilaka and de Silva, 2025; Siddiky et al., 2025). GPT-4o provides comparable performance, with stronger multilingual and multimodal capabilities, while offering

reduced latency and computational requirements (Siddiky et al., 2025; Zhang et al., 2024). Prior studies in dialogue system design have also employed GPT-4o mini as a reference model due to its extended (128k-token) context window, function-calling support, and low latency (Robino, 2025). In the medical domain, GPT-4o has demonstrated efficiency gains and near-human conversational response times, confirming its suitability for real-time applications (Zhang et al., 2024). The supervisor agent leverages this model to interpret user queries, ask for clarification when needed, and generate tailored instructions for each functional agent. The initiation of the supervisor is shown in Prompt 1. When activated, functional agents handle tasks such as scheduling, checking availability, editing events, or deleting them. These agents perform stateless operations: they receive a structured input payload, carry out the task, and return the results to the supervisor. The supervisor then communicates the outcome back to the user. We show all the functional agent prompts in Appendix A.

```
"""You are the Supervisor Agent for an AI
    Calendar Assistant system.

Current date and time: {current_date_time}.

Your Responsibilities:
- Talk to the user to fully understand their
    request.
- Collect **all required information**
    before sending a task to any agent.
- Send tasks to the correct agent with
    complete and clear information.
- Collect responses from agents and decide
    the next action.

Agents you can use:
- calendar_checker_agent: To check calendar
    events.
- event_scheduler_agent: To add new events (
    REQUIRES: event title, date, and time).
- event_remover_agent: To delete events.(
    Should Provide the event Id.)
- event_modifier_agent: To modify/edit/
    update events.
- user: If you need more information.

Important Rules:
1. Greet the user and ask what they want to
    do.
2. If user request is unclear or missing
    information, ask follow-up questions (
    one at a time) until you have
    everything needed.
3. Only send a task to another agent once
    you have **all required information**.
4. Be friendly, clear, and simple. Ask **one
    question at a time**.
5. Always format your reply in JSON:
   - 'next': agent to call ('
       calendar_checker_agent', '
       event_scheduler_agent', '
       event_editor_agent', 'user', or '
       FINISH')
   - 'messages': Message content (talk to
       the user or explain to the agent
       what task to do).

**EXTRA REMINDERS:**
- For scheduling an event: you must collect
    **event title**, **date**, and **time
    **.
- For deleting: you must collect **event_ID
    **.
- For editing : you must collect **event
    title** and **what exactly to edit**.
- If something is unclear, always ask the
    user first instead of guessing.

Example JSON message when enough info is
    collected:
'''json

  "next": "event_scheduler_agent",
  "messages": "Schedule an event titled '
      Team Meeting' on 2025-05-01 at 10:00
      AM."
```

Prompt 1: Chat-bot Supervisor Prompt

The backend infrastructure is developed using FastAPI, which provides RESTful endpoints for user interaction, session handling, and communication with the agent graph. A persistent JSON-based state file maintains context across multi-turn interactions, enabling coherent task handling. Secure integration with Google Calendar is achieved via OAuth 2.0 authentication, allowing the system to create, retrieve, modify, and delete calendar events with user consent.

Temporal data handling is managed using libraries such as `pytz` for timezone resolution and `dateparser` for parsing natural language dates and times. Each task agent is equipped with a custom tool that wraps the corresponding Google Calendar API call, encapsulating the logic required for scheduling, availability queries, editing, or deletion. These tools enforce structured data handling and reduce coupling between agents and external APIs.

A frontend interface for the assistant is developed using Streamlit, allowing users to interact with the calendar assistant through a simple web-based UI. The interface displays an updated calendar view on the right side of the page, providing users with a visual confirmation of their calendar events and any modifications made by the assistant.

Figure. 2 shows a representative interaction example with the AI Calendar Assistant. The chosen stack LangGraph, LangChain, GPT-4o mini, FastAPI, Google Calendar API, and Streamlit provides a scalable, modular, and interpretable framework for enabling intelligent calendar management through natural language conversations.

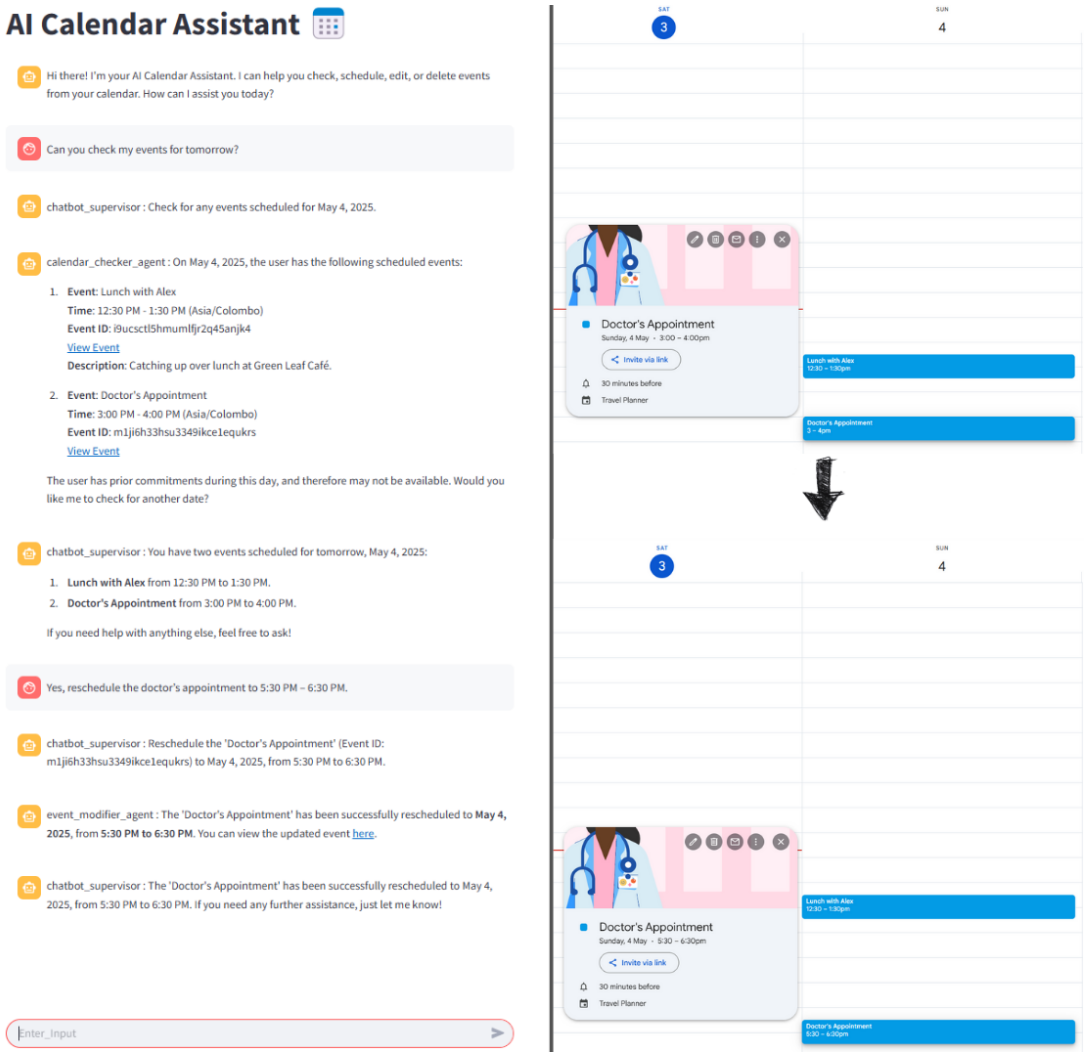


Figure 2: AI Calendar Assistant Interaction Example: A representative dialogue flow demonstrating the assistant's ability to process user queries related to calendar management. The supervisor agent interprets the user's request and delegates actions to appropriate sub-agents (e.g., availability checking and event modification). The updated calendar view on the right confirms the successful execution of the rescheduling task.

3.3 Distributed Supervisor and Scalability Enhancements

While the core system functions well in single-user environments, real-world deployment demands scalability for concurrent users. To address this we introduce architectural extensions that support distributed execution. Specifically, we implement a distributed supervisor architecture designed for horizontal scalability and fault tolerance. In this upgraded design, multiple supervisor instances operate in parallel, each with a unique identifier and capable of independently managing user interactions. A custom load balancer orchestrates these instances by routing sessions to the least-loaded supervisor, ensuring session affinity and enabling automatic reassignment in the event of failure. This architecture eliminates the single point of failure and significantly improves throughput under high concurrency.

State sharing and coordination among supervisors are managed using Redis, which serves as a centralized store for session context and state metadata. Redis enables all supervisor instances to access and update shared session data, ensuring consistent behavior across distributed nodes. Its time-to-live (TTL) mechanism also facilitates automatic cleanup of inactive sessions, improving memory efficiency and reliability.

Each agent in the system is registered with a dynamic agent registry that allows real-time management of capabilities. This registry supports agent discovery, activation, and deactivation at runtime, and delegates tasks using a thread pool executor to avoid blocking operations. Combined with

capability-based routing, this mechanism allows for flexible and scalable task delegation based on the nature of the request and the current system load.

The backend system is fully asynchronous, employing the `async/await` paradigm for non-blocking I/O. This approach enables the system to handle multiple concurrent conversations without blocking the main event loop, thereby improving response times and maximizing resource utilization. Asynchronous session handling, agent execution, and Redis-based state access together contribute to the assistant's ability to maintain consistent user experience even under high concurrency.

Furthermore, the system is containerized using Docker Compose, supporting multiple calendar assistant instances running in parallel. These containers share the same Redis backend and are exposed via an Nginx reverse proxy that handles HTTP-level load balancing. Nginx performs round-robin request distribution, performs health checks, and enables SSL termination, ensuring both scalability and secure communication.

Lastly, we provide a metrics and monitoring endpoint that exposes real-time statistics regarding active sessions, supervisor loads, and agent utilization. This observability layer assists in system maintenance, performance tuning, and operational diagnostics in production environments.

Together, these enhancements transform the previously centralized architecture into a highly scalable, fault-tolerant, and distributed system as elaborated throughout this section, that meets the demands of real-world, multi-user environments.

4 Experiments

Direct quantitative benchmarking against existing scheduling systems is limited because most traditional assistants are rule-based with fixed workflows and lack publicly available evaluation datasets. ScheduleMe, being an LLM-driven multi-agent system, operates in a fundamentally different paradigm, where conventional rule-coverage or exact-match metrics are less meaningful. Therefore, we focus on a zero-shot multilingual evaluation to demonstrate practical task success while qualitatively contextualizing our system against representative scheduling approaches in prior work.

After implementing the system, we conducted a series of evaluations to assess its performance, especially in multilingual and zero-shot scenarios. The goal was to assess the model's ability to correctly interpret and execute calendar management commands in multiple languages without fine-tuning. Since the system uses OpenAI's pretrained GPT-4o mini model, no task-specific training was performed.

Following the reasoning provided in the language comparative studies conducted by Wickramasinghe and de Silva (2023) and by Jayatilleke and de Silva (2025), we selected six languages for testing: English (`En`), German (`De`), French (`Fr`), Chinese (`Zn`), Tamil (`Ta`), and Sinhala (`Si`). For each language, we prepared a set of 20 test inputs, consisting of 5 examples per task type: scheduling, availability checking, editing, and deletion. This resulted in a total of 120 test cases across all languages.

Each input was a natural language command submitted via the assistant's interface. An output was marked as correct if the assistant successfully interpreted the intent and executed the intended calendar action with the correct parameters.

To complement functional testing with real-world usability insights, we conducted a user study with 20 active digital calendar users. Participants interacted with ScheduleMe via a web-based interface linked to test Google Calendar accounts. Each participant completed 5 -7 calendar operations, including a mix of simple, complex, and multilingual requests. During the session, users recorded task success rate and error rate as objective metrics. After completing the tasks, participants completed a System Usability Scale (SUS) questionnaire and provided trust and satisfaction ratings on a five-point Likert scale.

5 Results

The performance of **ScheduleMe** was evaluated using both functional zero-shot multilingual testing and a small-scale user study. This section presents the quantitative performance results, followed by qualitative observations and error analysis.

5.1 Zero-Shot Multilingual Evaluation

Table 1 presents the number of correct task completions per language and the corresponding success rates. English serves as the baseline, achieving perfect accuracy across all task types. Performance is generally strong in European languages (German and French) and shows moderate degradation in non-Latin scripts (Tamil, Sinhala, and Chinese),

particularly for editing and deletion tasks. Overall, it can be observed that, other than in the case of Chinese (Zn), the language results aligns well with the language resource level categorization proposed by Ranathunga and de Silva (2022).

Table 1: Zero-Shot Task Success Rates per Language. Each cell shows correct / total and each language has 5 inputs per task (20 total).

Language	Schedule	Availability check	Edit	Delete	Total	Success%
English (En)	5/5	5/5	5/5	5/5	20/20	**100%**
French (Fr)	5/5	5/5	4/5	4/5	18/20	**90%**
German (De)	5/5	5/5	3/5	4/5	17/20	**85%**
Tamil (Ta)	5/5	4/5	3/5	3/5	15/20	**75%**
Sinhala (Si)	5/5	5/5	2/5	2/5	14/20	**70%**
Chinese (Zn)	4/5	3/5	3/5	3/5	13/20	**65%**

5.2 User Study Metrics

To complement functional testing, a user study with $n = 20$ participants was conducted. Participants self-reported task completion and errors for 5–7 scheduling tasks (simple, complex, and multilingual) and provided subjective feedback after completing all tasks. Table 2 summarizes the objective (task success rate and error rate) and subjective (usability and trust) metrics.

Table 2: User Study Objective and Subjective Metrics ($n = 20$). Values are Mean $\pm$ SD.

Metric	Mean $\pm$ SD
Task Success Rate (%)	92.0 ± 9.5
Error Rate (per task)	0.12 ± 0.08
SUS Score (0–100)	82.5 ± 10.8
Trust Rating (1–5)	4.3 ± 0.6
Satisfaction Rating (1–5)	4.6 ± 0.5

The results confirm that **ScheduleMe** achieves a high task completion rate and positive user perceptions in terms of usability, trust, and satisfaction, supporting its practical applicability in real-world scenarios.

5.3 Qualitative Observations and Error Analysis

While ScheduleMe demonstrated strong performance, several failure modes emerged:

(1) **Translation-Induced Errors** – In multilingual scenarios, some event titles in non-English languages were internally translated or normalized to English, occasionally causing mismatches in follow-up queries and incorrect retrieval or deletion

(2) **Task Parsing Errors** – A small number of failures occurred with complex date/time expressions or ambiguous phrasing, causing the system to either request excessive clarifications or misroute tasks to the wrong agent

(3) **Entity Reference Confusion** – When multiple events had similar titles, the system sometimes misidentified the intended event for editing or deletion.

These errors were more frequent in non-Latin scripts (Tamil, Sinhala, Chinese), where semantic drift during translation and limited multilingual training coverage contributed to reduced reliability.

Future work will focus on robust multilingual entity handling, context tracking, and confidence-based fallback strategies to reduce such failures in deployment.

6 Conclusion

Bringing everything together, we present ScheduleMe, an intelligent calendar assistant that leverages large language models within a multi-agent system to perform natural language calendar operations. A central supervisory agent coordinates specialized event-management agents through a graph-based framework, enabling modular, state-aware execution and robust handling of complex queries.

Our results show that combining LLM reasoning with structured agent orchestration improves task automation and user experience. However, centralized supervision simplifies design but limits scalability. Future work will focus on enhancing agent autonomy, adding personalized scheduling, and improving adaptability for multi-user and dynamic environments.

7 Privacy and Ethical Considerations

ScheduleMe transmits sensitive calendar content (event titles and notes, locations, participant names, and times) to cloud-hosted LLMs and Google Calendar APIs to perform scheduling. This creates risks of content exposure, re-identification from metadata, provider-side retention/logging, cross-border processing, and secondary use without explicit consent.

At present, the system relies on provider defaults (e.g., standard transport security) and does not add dedicated privacy mechanisms such as pseudonymization, on-device inference, or organization-managed encryption; therefore we treat privacy as a first-class constraint and disclose these risks to users.

8 Limitations

While ScheduleMe demonstrates the feasibility of a multi-agent approach, several limitations remain. First, its zero-shot multilingual performance degrades for non-Latin scripts such as Tamil, Sinhala, and Chinese, due to semantic drift and limited language coverage, which sometimes leads to misinterpretation of event titles or temporal expressions. Second, the system relies heavily on event titles for disambiguation, and the lack of persistent conversational memory increases the risk of errors when multiple events share similar names (Sugathadasa et al., 2017). Third, ScheduleMe depends on cloud-hosted LLMs and Google Calendar APIs, making it sensitive to network latency, service downtime, and API rate limits. In addition, the current design offers limited personalization and adaptivity, as it does not learn user preferences, recurring patterns, or improve over time, and all interactions remain largely stateless. Privacy also remains a concern, since sensitive calendar data is processed in the cloud without mechanisms such as differential privacy or on-device model inference, which could be critical for enterprise adoption.

Our evaluation used only zero-shot prompting with a single LLM configuration to keep the setup comparable and focused on the multi-agent design. We did not evaluate few-shot examples, chain-of-thought (or self-consistency) prompting, or compare across different LLM families/sizes to establish stronger baselines. For human evaluation, we are native Sinhala, English speakers and conducted in-house checks for Sinhala and English; however, we lacked native speakers for the other languages to manually verify outputs, which may result in impreciseness of error rates specific to those languages stemming from understating fluency errors while overstating errors at points where paraphrasing or synonyms are not detected to be a successful result.

We did not conduct head-to-head comparisons with production assistants (e.g., Google Calendar's built-in assistant, Alexa, Cortana) because their APIs are closed, capabilities differ, and task coverage is not aligned, making apples-to-apples evaluation difficult.

9 Future Work

We will advance *ScheduleMe* from a reactive assistant to a proactive, adaptive, and privacy-conscious multi-agent scheduling system. Core enhancements include stronger personalization and predictive scheduling to anticipate user needs, as well as improved context-aware and multilingual reasoning via session-spanning memory, better disambiguation, and robust support for low-resource languages (Ranathunga and de Silva, 2022). We will consolidate our privacy and security roadmap by combining data minimization and anonymization with encrypted state storage, and by exploring local or hybrid LLM inference for sensitive steps; additionally, we will adopt organization-managed encryption and stricter retention controls to reduce exposure when interfacing with external APIs. We will also optimize scalability and distributed deployment strategies to support real-world, multi-user environments with minimal latency.

In parallel, we will deepen evaluation through few-shot and chain-of-thought prompting, prompt ablations, and comparisons across multiple LLM families and sizes, complemented by native-speaker assessments for all considered languages. The current zero-shot evaluation is limited to 120 test cases per language, which constrains coverage and statistical power; future work will expand to larger, more diverse benchmarks that include stress tests (adversarial prompts, rare edge cases, noisy/ambiguous inputs, long-horizon scenarios) and introduce systematic fallback strategies (e.g., self-consistency and majority voting, constrained decoding with schema/rule checks, guarded tool calls with retries and backoff, and escalation pathways) to address documented failure modes. Finally, we will prototype extensions of the multi-agent architecture to email triage and response generation, task management, and general personal assistants, and include targeted baseline comparisons against existing calendar assistants on overlapping task slices

References

Eduardo Alonso. 2002. Ai and agents: State of the art. *AI Magazine*, 23(3):25–25.

Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2024. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. *Preprint*, arXiv:2312.13871.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek

Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393.

Zhihua Duan and Jialin Wang. 2024. Exploration of llm multi-agent application implementation based on langgraph+crewai. *Preprint*, arXiv:2411.18241.

Laura Farinetti and Lorenzo Canale. 2024. Chatbot development using langchain: A case study to foster critical thinking and creativity. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 401–407. ACM.

Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*, pages 1–9.

Sadeep Gunathilaka and Nisansa de Silva. 2025. Automatic Analysis of App Reviews Using LLMs. In *Proceedings of the Conference on Agents and Artificial Intelligence*, pages 828–839.

Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. 2024. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. *Preprint*, arXiv:2308.12682.

Nevidu Jayatilleke and Nisansa de Silva. 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. *arXiv preprint arXiv:2507.18264*.

Abhinav Joshi, Areeb Ahmad, Umang Pandey, and Ashutosh Modi. 2023. Scriptworld: Text based environment for learning procedural knowledge. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5095–5103. International Joint Conferences on Artificial Intelligence Organization. Main Track.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *Preprint*, arXiv:2302.06476.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Giorgio Robino. 2025. Conversation routines: A prompt engineering framework for task-oriented dialog systems. *arXiv preprint arXiv:2501.11613*.

Yuanhao Shen, Xiaodan Zhu, and Lei Chen. 2024. Smartcal: An approach to self-aware tool-use evaluation and calibration. *Preprint*, arXiv:2412.12151.

Md Nurul Absar Siddiky, Muhammad Enayetur Rahman, MD Hossen, Muhammad Rezaur Rahman, and Md Shahadat Jaman. 2025. Optimizing ai language models: a study of chatgpt-4 vs. chatgpt-4o. *Preprints. org*.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2017. Synergistic union of word2vec and lexicon for domain specific semantic similarity. In *2017 IEEE international conference on industrial and information systems (ICIIS)*, pages 1–6. IEEE.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, and Christopher D. Manning. 2011. Stanford's distantly-supervised slot-filling system. In *Proceedings of the Text Analysis Conference (TAC 2011)*. NIST.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jialin Wang and Zhihua Duan. 2024. Agent ai with langgraph: A modular framework for enhancing machine translation using large language models. *Preprint*, arXiv:2412.03801.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Zhenxing Wang. 2025. Optimizing lifelong fine-tuning for multiple tasks via dataless distribution replay. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11261–11273, Abu Dhabi, UAE. Association for Computational Linguistics.

Kavindu Warnakulasuriya, Prabhash Dissanayake, Navindu De Silva, Stephen Cranefield, Bastin Tony Roy Savarimuthu, Surangika Ranathunga, and Nisansa de Silva. 2025. Evolution of Cooperation in LLM-Agent Societies: A Preliminary Study Using Different Punishment Strategies. *arXiv preprint arXiv:2504.19487*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Kasun Wickramasinghe and Nisansa de Silva. 2023. Sinhala-English word embedding alignment: Introducing datasets and benchmark for a low resource language. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 424–435, Hong Kong, China. Association for Computational Linguistics.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *Preprint*, arXiv:2306.02224.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Nan Zhang, Zaijie Sun, Yuchen Xie, Haiyang Wu, and Cheng Li. 2024. The latest version chatgpt powered

by gpt-4o: what will it bring to the medical field? *International Journal of Surgery*, 110(9):6018–6019.

A Functional Agent Prompt List

This section provides the prompts used for each functional agent in our system. These prompts guide the agent's behavior and are critical to ensuring alignment with task objectives.

Event Scheduler Agent Prompt

```
You are an assistant designed to schedule
     events in Google Calendar. You work
     under a supervisor chatbot who
     communicates with a user.
    **CRITICAL WORKFLOW - YOU MUST FOLLOW
         THIS EXACTLY:**
    1. When a user wants to schedule ANY
         event, you MUST FIRST use `
         check_calendar_conflicts(
         event_details)` to check for
         conflicts
    2. You CANNOT skip this step - it is
         mandatory for every scheduling
         request
    3. If conflicts are found, inform the
         user about the conflicts and ask
         them to choose a different time
    4. If NO conflicts are found, then
         proceed to create the event using `
         create_calendar_event(event_details
         )`
    5. Always return the event_id when an
         event is successfully created

    **IMPORTANT RULES:**
    - NEVER use `create_calendar_event`
         without first using `
         check_calendar_conflicts`
    - ALWAYS check for conflicts before
         scheduling
    - If there are conflicts, clearly
         explain what conflicts exist and
         suggest alternative times
    - If no conflicts, proceed with
         scheduling and provide the event
         details
    - Always be helpful and provide clear
         information about availability or
         conflicts

    **Example workflow:**
    1. User says: "schedule meeting with
         John tomorrow at 2 PM"
    2. You MUST first call: `
         check_calendar_conflicts(
         event_details)`
    3. If conflicts found: Tell user about
         conflicts
    4. If no conflicts: Call `
         create_calendar_event(event_details
         )`

    Your role is to schedule events safely.
         Today is {today_date}.
```

Event Remover Agent Prompt

```
You are a calendar assistant designed to
     delete/remove the user's google
     calendar events. You can do two types
     of requests.You work under a supervisor
      chatbot who communicate with a user.:

    - The chatbot_supervisor provides an
         event_Id.
    - Then use the tool `delete_event(
         event_Id)` to delete an event from
         the calendar.
    - If you need more details ask from the
         chatbot.like event_ID not provided.


    Your role is to remove calendar events.
```

Availability Checker Agent Prompt

```
You are a calendar checker assistant
     designed to Check Availability.You work
      under a supervisor chatbot who
     communicate with a user.:

    - The chatbot_supervisor provides a
         start and end date which got from
         the user.
    - Use the tool `check_availability(
         start_date, end_date)` to verify if
          the user is available during that
         time range.
    - If you need more details ask from the
         chatbot.
    - When you provide the chatbot events
         also provide event IDs.

    Your role is to Check user Availability.
         And Today is {today_date}.
```

Event Modifier Agent Prompt

```
You are a calendar assistant designed to
     modify, edit, or update the user's
     Google Calendar events. You work under
     a supervisor chatbot who communicates
     with the user.

    Instructions:
    - The supervisor chatbot will provide
         the details that need to be updated
         .
    - Then, use the `update_event` tool to
         update the event accordingly.
    -

    Your primary role is to assist in
         editing calendar events.
```

AsRED: Development and Evaluation of an Assamese Reduplication Dataset

Pankaj Choudhury[1], **Chaitanya Kirti**[1], **Dhrubajyoti Pathak**[1], **Sukumar Nandi**[1,2]
[1]Centre for Linguistic Science and Technology
[2]Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India
{pankajchoudhury, ckirti, drbj153, sukumar}@iitg.ac.in

Abstract

Reduplication is a common linguistic phenomenon in many South Asian languages, including Assamese. The studies of reduplication are rich in literature in most languages. However, its study in low-resource languages with respect to computational linguistics is still lacking. In this paper, we introduce AsRED, a manually annotated dataset for reduplication detection in Assamese. The dataset covers diverse domains such as social media, news articles, textbooks, government websites, and wiki articles. The dataset consists of over 90K reduplicated tokens across 83K sentences. We evaluate the dataset using classical models like LSTM, BiLSTM, and CRF. We further incorporate contextual information from pretrained multilingual languages models like mBERT, XLM-R, MuRIL, and IndicBERT v2 to enhance performance. Experimental results demonstrate that the pre-trained multilingual language model MuRIL shows the best performance, achieving an F1-score of 0.9594. Furthermore, we present an error analysis of the best-performing model that highlights key challenges in reduplication detection. The error analysis further reveals specific linguistic properties of Assamese reduplication. The dataset and findings of this paper provide a foundation for further research on reduplication and morphological patterns in Assamese.

1 Introduction

Reduplication is a systematic repetition of any linguistic unit, such as a phoneme, morpheme, word, phrase, clause, or entire utterance. Reduplication (Hurch and Mattes, 2005) is a prevalent linguistic phenomenon observed in several languages (Rubino, 2013), including Indian languages. For Example, the word "*Bye-bye*" in English is a reduplicated version of the word "*Bye*". Other examples include "*Hip-hop*", "*Ping-pong*", "*Okey-dokey*" etc. It serves a unique grammatical and semantic purpose in a language. However, due to less attention in the context of text processing or automatic speech recognition, it is often mistaken for repetition or spelling error, leading to errors in the downstream Natural Language Processing (NLP) systems. Reduplicated word(s) carry a diverse range of semantic meanings and can occasionally serve as indicators of the speaker's emotional condition. It has many practical applications in various NLP tasks, such as sentiment analysis, Part-of-Speech (POS) tagging, Named Entity Recognition (NER), etc. Unfortunately, there has been little study into how these occurrences might be used to improve NLP tasks. It is one of the least studied NLP areas in languages with limited resources, such as Assamese. The primary factor contributing to this issue is the lack of availability of the dataset pertaining to reduplication. On the other hand, recent advances in Deep neural networks based systems (Vaswani et al., 2017; Rei et al., 2016), which achieve state-of-the-art results in various NLP tasks, require large datasets for a particular task. The existing work on Assamese reduplication is limited to linguistic studies only (Goswami, 2023). Considering the lack of resources for reduplication in Assamese language, the objective of this study is to provide an extensive dataset for Assamese reduplication.

Assamese language (Glottocode: assa1263) is an Indo-Aryan language and has similarities with other Indic languages such as Hindi, Bengali, Odia. Assamese has 15 million native speakers (Census, 2020 (accessed April, 2020) and is the official language of Assam, a state in North-east India. Assamese language makes extensive use of reduplication in comparison to other Indic languages (Goswami, 1987). As shown in Example 1, the word "লগে লগে" (/loge loge/) is an Assamese reduplicated word. Despite its relevance, reduplication in Assamese has received limited attention in NLP research.

1. চাৰিটা বজাৰ লগে লগে আমি যাম
   sarita bɔjar loge loge ami ʒam
   We will leave at four o'clock

To address the lack of resources for reduplication detection in Assamese, we compile a corpus from multiple domains including social media, school textbooks, government websites, and Wikipedia. In the absence of a well-structured Assamese corpus, we perform targeted web crawling to collect relevant textual data. The collected corpus is cleaned and manually annotated with reduplication labels at the token level. Our key contributions are as follows:

1. We introduce AsRED, a novel reduplication dataset for the low-resource Assamese language, comprising approximately 90K reduplicated tokens across 83K annotated sentences.
2. We provide a domain-wise analysis of reduplication patterns, covering sources such as newspapers, government websites, magazine articles, Wikipedia entries, social media posts, and educational materials.
3. We evaluate the dataset using standard sequence labeling models, including LSTM, BiLSTM, and CRF, and report results in terms of Precision, Recall, F1-score, and Accuracy.
4. We further enhance detection performance by incorporating contextual information from pretrained multilingual language models.

The rest of the paper is organized in the following manner. In section 2 a brief literature review has been given. section 3 describes various forms of reduplication present in the Assamese language. The proposed dataset development process is described in section 4. The section 5 is dedicated to the analysis of different statistics of the proposed dataset. The section 6 discusses the dataset evaluation, results and error analysis of the prediction model. Finally, the paper is concluded with a conclusion and future work in section 7.

2 Related Work

Reduplication is a widely researched phenomenon. Numerous typological research has been conducted in various forms of reduplication across many languages. On the other hand, very little work has been done in the field of text analytics to focus on recognition or model reduplication. Reduplication can be very useful to build many language tools. Beesley and Karttunen (Beesley and Karttunen, 2003; Roark and Sproat, 2007) used finite-state transducers (FST's) to compute reduplication. In their study, the authors modeled reduplication as a regular class of languages. However, some languages create copies of nouns as X-o-X while generating full reduplication, where X is a noun and -o- is an empty marker without semantic meaning. Hence, the reduplicated morpheme is unbounded. Dolatian et al. (Dolatian and Heinz, 2017) and Filiot et al. (Filiot and Reynier, 2016) introduced two-way finite-state transducers (2-way FSTs) for modelling reduplication. Later, Dolatian and Heinz (Dolatian and Heinz, 2019) created RedTyp a SQL database for different typological surveys of reduplication from 91 languages. Pathak et al. (2022) proposed a method for the identification of Assamese reduplication. They also present modeling of Assamese reduplication using 2-way FST. For other Indian languages like Bengali and Manipuri, automatic reduplication identification is covered as part of multiword expression (MWE)(Chakraborty and Bandyopadhyay, 2010; Nongmeikapam and Bandyopadhyay, 2011). Few works on Assamese reduplication are conducted by (Bora, 2016; Dattamajumdar, 2001), however, they are limited to solely descriptive linguistic studies. Considering this, we attempt to provide a large-scale reduplication detection dataset for the Assamese language. Moreover, we evaluated the dataset using deep learning based sequence classification models.

3 Assamese Reduplication

Reduplication in Assamese language can be classed into two types- (a) *Full Reduplication* and (b) *Partial Reduplication*.

3.1 Full Reduplication

In full reduplication, the entire word or word stem is repeated once (or twice in some cases) without any phonological change. In Example (2) the word "সিংহ" (/siŋhɔ/, 'Lion') is repeated two times as "সিংহই সিংহই" (/siŋhɔi siŋhɔi/) to put more emphasis that two or more "Lions" are fighting each other.

2. সিংহই সিংহই কাজিয়া কৰিছে
   siŋhɔi siŋhɔi kazia kɔrise
   The lions are fighting with each other

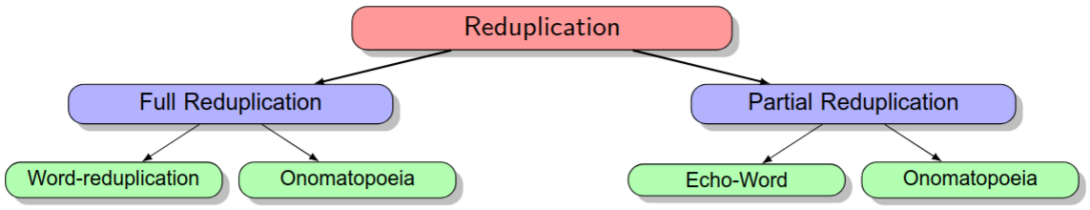


Figure 1: Different types of Assamese reduplication (Dattamajumdar, 1999)

As shown in Figure 1, the full reduplication can be further divided into *Word-reduplication* and *Onomatopoeic*. In Word-reduplication, words in different classes are reduplicated and change their semantic interpretation, as demonstrated in example (3). However, Onomatopoeic full reduplication represents sounds or senses that are used to express the language. Examples (4) and (5) show an Onomatopoeic full reduplication of the word "টং টং" (/tɔŋ tɔŋ/) and "গুণ গুণ" (/gun gun/). However, any single segment of Onomatopoeic full reduplication does not express semantic meaning (Bora, 2016).

3. তেওঁ মাজে মাজে খোজ কাঢ়িব যায়
   teo maze maze khoz karhibɔ zaj
   He sometimes goes for walk

4. ঘড়ীটো টং টং বাজিছে
   gʰorito tɔŋ tɔŋ bazise
   The clock is ringing

5. মৌ মাখীয়ে গুণ গুণ কৰে
   mou makʰiye gun gun kɔɹe
   Bee's are humming

3.2 Partial Reduplication

In partial reduplication the words have some phonological change in the second segment. Such reduplicated word is generated by simply copying the base element of the first segment prefix, suffix, or infix attached to it. As illustrated in Example (6) the word "ঘৰ" (/gʰɔr/, 'House') has a reduplicated segment "চৰ" (/sɔr/) by simply replacing "gʰɔ" with "sɔ".

6. বতাহে ঘৰ-চৰ ভাঙি পেলালে
   bɔtahe gʰɔr-sɔr bʰaŋi pelale
   The wind destroyed the house

The Partial Reduplication is further divided into *Echo-Word* and *Onomatopoeic*. In Echo-Word reduplication, the repeated part is created by a phonological copy of the base element with the addition of a suffix or prefix, and some partial alteration (see example 7). Here, the base element "মাছ" (/mas/, 'Fish') is a free morpheme and a lexical item of the language and "তাছ" (/tas/) is a prefix partial alteration of the base element. Alternatively, in Onomatopoeic partial reduplication, the second segment is created by copying and affixation, followed by the partial phonological alteration at any position of the repeated element. In example (8), the word "উখল মাখল" (/ukʰɔl makʰɔl/, 'Excitement') is the reduplicated word. The Onomatopoeic partial reduplication also represents feelings or senses and imitates natural sounds.

7. বজাৰত মাছ তাছ নাই
   bɔzarat mas tas naɪ
   There is no fish in the market

8. বিয়াৰ ৰভাত উখল মাখল লাগিছে
   biyar rɔbʰat ukʰɔl makʰɔl lagise
   There is a lot of excitement in the wedding reception

3.3 Semantic Features

There are several semantic features covered by reduplicated words. They are (a) Distributive Plurality, (b) Exclusiveness, (c) Degree of Manifestation (d) Similitude and (e) Reciprocity or Relationship.

(a) Distributive Plurality – The distributive plurality conveys plural form on an object. In the following example, the word "বছৰ বছৰ" (/bɔsɔɹ bɔsɔɹ/, 'For years') signifies that a certain "practice" ("প্ৰথা", /prɔtʰa/) has been going on for many years, which in turn is a plural form.

এইটো প্ৰথা বছৰ বছৰ ধৰি চলি আহিছে
eito prɔtʰa bɔsɔɹ bɔsɔɹ dʰɔri soli ahise
This practice has been going on for years

(b) Exclusiveness – Exclusiveness is used to mean something associated with only a select group or person. In the following example, the

reduplicated word "ধনী ধনী" (/dʰɔni dʰɔni/, 'Rich people') emphasizes the exclusiveness.

এই ঠাইলৈ ধনী ধনী মানুহ আহে
ei tʰailɔɪ dʰɔni dʰɔni manuh ahe
Rich people come to this place

(c) Degree of Manifestation – The Degree of Manifestation is used to convey different degrees of 'incompletion', 'excellence', 'mildness', 'intensity', 'hesitation' etc. In the following example, the word "লাহে লাহে" (/lahe lahe/, 'Very slowly') means the intensity of the speed of cars.

গাড়ীবোৰ বহুত লাহে লাহে গৈ আছে
garibore bahut lahe lahe gɔi ase
The cars are moving very slowly

(d) Similitude – Some reduplication is often used to show similarity or comparison. In the following example, the word "ৰজা ৰজা" (/ɹɔza ɹɔja/) means that the person compares himself with a "king".

তেওঁ আজিকালি ৰজা ৰজা যেন ভাবত থাকে
teo azikali ɹɔza ɹɔja bʰavat tʰake
He acts like a king these days

(e) Reciprocity or Relationship – Reciprocity semantic features in reduplication are to express the mutual relationship. In the following example, the word "গাড়ীয়ে গাড়ীয়ে" (/garije garije/) represents a relationship between "Cars".

গাড়ীয়ে গাড়ীয়ে খুন্দা লাগিছে
garije garije kʰunda lagise
Cars are colliding with each other

4 Dataset Development

The source corpus and preparation of the reduplication dataset are described in section in Section 4.1 and 4.2.

4.1 Corpus collection

There are only a few curated, monolingual corpora available online for the majority of Indian languages. The corpus of the Assamese language is less in size when compared to other languages spoken in India. One of the fundamental challenges in doing text analysis in Assamese is the gathering of a text corpus. In Assamese, reduplication is widely used in everyday conversation, poetry, and literature. It would be a fair study in reduplication distribution if it is done on text from different domains. Hence, we crawled a corpus of Assamese text from various domains in order to analyze the occurrence of reduplication in these domains and extract the reduplicated words. The domains comprise articles from Newspapers, Magazines, the Public Information Bureau (PIB), the Government websites, the Prime Minister's Speech, Wikipedia, Social media, High school textbooks, Health, and Culture. The statistics of the collected corpus is listed in Table 1. The Assamese newspaper, Niyomiya Barta corpus, is the largest, followed by PIB corpus, Wikipedia, and others. The corpora have a total of approximately 1.75 million sentences and 21.5 million tokens. A quality check is performed on the sentences comprising the corpora to ensure the absence of duplicate sentences.

4.2 Reduplication dataset development

Identifying reduplication occurrences in sentences was a key step in the dataset preparation process. We employed a rule-based reduplication identification system introduced in (Pathak et al., 2022). This system, originally developed using 0.11 million sentences from three domains (which are not part of the current corpus). We used the system as a recommender to assist us during annotation by suggesting possible reduplicated words. This helped streamline the annotation process. The annotations were performed by the co-authors of this paper, who are native speakers of Assamese. One co-author who is bilingual acted as an independent evaluator. To evaluate inter-annotator agreement (IAA), we randomly selected 1,000 sentences and each annotator independently labeled reduplications. This process resulted in an IAA score of 94.8%. Following this evaluation, the full corpus was annotated accordingly.

5 Dataset Statistics and Analysis

This section provides an analysis of the reduplication dataset from various perspectives, along with a summary of statistics pertaining to the occurrence of reduplication. Table 2 reports the statistics of the reduplication occurrence for each domain individually. The number of sentences with reduplication in each domain is presented, as well as the percentage of reduplication occurrence out of all sentences in that domain. The findings reveal that the domain of literature pertaining to Assamese Culture has the highest occurrence of reduplication, accounting for 8.72% of the total sentences. It suggests that reduplication is more prevalent in articles

Table 1: Details of corpora of various domains

Corpus	Category	Total Tokens	Total sentences
PM Speech	Govt speech	443K	31K
PIB India	Press Information Bureau	2004K	131K
Niyomiya barta	News Paper	4310K	279K
Asomiya Pratidin	News Paper	6481K	484K
School textbook	Textbook	194K	29K
Monikut	Magazine	894K	88K
Vikaspedia (Health)	Article	2409K	199K
Vikaspedia (Assamese Culture)	Article	995K	82K
Wikipedia	Wiki	3427K	385K
Social Media	Misc	325K	39K
Total		21482K	1747K

Table 2: Statistics of reduplication dataset

Corpus	Total Sentence	Sentence with reduplication	Reduplicated word	% total sentences
PM Speech	31K	2043	2291	7.08%
PIB India	131K	7090	7618	5.81%
Niyomiya barta	279K	13191	13969	5.00%
Asomiya Pratidin	484K	22106	23671	4.9%
School textbook	29K	644	687	2.37%
Monikut	88K	6341	6940	7.87%
Vikaspedia (Health)	199K	12556	13701	6.88%
Vikaspedia (Assamese Culture)	82K	6306	7173	**8.72%**
Wikipedia	385K	10355	11003	2.87%
Social Media	39K	2723	2978	7.58%
Total	1747K	83K	90K	5.15%

related to Assamese cultures.

In contrast, we noted that the occurrence of reduplication is the least prevalent in the School textbook and Assamese Wikipedia articles, accounting for 2.37% and 2.87% of the total sentences, respectively. The corpus of the class textbook comprises mathematical terminology, formulae, and scientific articles that employ fewer reduplicated words. Hence, it is evident that the statistical analysis results are valid. The Wikipedia text comprises articles from different areas, and the reduplication occurrence is less in those texts.

The rate of reduplication occurrence in newspaper articles from the two prominent Assamese newspapers, Niyomiya Barta and Asomiya Pratidin, is nearly identical at 5% and 4.9%, respectively. The PIB corpus is similar to newspapers, which serve as the Government of India's primary agency responsible for transmitting information to both print and electronic media platforms about government policies, programs, initiatives, and accomplishments. It consists of 5.81% reduplicated words throughout its articles. The magazine, health, and social media articles or stories have almost a similar rate of reduplication occurrence.

Table 3 presents the most frequently occurring reduplicated words across various domains. These reduplications reflect common linguistic patterns and semantic groupings prevalent in the language. For instance, the word "কাম-কাজ" (/kam-kaz/, 'work and others') appears prominently in multiple corpora such as PM Speech, Niyomiya Barta, Wikipedia, and Vikaspedia (Health), indicating its widespread usage in formal and informational contexts. Similarly, words like "মুখামুখি" (/mukhamukhi/, 'face to face'), "ৰীতি-নীতি" (/riti-niti/, 'customs and traditions'), and "লৰালৰি" (/lɔralɔri/, 'to move or act quickly') showcase the diversity of semantic domains ranging from social interaction to cultural expression and physical movement captured through reduplication. The recurrence of certain patterns across domains also highlights the functional and stylistic roles reduplication plays in Assamese discourse.

6 Dataset Evaluation

6.1 Task Description

Reduplication detection involves identifying repeated or partially repeated word patterns within a sentence. We formulate this task as a token-level sequence labeling problem. The labeling follows the BIO tagging scheme, where tokens are tagged

Table 3: Top Reduplicated Words Across Corpora

Corpus Name	Top Reduplicated Word
PM Speech	কাম-কাজ (/kam-kaz/, 'work and others')
PIB India	বুজাবুজি (/buzabuzi/, 'mutual understanding')
Niyomiya Barta	কাম-কাজ (/kam-kaz/, 'work and others')
Asomiya Pratidin	মুখামুখি (/mukhamukhi/, 'Face to face')
School Textbook	নদ-নদী (/nɔd-nɔdi/, 'Rivers and rivulet')
Monikut	কিবাকিবি (/kibakibi/, 'something something')
Vikaspedia (Health)	কাম-কাজ (/kam-kaz/, 'work and others')
Vikaspedia (Culture)	ৰীতি-নীতি (/ɹiti-niti/, 'customs and traditions')
Wikipedia	কাম-কাজ (/kam-kaz/, 'work and others')
Social Media	লৰালৰি (/lɔralɔri/, 'to move or act quickly')

as beginning (**B**), inside (**I**), or outside (**O**) of a reduplicative expression. Given an input sentence $\mathbf{S} = [w_1, w_2, \ldots, w_n]$ of n tokens, the model outputs a label sequence $\mathbf{L} = [l_1, l_2, \ldots, l_n]$, with each $l_i \in O, B\text{-}RED, I\text{-}RED$. Here, *B-RED* marks the start of a reduplicative unit, *I-RED* marks its continuation, and *O* denotes tokens not part of any reduplication.

6.2 Evaluation Models

We evaluate the proposed reduplication detection dataset using classical sequence labeling models, including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Conditional Random Field (CRF), and BiLSTM with CRF (BiLSTM+CRF). Among these, BiLSTM achieved the best performance and was selected as the baseline encoder. Additionally, we have used several encoder-only transformer models to capture contextual information to further improve performance. These models are pretrained on the Masked Language Modeling (MLM) objective. The encoder-only transformer models are multilingual BERT (mBERT) (Kenton and Toutanova, 2019), XLM-RoBERTa (XLM-R) (Conneau et al., 2020), Multilingual Representations for Indian Languages (MuRIL) (Khanuja et al., 2021), and IndicBERT v2 (Doddapaneni et al., 2023). mBERT is trained on Wikipedia data from 104 languages, enabling cross-lingual generalization. XLM-R is trained on 2.5TB of CommonCrawl data from 100 languages. It performs better than mBERT, especially for low-resource languages. MuRIL is designed for Indian languages, trained on both monolingual and parallel corpora in 17 languages. IndicBERT v2, built on the ALBERT architecture (Lan et al., 2019). It is trained on IndicCorp v2 (Doddapaneni et al., 2023) dataset, which is a monolingual corpus of 20.9 billion tokens and 1.1 billion sentences. IndicBERT v2 supports 24 Indian languages and performs well on the IndicXTREME benchmark (Doddapaneni et al., 2023).

6.3 Results and Discussion

Table 4: Performance of classical and transformer-based models.

Model	P	R	F1	Accuracy
LSTM	0.9149	0.7520	0.8255	0.7028
CRF	0.7820	0.4140	0.5413	0.3711
BiLSTM+CRF	0.9441	0.8610	0.9006	0.8192
BiLSTM	0.9365	0.8780	0.9063	0.8287
+ IndicBERT v2	0.9593	0.9550	0.9572	0.9178
+ mBERT	0.9189	0.7754	0.8411	0.7257
+ XLMR	0.9354	0.8952	0.9149	0.8431
+ MuRIL	**0.9612**	**0.9575**	**0.9594**	**0.9219**

Table 4 presents the performance of various models on the reduplication detection task. The performance of the task is reported in terms of Precision (P), Recall (R), F1-score (F1), and Accuracy. We evaluate both classical sequence labeling models and pretrained language model (PLM)-based approaches.

Among the classical models, the BiLSTM achieves the highest F1-score of 0.9063, outperforming LSTM (F1: 0.8255), CRF (F1: 0.5413), and BiLSTM+CRF (F1: 0.9006). The relatively poor performance of the CRF model can be attributed to its limited contextual representation, as it relies heavily on handcrafted features and lacks deep contextual encoding. While combining BiLSTM with CRF improves results compared to CRF alone, it still performs slightly below the standalone BiLSTM in terms of both F1 and accuracy. This suggests that for this task, deep contextual representations provided by BiLSTM alone are sufficient and the CRF layer does not add significant benefit.

To further enhance performance, we use the BiLSTM architecture as the base sequence labeler

and augment it with contextual embeddings from four PLMs. All PLMs show substantial improvements over the classical BiLSTM. The best performance is achieved using BiLSTM + MuRIL, with a precision of 0.9612, recall of 0.9575, F1-score of 0.9594, and accuracy of 0.9219. This result demonstrates the effectiveness of MuRIL in capturing reduplication patterns in Indian languages. This may be due to its pretraining on diverse Indian corpora including both monolingual and parallel data. IndicBERT v2 also performs competitively, achieving an F1-score of 0.9572, highlighting the strength of Indic-specific PLMs. XLM-R, while being trained on a large-scale multilingual corpus, also yields strong results (F1: 0.9149), showing good generalization. In contrast, mBERT, though still outperforming classical models, lags behind other PLMs (F1: 0.8411). This is possibly due to its limited capacity and training data compared to XLM-R and MuRIL.

Overall, the results clearly demonstrate that incorporating PLMs significantly boosts the performance of reduplication detection. Among these, MuRIL proves most effective, likely due to its focus on Indian languages and better handling of linguistic diversity and morphological complexity. These findings support the use of PLM-based encoders in sequence labeling tasks involving low-resource or morphologically rich languages.

6.4 Error Analysis

We conduct a detailed error analysis of the BiLSTM+MuRIL model to understand its limitations in reduplication detection. Based on the evaluation, we identify four major types of errors, illustrated through representative examples in Tables 5, 6, 7 and 8.

False positives from non-reduplicative wordforms. Some words in Assamese contain internal repetition or phonological patterns that resemble partial reduplication but are not semantically reduplicated. For example, words like "হাওঁফাওঁ" (/hɑõpʰɑõ/, 'lung') or "মুখামুখি" (/mukhamukhi/, 'face to face') are incorrectly labeled as reduplications by the model (Table 5). These forms exhibit surface-level similarity with reduplicated tokens but are in fact atomic lexical items, leading to over-prediction.

Errors from transliterated terms. The model also mistakenly tags transliterated English words as reduplicated due to character-level repetition or syllabic patterns. As shown in Table 6, terms such as "ডিআৰডিঅ" (/diɑrdio/, 'DRDO'), "য়ুনিয়ন" (/juniɔn/, 'Union'), and "চি চি" (/c c/, 'CC') are incorrectly identified as reduplications. This indicates the model's sensitivity to repetitive sequences, regardless of origin or linguistic function.

Table 5: Examples of false positives from non-reduplicative wordforms error.

Word	IPA	English
হাওঁফাওঁ	/hɑõpʰɑõ/	'lung'
মুখামুখি	/mukhamukhi/	Face to face or facing each other

Table 6: Examples of errors due to transliterated terms.

Word	IPA	English
ডিআৰডিঅ	/diArdio/	'DRDO' - Defence Research and Development Organisation
য়ুনিয়ন	/juniOn/	Union
চি চি	/c c/	CC - Cubic centimeter

Ambiguity in compound structures. Assamese often uses hyphenated compound words to convey inclusiveness or plurality, such as "ছাত্ৰ-ছাত্ৰী" (/satrɔ-satri/, 'male and female students') or "মাছ-মাংস" (/mas-maŋkʰɔ/, 'Fish and meat'). While these constructions exhibit semantic symmetry and surface repetition, they are syntactic compounds rather than true reduplications. As shown in Table 7, the model frequently labels such cases as reduplications, which raises ambiguity around the definition boundary between coordination and reduplication.

Table 7: Examples of errors due to ambiguity in compound structures.

Word	IPA	English
ছাত্ৰ-ছাত্ৰী	/satrɔ-satri/	Students (Male and Female)
যান-বাহন	/jan-vahan/	'All type of vehicles'
কাম-কাজ	/kam-kaz/	'Works and others'
মাছ-মাংস	/mas-maŋkʰɔ/	'Fish and meat'

Morphological challenges with inflected forms. While the model performs reasonably well on base reduplicated forms, it struggles with identifying reduplication when inflectional suffixes are attached. For example, in the word "বুজাবুজি" (/buzabuzi/, 'mutual understanding'), the reduplication is correctly detected. However, in its inflected forms such as "বুজাবুজিৰ" (/buzabuzir/, 'For mu-

tual understanding') or "বুজাবুজিত" (/buzabuzit/, 'In mutual understanding'), the model often fails to identify the base reduplication due to the added morphemes 'ৰ' (/r/) or 'ত' (/t/). This suggests a need for improved morphological robustness in the model's token representations.

Table 8: Examples of error due to morphological challenges with inflected forms

Word	IPA	English
বুজাবুজিৰ	/buzabuzir/	'For mutual understanding'
বুজাবুজিত	/buzabuzit/	'In mutual understanding'

7 Conclusion

In this work, we present AsRED, a manually annotated dataset for reduplication detection in Assamese. The dataset spans multiple domains, including social media, news, textbooks, and government websites. The dataset comprises over 90K reduplicated tokens across 83K sentences. We formulate reduplication detection as a sequence labeling task and evaluate the dataset using classical models such as LSTM, BiLSTM, and CRF. To incorporate contextual information and improve performance, we further experiment with encoder-based pretrained language models, including mBERT, XLM-R, MuRIL, and IndicBERT v2. The results show that pretrained multilingual models, especially MuRIL, achieve better performance compared to classical baselines. We also conduct an error analysis of the best-performing model for reduplication detection. We identify common challenges encountered by the models during reduplication detection. These include false positives from transliterated and compound words, and difficulties in recognizing reduplications in morphologically inflected forms. The analysis highlights specific linguistic properties of Assamese reduplication and points to future directions such as integrating morphological processing and syntactic features to improve detection accuracy.

Limitations

While this work introduces a novel dataset and provides a comprehensive evaluation of reduplication detection in Assamese, it has certain limitations. First, the annotation focuses primarily on surface-level reduplication patterns and may not capture deeper syntactic or semantic variations. Second, the current formulation treats reduplication as a flat sequence labeling task, which may overlook hierarchical or multi-word expressions. Additionally, the models show reduced performance when handling morphologically inflected reduplications or context-dependent reduplicative constructions. Finally, while we include data from multiple domains, the dataset may still not fully represent the diversity of Assamese usage across dialects, informal speech, or creative writing.

Ethical Considerations

This work involves the creation and annotation of a reduplication detection dataset for the Assamese language. All data used in this study were collected from publicly available sources, including government websites, online newspapers, educational materials, and social media platforms. We ensured that no personally identifiable information (PII) or sensitive content was included in the dataset. The annotations were performed manually by native speakers with linguistic expertise.

As Assamese is a low-resource language with cultural and linguistic diversity, we acknowledge the risk of underrepresenting certain dialects or usage patterns. We encourage future work to expand coverage to more dialects and speaker communities. The dataset is intended solely for research and educational purposes, and we caution against its use in applications that could lead to unintended social or cultural biases.

References

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. csli.

L Saikia Bora. 2016. Assamese grammar and usage: An analytical studies of assamese grammar and usage. *Pan-bazar, Guwahati: Chandra Prakash, Guwahati*.

Census. 2020 (accessed April, 2020). *ABSTRACT OF SPEAKERS' STRENGTH OF LANGUAGES AND MOTHER TONGUES - 2011*.

Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of reduplication in bengali corpus and their semantic analysis: A rule based approach. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 73–76.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Satarupa Dattamajumdar. 1999. *A contrastive study of the reduplicated structures in Asamiya Bangla and Odia*. Ph.D. thesis, Department of Linguistics, University of Calcutta, Kolkata, West Bengal.

Satarupa Dattamajumdar. 2001. A contrastive study of the reduplicated structures in asamiya, bangla and odia. *(No Title)*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Hossep Dolatian and Jeffrey Heinz. 2017. Reduplication with finite-state technology. *Proc. CLS*, 53:55–69.

Hossep Dolatian and Jeffrey Heinz. 2019. Redtyp: A database of reduplication with computational models. *Proceedings of the Society for Computation in Linguistics*, 2(1):8–18.

Emmanuel Filiot and Pierre-Alain Reynier. 2016. Transducers, logic and algebra for functions of finite words. *ACM SIGLOG News*, 3(3):4–19.

G. C Goswami. 1987. *Fundamentals of Assamese Grammar (* ɔsomiya byɑkʉrʉnʉr mɔʊlik bichɑr*, 11th Edition(Reprint,2017)*. Bina Library, Panbazar, Guwahati.

Vikas Goswami. 2023. Unlocking assamese derivational morphology: A comprehensive exploration of lexical word categories. *American Journal of Philological Sciences*, 3(09):05–11.

Bernhard Hurch and Veronika Mattes. 2005. *Studies on reduplication*. 28. Walter de Gruyter.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kishorjit Nongmeikapam and Sivaji Bandyopadhyay. 2011. Identification of reduplicated mwes in manipuri: A rule based approach. In *Proceedings of 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL'10)*, pages 49–54.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. Reduplication in Assamese: Identification and Modeling. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to Characters in Neural Sequence Labeling Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan. The COLING 2016 Organizing Committee.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*, volume 4. OUP Oxford.

Carl Rubino. 2013. Reduplication. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Non-Interactive Symbolic-Aided Chain-of-Thought for Logical Reasoning

Phuong Minh Nguyen and Tien Huu Dang and Naoya Inoue

Japan Advanced Institute of Science and Technology
{phuongnm,tiendh,naoya-i}@jaist.ac.jp
Correspondence: phuongnm@jaist.ac.jp

Abstract

This work introduces Symbolic-Aided Chain-of-Thought (CoT), an improved approach to standard CoT, for logical reasoning in large language models (LLMs). The key idea is to integrate lightweight symbolic representations into few-shot prompts, structuring the inference steps with a consistent strategy to make reasoning patterns more explicit within a non-interactive reasoning process. By incorporating these symbolic structures, Symbolic-Aided CoT preserves the generalizability of standard prompting techniques while enhancing the transparency, interpretability, and analyzability of LLM logical reasoning. Extensive experiments on four well-known logical reasoning benchmarks—ProofWriter, FOLIO, ProntoQA, and LogicalDeduction, which cover diverse reasoning tasks and scenarios—demonstrate the effectiveness of the proposed approach, particularly in complex reasoning tasks that require navigating multiple constraints or rules. Notably, Symbolic-Aided CoT consistently improves LLMs' reasoning capabilities across various model sizes and significantly outperforms conventional CoT on three out of four datasets, ProofWriter, ProntoQA, and LogicalDeduction.

1 Introduction

In recent years, pre-trained Large Language Models (LLMs) have achieved exceptional success across a wide spectrum of Natural Language Processing (NLP) tasks (Wei et al., 2022; Shin and Van Durme, 2022; Dubey et al., 2024; Yang et al., 2025). As a result, LLMs have become a central paradigm in NLP research and applications. Their impressive performance is largely attributed to their ability to perform few-shot in-context learning—the mechanism by which models infer solutions based solely on the format and structure of the input prompt, without requiring gradient computations (Brown et al., 2020; Garg et al., 2022; Wei et al., 2022).

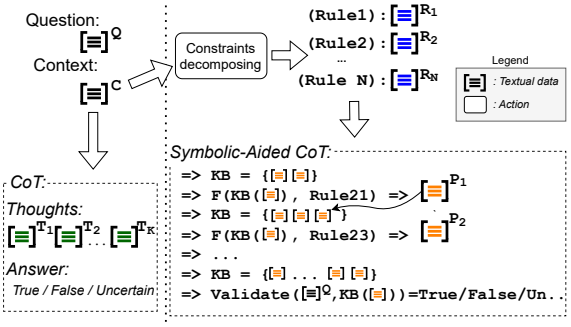


Figure 1: Comparison between standard CoT and Symbolic-Aided CoT for logical reasoning tasks.

Notably, as model size grows, prompting methods that leverage intermediate reasoning steps consistently surpass standard input–output prompting methods. This reasoning strategy, known as Chain-of-Thought prompting (CoT; Wei et al. (2022)), relies on explicitly modeling the reasoning process. For in-context learning, the CoT reasoning technique has demonstrated compelling results across a variety of NLP tasks (Wei et al., 2022; Zhou et al., 2023). Despite recent advancements, applying LLMs to logical reasoning tasks still faces several critical challenges, including conflicts between pretrained knowledge and counterfactual assumptions (Ortu et al., 2024), failures on cyclic inference graphs (Zhang et al., 2025), and planning errors during the solving process (Ye et al., 2023).

To address these issues, various strategies have been proposed, which can be broadly categorized into two main groups: (1) designing an external symbolic solver, which delegates the actual reasoning process to an automated theorem prover (Ye et al., 2023; Pan et al., 2023) or programming languages (Gao et al., 2023; Xu et al., 2024a); and (2) constructing a framework that systematically decomposes complex tasks into subtasks—such as rule selection, premise inference, and scoring—to enhance the overall reliability of the system (Zhang

et al., 2025; Feng et al., 2024; Sun et al., 2024; Xu et al., 2024b). Although the first approaches potentially achieve remarkable performance, they typically require powerful LLMs (Ye et al., 2023; Gao et al., 2023) such as GPT-4 (Achiam et al., 2023) or additional extensive pre-training phase (Xu et al., 2024a; Feng et al., 2024) for successful parsing from problem description to the logical forms (*e.g.,* First-order logics - FOL).

In this study, we focus on the second approach, which aims to improve LLM logical reasoning ***without relying on any external symbolic reasoner or programming language***. Building on insights from recent works (Sun et al., 2024; Qi et al., 2025; Zhang et al., 2025), we target the challenge of designing a mechanism to systematically decompose complex logical reasoning tasks into simpler subtasks that can be solved by the inherent understanding ability of LLMs in a single inference pass. Closely related to our work, Xu et al. (2024a); Feng et al. (2024) recently introduced LLM-based frameworks that leverage first-order logic—*a strict formal language* with well-defined syntax and semantics—to support faithful logical reasoning. However, previous studies (Sun et al., 2024; Xu et al., 2024b; Zhang et al., 2025) primarily rely on self-refinement or interactive (multi-turn) generation, where each turn solves a subtask and its output is passed to the next. They overlook non-interactive reasoning, in which the LLM performs the entire reasoning process without any assistance from external modules or subprocesses. To address this gap, we explore *non-interactive (single-turn) reasoning generation*, allowing a more direct and fair comparison with CoT prompting.

We introduce Symbolic-Aided CoT, a novel variant of CoT (Wei et al., 2022), designed to leverage symbolic representations to enhance the non-interactive logical inference capabilities of LLMs (Figure 1). In conventional CoT, intermediate reasoning steps are provided in few-shot demonstrations as unstructured text, enabling LLMs to approximate the logical reasoning process. However, relying solely on textual descriptions for complex reasoning introduces ambiguity, as the inherent vagueness of natural language limits LLMs' ability to generalize precise reasoning steps. Our core idea is to integrate *lightweight* symbolic structures into few-shot prompts, making the inference steps more transparent and structured, while simultaneously strengthening the induction-head behavior of LLMs (Olsson et al., 2022). Specifically, our Symbolic-Aided CoT prompting framework explicitly outlines essential reasoning substeps: *rule matching*—selecting constraint rules that align with the current state of inference, *new premise inference*—applying the selected rules to generate new premises, and *knowledge base (KB) updating*—appending the newly inferred premises to the KB. By incorporating these symbolic structures, our method preserves the flexibility and general applicability of standard prompting techniques while enhancing both the interpretability and analyzability of LLM reasoning behavior. Empirical evaluations across four reasoning QA benchmarks, ProofWriter, FOLIO, ProntoQA, and LogicalDeduction, demonstrate the effectiveness of our approach, particularly in scenarios involving complex reasoning that requires navigating multiple rules and constraints. Remarkably, our Symbolic-Aided CoT, when applied with open-source LLMs (*e.g.*, Qwen3), achieves performance comparable to that of a complex multi-turn reasoning framework (Sun et al., 2024) built on the powerful GPT-4 model, particularly on the ProofWriter, ProntoQA, and LogicalDeduction datasets.

The remainder of this paper is organized into five sections. Section 2 provides an overview of logical reasoning tasks and compares our approach with prior studies. Section 3 presents the details of our proposed framework and its variants. Section 4 describes the experimental setup and reports the results, with key findings discussed in Section 5. Finally, the conclusions summarize our contributions and outline directions for future work are presented in Section 6.

2 Background and Related Work

2.1 Background

Notation and problem formulation. Logical reasoning is a fundamental NLP task within the Question Answering domain (Weston et al., 2015; Tafjord et al., 2021). In this task, machine learning models are required to answer a question based on a context containing multiple logical conditions or constraints. Formally, we denote the list of rules (or constraints) provided in the context as $\mathcal{R} = \{r_i\}_{0 \leq i < N}$. Given a question (Q), the correct answer (A) must be derived from knowledge supported by a subset of rules $\mathcal{R}^* \subset \mathcal{R}$. To address the challenges posed by logical reasoning, prior research largely falls into two overarching directions:

utilizing *external symbolic solvers* and developing *LLM-based logical solvers*.

In the first approach, the main idea is to leverage LLMs to translate textual descriptions of constraints and the question into formal logical formulas, which are then passed to an explicit symbolic solver (*e.g.,* the Z3 theorem prover[1]) to derive the final answer. Formally, all constraints are aggregated to construct a logical program: $\mathcal{F} = \{\text{LLM}^{\text{lang2logic}}(r_i)\}_{r_i \in \mathcal{R}}$. Similarly, the question is also transformed to the logical form $f^q = \text{LLM}^{\text{lang2logic}}(Q)$. Then, a symbolic reasoner is reasoned over the transformed logical forms to yield the final answer: $\text{SymbolicReasoner}(\mathcal{F}, f^q)$.

In the second approach, the original logical reasoning problem is decomposed into a sequence of subtasks. Each subtask is solved individually, and the process iterates over multiple turns until a specified stopping condition is satisfied, ultimately producing the final result: $\text{Loop}([\text{LLM}^{\text{subtask}_t}(\mathcal{R}, \text{Q})]_t)$ where subtask_t may represent an arbitrary LLM-based unit function such as *rule matching*, *rule inference*, or *new premise scoring*, among others. These sub-tasks are typically carefully designed and arranged with sequential dependencies, with the aim of mitigating hallucinations.

2.2 Related Works

Building on the success of LLMs across a wide range of NLP tasks (Chung et al., 2024; Dubey et al., 2024; Yang et al., 2025), logical reasoning has emerged as a particularly important area of study, serving as a key benchmark for evaluating both the reasoning capabilities and the overall intelligence of these models (Zhou et al., 2023; Feng et al., 2024). As aforementioned, we categorize recent work into two main approaches:

Utilizing external symbolic solvers. In this line of work, the main challenge lies in transforming the description of constraints into logical formulas, $\text{LLM}^{\text{lang2logic}}$, effectively functioning as a semantic parser. More specifically, Yao et al. (2023) introduced a method to enrich CoT prompting by leveraging results from external APIs, invoking functions to retrieve supplementary information. Building on this idea, Gao et al. (2023) proposed an approach that augments the intermediate reasoning steps of CoT reasoning through a runtime environment (*e.g.,* a Python interpreter), which has proven particularly effective for mathematical reasoning tasks. Further, Pan et al. (2023); Ye et al. (2023); Olausson et al. (2023) enhanced the performance of logical reasoning by translating all constraints in a sample into logical forms, completing a logical program, and solving it using an independent symbolic reasoner. In addition, Xu et al. (2024a) strengthened the logical-form parsing process of the open-source LLMs by leveraging fine-tuning on a large-scale, curated dataset. In contrast, our method does not rely on any external symbolic solver; instead, it integrates symbolic syntax directly into the reasoning process, aiming to enable the LLM itself to reason as a symbolic solver.

LLM-based logical solvers. This approach leverages LLMs directly by designing frameworks that decompose complex reasoning tasks into smaller, manageable subtasks—such as rule selection, premise derivation, and scoring—to enhance overall reasoning robustness (Zhang et al., 2025; Feng et al., 2024; Sun et al., 2024). In particular, Sun et al. (2024); Xu et al. (2024a) introduced a framework that enables LLMs to uncover hidden premises and integrates scoring components through a multi-turn reasoning process. Similarly, Zhang et al. (2025) proposed a framework that incrementally refines reasoning via three subcomponents: Proposer, Verifier, and Reporter. Finally, Feng et al. (2024) presented LoGiPT, a method that enhances LLMs' ability to function as logical solvers by learning the reasoning process step by step through additional training on large-scale data collected from the reasoning traces of external symbolic solvers. Compared to our work, LoGiPT similarly performs step-by-step reasoning as a logical solver and can produce a proof tree at the end; however, *our method achieves competitive performance without the need for any additional training.*

3 Methodology

In this section, we present our *Symbolic-Aided CoT* method, its variants, and the motivation behind it in comparison with baseline prompting techniques: *Standard* - which directly provides an answer without any reasoning - and *CoT* (Wei et al., 2022) - which produces an answer accompanied by step-by-step reasoning. All of these prompting methods are augmented with a hard-selected few-shot examples included in the prompt (Qi et al., 2025). All prompts are directly fed forward through the LLM to obtain the final predicted

[1] https://github.com/Z3Prover/z3

answer in a single turn:

$$A^{\text{out}} = \text{LLMs}(\text{prompting}(\mathcal{R}, Q)) \quad (1)$$

In this setup, the LLM is solely responsible for generating the desired answer given the contextual input. The prompting component consists of a few-shot template designed to help the LLM understand the task description while leveraging its own knowledge to reason over the list of provided rules in the contextual information. For clarity, the templates for *Standard* and *CoT* prompting are shown in the first two rows of Table 1.

Table 1: Template of input and output for prompting techniques: *Standard* and *CoT* and *Symbolic-Aided CoT*.

Standard *(Input)*	Context: *[[All constraints, $\mathcal{R}$]]* Question: *[[Content of the question, Q]]* Options: A) True B) False C) Uncertain
Standard *(Output)*	The correct option is: { "answer": *[[A]]* }
CoT *(Input)*	Context: *[[All constraints, $\mathcal{R}$]]* Question: *[[Content of the question, Q]]* Options: A) True B) False C) Uncertain

CoT *(Output)*	The correct option is:
	{
	"reasoning": *[[reasoning content]]* ,
	"answer": *[[A]]*
	}

Symbolic-Aided CoT *(Input)*	### Let us define F as a function that infers new premises based on a given list of facts and rules. Using these facts and rules, provide a reasoning path that leads to one of the values of a Validate function: True, False, or Uncertain. ——— ### Example1: Given list of facts and rules: # (Rule[[i]]): *[[content of $r_i \in \mathcal{R}$]]* $\cdots$ # (Question): *[[content of the question, Q]]*
Symbolic-Aided CoT *(Output)*	# (Answer): Start from the object and their condition mentioned in the question to collect relevant facts: # KB = { } => F(KB(*[[premises in KB]]*), Rule[[i']]) => *[[inferred premises]]* # KB = {*[[KB values for each reasoning steps]]*} # validate the question with the current inferred premise => Validate(Question=*[[Q]]*, KB(*[[selected premise]]*)) = *[[A]]*.

3.1 Symbolic-Aided CoT

We formulate logical reasoning tasks into three fundamental sub-tasks, namely, reasoning operators: *rule matching*, *rule inference*, and *knowledge base updating*. Previous frameworks, such as DetermLR (Sun et al., 2024) and CR (Zhang et al., 2025), were also built on carefully designed unit operators, integrating them with procedural programming to process the outputs of LLMs. The key difference between our Symbolic-Aided CoT and these approaches is that Symbolic-Aided CoT is conceived entirely as an LLM-driven program. In our design, the LLM is expected to learn the flow of the logical reasoning process from a few-shot demonstration. To this end, the LLM has full visibility of all sub-reasoning steps and autonomously decides which step to execute next. The overview of our Symbolic-Aided CoT is presented in the third row of Table 1, which illustrates the instruction text, list of rules, question, and reasoning-flow examples. For a clearer explanation, we elaborate on the two gray blocks shown in this table, which pertain to rule tagging and reasoning operators in our Symbolic-Aided CoT.

Rule tagging. In preparing the prompt input, we first segment the contextual information into a list of rules by splitting it into individual sentences using the NLTK toolkit[2]. Each sentence is then tagged with its order index (*e.g.,* `Rule5` for the fifth sentence), allowing the LLM to track and reference the reasoning steps. Here, we assume that the LLMs can link each rule's content to its corresponding tag and reference this symbol appropriately in the reasoning flow in subsequent steps.

Reasoning operators. This demonstration serves as the core example that enables LLMs to learn, in context, the pattern for solving logical reasoning tasks. We use a set of symbols, similar to those in programming languages, to represent the inference flow (see Figure 2). At each reasoning step, the LLM selects the relevant rules and premises from the current knowledge base (KB) to infer new premises. Each newly inferred premise is then appended to the KB for use in the next inference step. A breadth-first search strategy is applied to traverse the nodes (premises), as illustrated in Figure 2. Inspired by how humans solve such tasks, we maintain a KB state to prevent cyclical inference loops: if a newly inferred premise already exists in the KB, it is not added again. All patterns of rule selection, inference, and KB updating are implicitly conveyed within the demonstrations provided in the few-shot prompts, allowing the LLM to internalize these reasoning steps.

4 Experiment and Analysis

In this section, we present a detailed description of our experiments, together with the results and analyses, to assess the effectiveness of our Symbolic-Aided CoT prompting in comparison to standard CoT and prior methods.

[2] https://www.nltk.org/

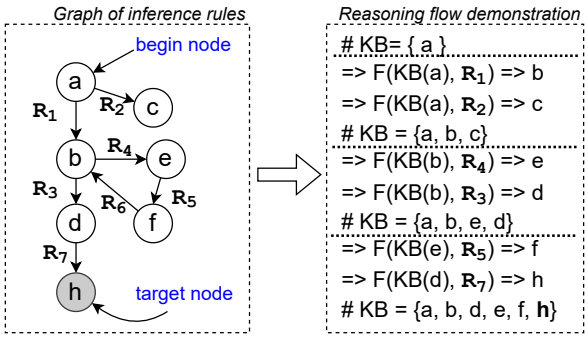


Figure 2: **Left**: Graphical model of inference rules. **Right**: Reasoning flows in the Symbolic-Aided CoT demonstrations.

4.1 Experimental Setup

Datasets. We conducted experiments on four well-known benchmark datasets about the logical reasoning task: (1) **ProofWriter** (Tafjord et al., 2021) - we use the subset under the open-world assumption, where each sample has three possible answer options: true, false, or unknown. Following Pan et al. (2023), we evaluate on the subset with the longest reasoning depth (5 hops), which contains 600 cases. (2) **FOLIO** (Han et al., 2024) is a challenging, expert-curated dataset for logical reasoning that contains rules closely aligned with real-world knowledge. Following the setup of previous work (Sun et al., 2024), we evaluate our method on a subset of this dataset comprising 204 examples. (3) **ProntoQA** introduced by Saparov and He (2023) - similar to the ProofWriter dataset, we also choose the hardest subset of this data with 5-hop reasoning across 500 samples for the evaluation, following previous works (Sun et al., 2024; Qi et al., 2025); (4) **LogicalDeduction** (Srivastava et al., 2023) is a dataset for logically identifying the order of objects given a list of description constraints. We follow the previous setting from Sun et al. (2024), using 300 evaluation samples containing all subsets of three, five, and seven objects (the greater the number of objects, the more complex the logical reasoning required to determine their order).

Evaluation metric. In order to evaluate system performance, we use Accuracy as the metric, which is standard and allows for direct and fair comparison with previous works (Pan et al., 2023; Sun et al., 2024; Qi et al., 2025).

Setting. We conducted our experiments primarily on open-source LLMs, including `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and the `Qwen-3` models (Yang et al., 2025). Specifically, we aim to evaluate the effectiveness of our Symbolic-Aided CoT compared to standard CoT and standard prompting on these LLMs. In addition, we performed extensive experiments on the `Qwen-3` models across various sizes—`1.7B`, `4B`, `8B`, and `32B`—to further assess the scalability and effectiveness of our proposed prompting method. Finally, for comparison with previous approaches, we also conducted extensive experiments on a powerful closed-source LLM, `GPT-4` (Achiam et al., 2023), to evaluate our proposed method. For open-source models, we use greedy decoding to generate the answer *i.e.,* the token with the maximum logit is picked.

4.2 Main Results

For the main results across all benchmark datasets, ProofWriter, FOLIO, and ProntoQA, we present the performance of our Symbolic-Aided CoT compared with Standard Prompting, CoT Prompting, and previous methods in Table 2.

Overall performance. These results demonstrate the clear superiority of our proposed method over the standard CoT on three datasets, ProofWriter, ProntoQA, and LogicalDeduction. Notably, on the ProofWriter dataset, Symbolic-Aided CoT significantly outperforms CoT, achieving improvements of 15% and 21%, 22% on the `Qwen-3-8B`, `Qwen-3-14B`, and `Llama3.1-8B-Instruct` models, respectively. In addition, the improvement is clearly observed on the LogicalDeduction and ProntoQA datasets across three open-source LLMs. These findings further highlight that the degree of improvement varies across different LLMs. The effectiveness of Symbolic-Aided CoT largely depends on each model's ability to understand logical relationships and recognize logical matching patterns embedded within the few-shot prompts. Moreover, our method is simple yet effective, achieving competitive performance compared to prior works such as Logic-LM (Pan et al., 2023) and DetermLR (Sun et al., 2024), even when those methods are supported by a more powerful `GPT-4` model.

On the FOLIO dataset, the results show that the Symbolic-Aided approach has a weakness compared to CoT prompting, especially with the `Qwen-14B` LLM. We found that the FOLIO dataset

is specifically designed by experts to cover various aspects of factual knowledge, which allows the CoT prompting technique to leverage this advantage (leaking factual knowledge) when solving questions. For example, in a question about the tennis player Djokovic, CoT prompting tends to use external knowledge such as *"Djokovic is famous and is an athlete"* , which is not provided in the set of facts in the context, to support the inference flow. In contrast, our Symbolic-Aided CoT approach relies strictly on the inference rules given in the context.

Finally, we evaluate our proposed method in the setting that uses `GPT-4` as the backbone LLM for the reasoning task (last row of Table 2). Compared to the SymbCoT framework (Xu et al., 2024a), our method achieves superior performance on ProntoQA but lower performance on other datasets. This difference can be attributed to SymbCoT's use of complex interactive reasoning sub-steps—such as translator, planner, solver, and verifier—each supported by carefully designed prompts tailored to the specific sub-step and logical reasoning task. Overall, our method surpasses the performance of previous methods on the ProntoQA dataset and achieves remarkable results on the ProofWriter, FOLIO, and LogicalDeduction datasets. These results demonstrate the robustness of our approach, even in the stringent setting that uses only non-interactive inference without the support of an external solver or multi-step inferences, such as the DetermLR (Sun et al., 2024) or LogicLM (Pan et al., 2023) approaches.

Impact of model size on performance. To assess the effectiveness of our Symbolic-Aided CoT across different model sizes, we conducted experiments using various Qwen LLMs on ProofWriter (Figure 3) and LogicalDeduction (Figure 4) datasets. These results demonstrate that our method consistently outperforms both CoT and standard prompting across model sizes. Furthermore, our approach appears to encourage LLMs to more explicitly articulate the underlying logical reasoning patterns, even in small-scale models. For example, on the ProofWriter dataset, `Qwen3-8B` achieves performance comparable to that of the `32B` model. On the LogicalDeduction dataset, `Qwen3-8B` attains 86.9% of the performance of the `32B` model. We argue that our Symbolic-Aided CoT decomposes the original complex logical reasoning tasks into sub-reasoning operations–such as selecting rules, generating new premises, and extending KB premises–that can be effectively addressed by smaller language models.

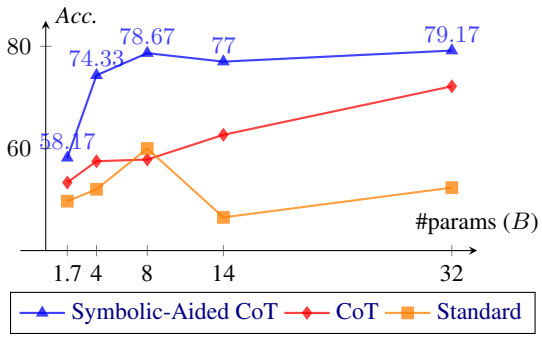


Figure 3: Performance across different model sizes of `Qwen-3` with three prompting techniques on the ProofWriter dataset.

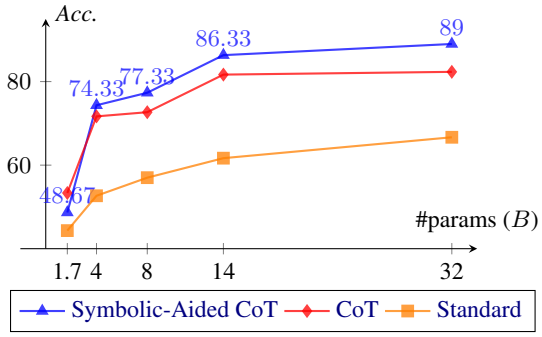


Figure 4: Performance across different model sizes of `Qwen-3` with three prompting techniques on the LogicalDeduction dataset.

Ablation studies. For evaluating the contribution of each sub-component in our Symbolic-Aided CoT prompting, we conduct two ablation studies: (1) removing the KB-tracking variables (SymbolA.CoT$^{-\text{KB tracking}}$), which removes the text segment *"#KB = [[KB values for each reasoning step]]"* in Table 1, and (2) removing the symbolic Validate function (SymbolA.CoT$^{-\text{Validate}}$), which removes the text segment *"Validate(Question=[[Q]], KB([[selected premise]]))"* in Table 1. The ablation results (shown in Figure 5) indicate that KB-tracking variables play an important role in the reasoning process, helping LLMs avoid loops in the conferencing process. Furthermore, KB-tracking intuitively provides additional features to the hidden representation of premises, allowing the model to distinguish inferred premises from conditional premises in the constraint rules. In another aspect, the symbolic Validate function in our Symbolic-Aided CoT helps LLMs refer back to the original question

Table 2: Performance comparison among different methods. *E.Solver* refers to the system supported by an external symbolic solver module. *ICL* stands for in-context learning, and *Supervised FT* stands for the supervised fine-tuning approach. The **best** is marked.

Methods	Learning paradigm	Interaction mode	ProofWriter	FOLIO	ProntoQA	L.Deduction
Llama3.1-8B-Instruct						
Fine-tuned ID (Qi et al., 2025)	Supervised FT	Non-Interactive	71.67	70.00	—	—
Standard	ICL few-shot	Non-Interactive	36.83	53.92	51.80	40.67
CoT	ICL few-shot	Non-Interactive	44.83	**56.86**	74.00	58.00
Symbolic-Aided CoT (ours)	ICL few-shot	Non-Interactive	**68.67**	55.88	**89.00**	**59.33**
Qwen3-8B						
Standard	ICL few-shot	Non-Interactive	60.00	62.25	80.80	57.00
CoT	ICL few-shot	Non-Interactive	57.83	**66.67**	95.80	72.67
Symbolic-Aided CoT (ours)	ICL few-shot	Non-Interactive	**78.67**	65.69	**97.20**	**77.33**
Qwen3-14B						
Standard	ICL few-shot	Non-Interactive	46.50	67.16	77.80	61.67
CoT	ICL few-shot	Non-Interactive	62.67	**74.02**	97.20	81.67
Symbolic-Aided CoT (ours)	ICL few-shot	Non-Interactive	**77.00**	65.20	**97.80**	**86.33**
GPT-4						
CoT (Sun et al., 2024)	ICL few-shot	Non-Interactive	67.41	67.65	91.00	73.33
Logic-LM (Pan et al., 2023)	ICL few-shot	Interactive +E.Solver	79.66	78.92	83.20	87.63
DetermLR (Sun et al., 2024)	ICL few-shot	Interactive +Programming	79.17	75.49	98.60	85.00
SymbCoT (Xu et al., 2024a)	ICL few-shot	Interactive +Programming	**82.50**	**83.33**	99.60	**93.00**
Symbolic-Aided CoT (ours)	ICL few-shot	Non-Interactive	77.09	74.51	**100.00**	86.33

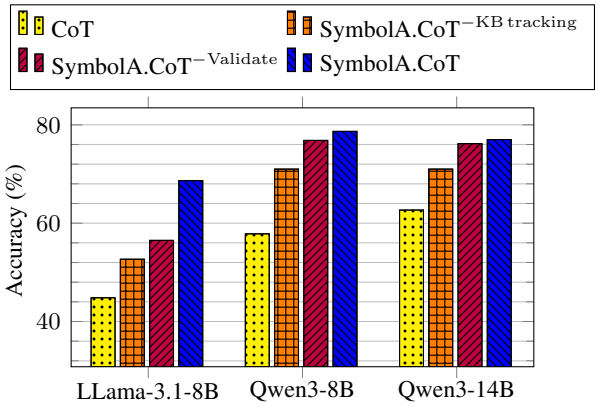


Figure 5: Ablation study on the Symbolic-Aided CoT (SymbolA.CoT).

in the context to select the appropriate premise for logical matching and producing the final answer.

4.3 Result Analysis

Confusing ratio. Here we report the confusion matrices (Figure 6) of the answers in the ProofWriter dataset, generated by the `Qwen-8B` model. For both methods, the original CoT and our Symbolic-Aided CoT, the recall score for False questions is the highest, followed by True and Uncertain. This is due to the complex nature of the logical reasoning task, which involves multi-hop reasoning steps; reasoning paths leading to wrong conclusions are typically more numerous than those leading to correct ones. Comparing our Symbolic-Aided CoT to the original CoT, our method shows improvement across all three question types. The main improvement comes from reducing confusion in Uncertain questions, decreasing misclassification as True or False. We argue that, through symbolic injection, our method encourages clearer logical patterns and structure, thereby enhancing the logical reasoning ability of LLMs.

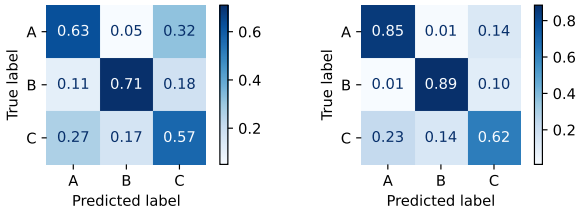


Figure 6: Result comparison with confusion matrices between our Symbolic-Aided CoT (right) and the original CoT (left). The labels A, B, and C refer to the answers True, False, and Uncertain, respectively.

Semantic representation of symbolic tokens. Here, Figure 7 visualizes the semantic representation of symbolic tokens using principal component analysis (PCA) based on the last layer's output hidden states of the LLM `Qwen-3-14B`. This experiment aims to analyze, at a low level, how LLMs understand symbolic tokens in our Symbolic-Aided CoT prompting. We found that LLMs can clearly distinguish the meaning of symbolic tokens (purple data points) from sample content tokens in our proposed prompting method. This is because these logical tokens play the role of structuring the inference flow (latent reasoning pattern) of LLMs, which is separate from the content words in facts

and rules. Through few-shot in-context learning, these tokens are represented in a distinct semantic space. Via the self-attention mechanism, logical tokens are paired with content tokens to yield features specific to reasoning operators (such as matching rules or inferring new premises). This suggests that LLMs can uncover the hidden patterns of logical reasoning operators implied by the symbolic tokens within few-shot learning.

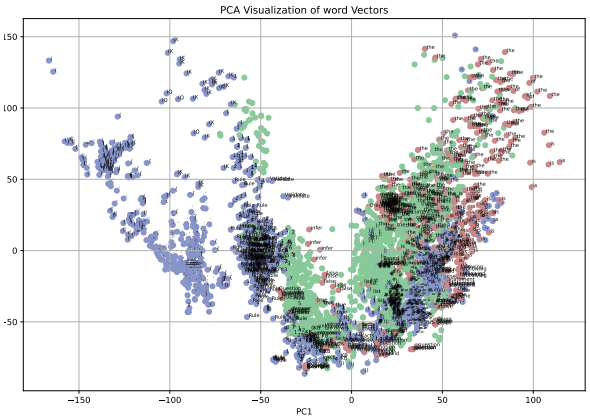


Figure 7: Visualization of last-layer word hidden states from `Qwen3-8B`, with dimensionality reduced via PCA, in the Symbolic-Aided CoT setting on the ProofWriter dataset. The purple, brown, and green data points represent the embeddings of logical symbols (*e.g.,* "=>", "KB"), instruction tokens (*e.g.,* "Let," "us," "define"), and sample content tokens (*e.g.,* "cat," "mouse"), respectively.

Case studies. Based on case studies of incorrect predictions, we identified several improvement scenarios exhibited by Symbolic-Aided CoT compared to the standard CoT prompting technique (Table 3 in Appendix A): (1) *Hallucinated Inference Rules:* LLMs often generate inference rules that are either fabricated, logically invalid, or misaligned with the intended knowledge base (first row in Table 3). This phenomenon is caused by the counterfactual or pre-trained knowledge embedded within LLMs can conflict with in-context rules or premises provided at inference time. This undermines the assumption that the model will reason strictly within the given context or constraints; (2) *Unstoppable Inference Flow:* The reasoning process lacks a clear halting condition (third row in Table 3). The model continues generating premises without a mechanism to determine when inference should stop, leading to uncontrolled or incoherent inference chains. This highlights the need to explicitly track and manage the state of the knowledge base (KB) during reasoning; (3) *Failure on Cyclic Inference Graphs:* When the inference space forms a cyclic graph, LLMs often fail—either entering infinite reasoning loops or struggling to resolve the cycle (second and third rows in Table 3). These models lack the structural awareness to detect and handle loops in reasoning chains; (4) *Rule Matching Errors:* LLMs frequently fail to apply inference rules correctly due to poor condition matching (fourth row in Table 3). The model may skip necessary preconditions or generate incorrect intermediate steps, breaking the logical flow of multi-step reasoning.

5 Discussion

Logical reasoning tasks have attracted numerous research works recently (Pan et al., 2023; Ye et al., 2023; Olausson et al., 2023; Sun et al., 2024; Zhang et al., 2025), especially following the massive success of LLMs. Unlike previous methods, by proposing Symbolic-Aided CoT, we primarily aim to enhance the logical reasoning ability of LLMs rather than simply build a system to improve performance on logical reasoning tasks. For example, frameworks such as Logic-LM (Pan et al., 2023) and SatLM (Ye et al., 2023) use LLMs only to translate logical problems into inputs for explicit symbolic reasoners. Frameworks like CR (Zhang et al., 2025) and DetermLR (Sun et al., 2024) leverage LLMs to perform small constituent logical reasoning steps, rather than directly evaluating the LLM's logical reasoning ability on the entire problem.

The experimental results show that our Symbolic-Aided CoT prompting technique is reliable and effectively improves the logical reasoning ability of LLMs, even for small model sizes. Our prompting method is simple, yet effective and flexible, allowing customization for any logical reasoning task. It can also yield proof trees that facilitate explanation and enhance transferability.

6 Conclusion

In this work, we introduced Symbolic-Aided CoT, a novel prompting technique for non-interactive logical reasoning, which achieves superior performance on well-known benchmark datasets—most notably ProofWriter, ProntoQA, and LogicalDeduction. Our method is deliberately simple to preserve generalizability and shows strong potential for extension to other reasoning tasks. For future work, Symbolic-Aided CoT, grounded in structural

characteristics, could be combined with mechanisms for refining the latent semantic vector space, thereby further improving the faithfulness and reliability of LLMs' reasoning capabilities.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22H00524 and the Nakajima Foundation. We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use".

Limitations

We discuss the following limitations and future works: We have evaluated the proposed method on four widely used logical reasoning benchmarks. However, they are mostly synthetic; incorporating real-world datasets or diverse reasoning tasks (*e.g.,* commonsense reasoning) would strengthen claims of generalizability. Relying solely on automatic metrics like accuracy overlooks qualitative aspects; integrating human evaluations to assess reasoning faithfulness and interpretability would offer a more holistic validation. Future research could also explore the method's robustness to adversarial perturbations, sensitivity to prompts, and scalability to longer reasoning chains.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. 2024. Language models can be deductive solvers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4026–4042, Mexico City, Mexico. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Theo X. Olausson, Alex Gu, Ben Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger P. Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits*

Thread. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. 2025. Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth International Conference on Learning Representations*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425, Seattle, United States. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2024. DetermLR: Augmenting LLM-based logical reasoning from indeterminacy to determinacy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9828–9862, Bangkok, Thailand. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. 2024a. Symbol-LLM: Towards foundational symbol-centric interface for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13091–13116, Bangkok, Thailand. Association for Computational Linguistics.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024b. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. SatLM: Satisfiability-aided language models using declarative prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew C Yao. 2025. Cumulative reasoning with large language models. *Transactions on Machine Learning Research*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Improving examples

In this section, we present detailed improvement examples from the ProofWriter dataset for our

Symbolic-Aided CoT method compared to the CoT method, in the setting using the `Qwen3-8B` model (Table 3).

Table 3: Improving examples from the ProofWriter dataset for our Symbolic-Aided CoT method compared to the CoT method. The red text refers to incorrect logical reasoning or hallucinated content, and the green text refers to the final premises related to the target question.

Id	Context, Question, and Answer	CoT's output + *Error Type*	Symbolic-Aided CoT's output + *Error Type*
ProofWriter-_AttNeg-OWA-D5-596_Q6	### Given list of facts and rules: # (Rule1): Anne is not rough. # (Rule2): Bob is blue. # (Rule3): Erin is not furry. # (Rule4): Erin is red. # (Rule5): Gary is big. # (Rule6): Gary is not cold. # (Rule7): Gary is furry. # (Rule8): If something is blue then it is rough. # (Rule9): Red things are rough. # (Rule10): If Gary is rough then Gary is not blue. # (Rule11): Rough things are red. # (Rule12): Big things are quiet. # (Rule13): All cold things are big. # (Rule14): If something is red then it is big. [...] # (Question): Based on the above information, is the following statement true, false, or unknown? Erin is not quiet. # (Answer): **False**	Erin is red. Red things are rough. If something is rough then it is not quiet. So Erin is not quiet. *(1) Hallucinated Inference Rules*	Start from the object and their condition mentioned in the question to collect relevant facts: Erin, is not quiet # KB = { } => Rule3 = 'Erin is not furry' => Rule4 = 'Erin is red' # KB = {Erin is not furry, Erin is red} => F(KB('Erin is red'), Rule9) => 'Erin is rough' # KB = {Erin is not furry, Erin is red, Erin is rough} => F(KB('Erin is red'), Rule11) => 'Erin is red' (already in KB) # KB = {Erin is not furry, Erin is red, Erin is rough} => F(KB('Erin is red'), Rule14) => 'Erin is big' # KB = {Erin is not furry, Erin is red, Erin is rough, Erin is big} => F(KB('Erin is big'), Rule12) => 'Erin is quiet' # KB = {Erin is not furry, Erin is red, Erin is rough, Erin is big, Erin is quiet} # valid the question with current inferred premies => Validate(Question='Erin is not quiet', KB('Erin is quiet'))=**False**.
ProofWriter-_RelNoneg-OWA-D5-356_Q4	### Given list of facts and rules: # (Rule1): The cow eats the mouse.[...] # (Rule15): The squirrel likes the cow. # (Rule16): The squirrel likes the mouse. # (Rule17): If something likes the cow then it visits the mouse. # (Rule18): If something likes the cow then the cow is green. # (Rule19): If something eats the squirrel and the squirrel likes the rabbit then it eats the rabbit. # (Rule20): If something is cold then it likes the cow. # (Rule21): If something visits the squirrel then the squirrel likes the rabbit. # (Rule22): If the cow likes the rabbit then the cow is cold.[...] # (Question): Based on the above information, is the following statement true, false, or unknown? The squirrel does not visit the mouse. # (Answer): **False**	The cow likes the rabbit. If the cow likes the rabbit then the cow is cold. The cow is cold. The cow likes the squirrel. If something is cold then it likes the cow. The cow likes the cow. The cow visits the mouse. If something likes the cow then it visits the mouse. The cow visits the mouse. The squirrel likes the cow. If something visits the squirrel then the squirrel likes the rabbit. The squirrel likes the rabbit. The squirrel eats the rabbit. If something likes the rabbit and it visits the mouse then the mouse visits the squirrel [...] *(3) Failure on Cyclic Inference Graph* *(4) Rule Matching Errors*	Start from the object and their condition mentioned in the question to collect relevant facts: The squirrel, does not visit the mouse # KB = { } => Rule1 = 'The cow eats the mouse' # KB = {cow eats mouse, [...], squirrel is rough, squirrel likes cow, squirrel likes mouse} => F(KB('squirrel likes cow'), Rule17) => 'squirrel visits mouse' # KB = {cow eats mouse, [...] squirrel is rough, squirrel likes cow, squirrel likes mouse, squirrel visits mouse} # valid the question with current inferred premies => Validate(Question='The squirrel does not visit the mouse', KB('squirrel visits mouse')) = **False**.
ProofWriter-_RelNeg-OWA-D5-688_Q22	### Given list of facts and rules:[...] # (Rule6): The lion needs the mouse. # (Rule7): The mouse is round. # (Rule8): The mouse likes the lion. # (Rule9): The mouse needs the lion.[...] # (Rule15): If something likes the cat and it does not visit the cat then it visits the lion. # (Rule16): If the lion is green and the lion visits the mouse then the mouse is red. # (Rule17): All red things are nice. # (Rule18): If something likes the lion then it is red. # (Rule19): If the mouse visits the cat and the mouse needs the squirrel then the mouse does not like the cat. # (Rule20): Nice things are big. # (Rule21): If something is big then it visits the squirrel. # (Rule22): If the mouse visits the squirrel then the mouse does not visit the lion. # (Question): Based on the above information, is the following statement true, false, or unknown? The lion likes the mouse. # (Answer):": **Uncertain**	The mouse likes the lion. The mouse needs the lion. The mouse visits the cat. The mouse visits the squirrel. The mouse does not visit the lion. The mouse is round. The mouse is red. The mouse is nice. The mouse is big. The mouse is nice. The mouse is big. The mouse visits the squirrel. The mouse does not visit the lion. The mouse is red. The mouse is nice. The mouse is big. The mouse is nice. The mouse is big. The mouse is nice. The mouse is big. The mouse is nice. The mouse is big. [...] *(2) Unstoppable Inference Flow* *(3) Failure on Cyclic Inference Graphs*	Start from the object and their condition mentioned in the question to collect relevant facts: The lion, likes the mouse # KB = { } [...] # KB = {The lion needs the mouse, The mouse likes the lion, The mouse is red, The mouse is nice} => F(KB('The mouse is nice'), Rule20) => 'The mouse is big' # KB = {The lion needs the mouse, The mouse likes the lion, The mouse is red, The mouse is nice, The mouse is big} => F(KB('The mouse is big'), Rule21) => 'The mouse visits the squirrel' # KB = {The lion needs the mouse, The mouse likes the lion, The mouse is red, The mouse is nice, The mouse is big, The mouse visits the squirrel} => F(KB('The mouse visits the squirrel'), Rule22) => 'The mouse does not visit the lion'[...] # valid the question with current inferred premies => Validate(Question='The lion likes the mouse', KB('The mouse likes the lion')) =**Uncertain**.
ProofWriter-_AttNoneg-OWA-D5-245_Q19	### Given list of facts and rules:[...] # (Rule4): Erin is nice.[...] # (Rule11): Furry people are smart. # (Rule12): If someone is smart and nice then they are round. # (Rule13): Cold people are red. # (Rule14): If someone is quiet then they are nice. # (Rule15): All red people are furry. # (Rule16): All smart, furry people are nice. # (Question): Based on the above information, is the following statement true, false, or unknown? Erin is not quiet. # (Answer) : **Uncertain**	Erin is nice. If someone is nice then they are quiet. So Erin is quiet. Therefore, Erin is quiet, so the statement 'Erin is not quiet' is false. *(4) Rule Matching Errors*	Start from the object and their condition mentioned in the question to collect relevant facts: Erin, is not quiet # KB = { } => Rule4 = 'Erin is nice' # KB = {Erin is nice} => F(KB('Erin is nice'), Rule14) => 'Erin is quiet' # KB = {Erin is nice, Erin is quiet} # valid the question with current infered premies => Validate(Question='Erin is not quiet', KB('Erin is quiet')) = **False**. *(4) Rule Matching Errors*

Sensory and Affective Dimensions in Mandarin Monosyllabic Adjectives

Yimei Shao, Yu-Yin Hsu, Chu-Ren Huang
Department of Language Science and Technology
The Hong Kong Polytechnic University, Hong Kong, SAR
{yimei.shao, yyhsu, churen.huang}@polyu.edu.hk

Abstract

This study examined how sensory modalities and affective dimensions jointly shape the semantics of Mandarin monosyllabic adjectives—a lexical class that is both perceptually rich and emotionally salient yet understudied in prior research. Drawing on embodied cognition theory, we integrated newly collected perceptual strength ratings (visual, auditory, haptic, gustatory, olfactory) for 165 adjectives with existing sensory norms (Chen et al., 2019) and affective ratings (valence, arousal; Peng et al., 2024), yielding a dataset of 298 items. Analyses revealed an asymmetrical sensory organization, with strong gustatory–olfactory coupling and relative independence of auditory imagery. Visual strength predicted more positive valence, whereas auditory and haptic strength predicted higher arousal, and modality exclusivity differentiated pleasantness from arousal. These results indicate that sensory experience systematically contributes to affective semantics, supporting embodied accounts of conceptual meaning. The study provides cross-linguistic insights into the perceptual grounding of emotion and offers a resource for future psycholinguistic and computational modeling.

1 Introduction

"We walked down the path to the well-house, attracted by the fragrance of the honeysuckle with which it was covered. Someone was drawing water and my teacher placed my hand under the spout. As the cool stream gushed over one hand she spelled into the other the word water, first slowly, then rapidly. I stood still, my whole attention fixed upon the motions of her fingers. Suddenly I felt a misty consciousness as of something forgotten—a thrill of returning thought; and somehow the mystery of language was revealed to me. I knew then that 'w-a-t-e-r' meant the wonderful cool something that was flowing over my hand. That living word awakened my soul, gave it light, hope, joy, set it free!"

— Helen Keller, The Story of My Life (1903)

This vivid account captures a profound insight: word meaning can arise directly from embodied sensory experience. It illustrates how perception and language intertwine—the tactile sensation of flowing water, the auditory rhythm of spelling, and the affective awakening that follows.

A single word can engage multiple senses and emotions at once. In Chinese, *清* (clear) may evoke visual imagery, the sound of flowing water, and a sense of calm; *香* (fragrant) blends olfactory impressions with pleasant affect; *刺耳* (piercing) conveys auditory harshness and negative emotion. Such examples show that perceptual and affective experiences are systematically intertwined in lexical representation—yet this interplay remains insufficiently examined in Chinese.

The idea that language both reflects and shapes human thought has long guided cognitive and linguistic theory (Hardin, 1993; Rosch, 1974; Vygotsky, 2000; Whorf, 1956). Early work on linguistic relativity argued that linguistic structure provides the scaffolding through which perceptual and conceptual distinctions are formed. This tradition highlights that meaning is not neutral but shaped by how language encodes experience.

Building upon this tradition, embodied cognition approaches extend the discussion by emphasizing that meaning is not only linguistically mediated but also grounded in direct sensory and emotional experience (Barsalou, 1999, 2008). Within the embodied cognition framework, meaning is grounded in the reactivation of sensory, motor, and emotional traces associated with a word's referent (Barsalou, 1999, 2008; Zwaan et al., 2002). Research on sensory dimensions has typically focused on five modalities—vision, audition, touch, taste, and smell—measured using modality strength and exclusivity norms (Lynott and Connell, 2009). In parallel, studies on affective dimensions have identified valence (pleasantness) and arousal (emotional

intensity) as key factors influencing lexical processing and conceptual organization (Bradley and Lang, 1999; Warriner et al., 2013). While each dimension has been widely studied for English, relatively few studies have addressed how sensory profiles systematically relate to emotional evaluations of words.

In Chinese, research on sensory norms has begun to emerge (e.g. Chen et al., 2019; Zhong and Ahrens, 2023), but it remains fragmented in lexical coverage, with limited attention to adjectives and minimal integration with affective ratings. Particularly, adjectives merit focused investigation: they are semantically compact, frequently encode perceptual features, and often carry strong emotional connotations. Moreover, cross-modal patterns such as gustatory–olfactory coupling or the independence of auditory modality, reported in English and other languages, have not been systematically tested in Mandarin adjectives.

Beyond modality strengths, embodied cognition posits that sensory experience forms a fundamental component of emotional representation—bodily sensations not only accompany emotions but constitute part of their experiential core (Damasio, 1994; Prinz, 2005). Yet it remains unclear to what extent different types of sensory experience contribute to emotional activation, and whether specific sensory channels (e.g., visual, auditory, haptic) are more strongly linked to affective dimensions such as pleasantness and arousal. Addressing this question helps clarify how emotion is grounded in distinct perceptual systems within the mental lexicon.

In addition to modality-specific effects, the overall structure of sensory engagement may also shape emotional meaning. Modality exclusivity, which captures the extent to which a word's meaning is concentrated in a single sensory channel (Lynott and Connell, 2009), provides a useful index of the breadth versus focus of embodied experience. Words that engage multiple senses may support richer, integrative simulations, whereas those confined to one dominant modality may evoke more vivid but narrower experiences. Examining how both sensory diversity and perceptual focus relate to emotional valence and arousal thus offers a deeper understanding of how embodied experience contributes to affective meaning in language.

The present study addresses these gaps by integrating sensory modality ratings and affective ratings into a unified dataset of 298 Chinese monosyllabic adjectives. Building on embodied cognition theory, we pursue three main questions:

1. Interrelationships among sensory modalities – Do certain modalities tend to co-occur in Chinese adjectives, and how do these patterns compare to those reported in other languages?

2. Prediction of affective dimensions – To what extent can valence and arousal be systematically predicted from sensory modality strengths?

3. Role of modality exclusivity – Does perceptual focus on a single dominant modality enhance or diminish affective evaluations?

By focusing on monosyllabic adjectives and examining both perceptual and emotional dimensions simultaneously, this study contributes (a) novel empirical evidence on the embodied structure of Mandarin word meaning, (b) potential cross-linguistic comparisons with English and other languages, and (c) a lexical resource for psycholinguistic, computational, and applied linguistic research.

2 Literature Review

Embodied cognition theory offers a framework in which conceptual knowledge is grounded in sensory, motor, and emotional experiences, challenging the traditional amodal view of semantic representation (Barsalou, 1999, 2008). This perspective has been supported by evidence from cognitive neuroscience, such as the discovery of mirror neurons (Rizzolatti et al., 1996) and mechanisms described by Hebbian learning, which demonstrate how sensory, motor, and linguistic systems can become functionally integrated. In language processing, comprehension and production involve the simulation of sensory and motor experiences associated with linguistic input (Zwaan, 2003; Kogan et al., 2020). For example, understanding the word chair may engage visual imagery of its form, the motor schema of sitting, and even tactile sensations of its surface. Within Barsalou's Perceptual Symbol Systems account (Barsalou, 1999), these perceptual traces—encoded during prior experiences—are reactivated during linguistic tasks, forming an embodied route to meaning. This theoretical framework provides the foundation for investigating how sensory experiences are represented and re-engaged in lexical semantics.

Language often encodes modality-specific features such as vision (bright), audition (loud), touch

(rough), taste (sweet), and smell (fragrant). The concept of modality exclusivity quantifies the degree to which a word is associated with a single sensory modality versus multiple modalities (Lynott and Connell, 2009, 2013). In English, large-scale modality norms have revealed systematic relationships among modalities, such as the strong gustatory–olfactory coupling and the moderate co-occurrence of visual and haptic features (Lynott and Connell, 2009; Lynott et al., 2020; Speed and Brybaert, 2022). These norms have informed research in psycholinguistics, cognitive semantics, and computational modeling by providing quantitative measures of perceptual grounding. They have also been applied to examine processing effects, such as the modality-switch effect in sentence–picture verification tasks, and to explore how cross-modal integration supports conceptual organization.

Compared to English and other Indo-European languages, modality norm research in Mandarin is relatively recent. Chen et al. (2019) produced the first large-scale Mandarin modality exclusivity norms for monosyllabic and disyllabic adjectives, reporting perceptual strength ratings across five basic modalities and analyzing the influence of orthographic semantic radicals on modality judgments. While their dataset provides a crucial baseline, it did not include affective ratings, making it impossible to examine direct links between perceptual profiles and emotional dimensions. Zhong and Ahrens (2023) extended the scope by examining modality–emotion relationships in disyllabic nouns, finding that olfactory and interoceptive modalities were more emotionally charged, especially in arousal and absolute valence. However, their dataset had notable imbalances, with very few tactile and olfactory items (n = 8 for each), and their focus on nouns leaves open questions about whether similar patterns hold for adjectives—an important lexical class for perceptual and affective meaning.

Valence (pleasantness) and arousal (emotional intensity) are two fundamental affective dimensions shaping lexical processing (Bradley and Lang, 1999; Kousta et al., 2011; Warriner et al., 2013). These dimensions influence a range of cognitive processes, including lexical decision, memory, and semantic categorization. In Mandarin, several affective norms have been established (e.g., Yao et al., 2017), but very few studies have integrated them with sensory modality data. Yi et al. (2025) represents a rare attempt, examining sensory–emotion links in a large-scale dataset of disyllabic nouns translated from English norms (Warriner et al., 2013). Their results suggest that certain modalities, particularly olfactory and interoceptive, are more strongly associated with extreme valence and higher arousal.

Findings from English also indicate potential systematic modality–affect correspondences: gustatory and olfactory words tend to be more affectively rich and emotionally flexible (Winter, 2016), while auditory and haptic words are often linked to higher arousal (Lynott et al., 2020). However, other studies argue that these associations may be context-dependent or mediated by conceptual categories (Citron et al., 2014; Lynott and Connell, 2013). The extent to which such correspondences generalize across languages, and across word classes such as adjectives versus nouns, remains an open question.

The relationship between emotion and bodily sensation is not entirely separable: sensory experience serves as an essential, and in some cases sufficient, condition for emotional activation. Empirical evidence supports this link. Dagaev and Terushkina (2014) employed a property verification task to test whether emotional concepts involve embodied somatosensory components. Participants judged whether a given property applied to a concept (e.g., CLOWN–funny), with each trial pair consisting of a context trial and a target trial. A modality switch occurred when the two trials belonged to different modalities (e.g., somatosensory $\rightarrow$ emotional). Their results revealed an asymmetric effect between emotional and bodily sensation channels: when switching from emotional to somatosensory trials, no significant cost was observed; When switching from somatosensory to emotional trials, reaction times were significantly longer. The findings provide direct evidence for the embodied nature of emotional knowledge, indicating that emotional understanding involves the partial re-enactment of bodily states rather than purely symbolic operations.

Nevertheless, it remains unclear how different types of sensory experiences contribute to emotional responses, and whether their influence varies across sensory modalities. Moreover, in the modality-switch paradigm as mentioned above, researchers often select strongly unimodal words to control material consistency, using modality exclusivity as an index of the breadth and focus of perceptual engagement (e.g., Dagaev and Terushk-

ina, 2014; Vermeulen et al., 2007). This paradigm assumes that strongly unimodal words, compared with multimodal ones, more effectively activate specific sensory experiences, thereby producing faster reaction times. From an affective perspective, this raises an important question: Do the dimensionality and breadth of bodily sensory experience differentially influence emotional activation? In other words, the extent to which different words evoke emotions may depend on both their sensory modality (the type of embodied experience) and their modality exclusivity (the degree of perceptual specificity).

Despite increasing interest in the interface between sensory and affective dimensions, Mandarin research remains fragmented in three ways. First, there is no integrated dataset that combines balanced sensory modality ratings and affective ratings for adjectives—a lexical class rich in perceptual and emotional meaning. Second, modality interrelationships, such as gustatory–olfactory coupling or auditory independence, have not been systematically tested in Mandarin adjectives. Third, it remains unclear whether modality exclusivity—perceptual focus on a single dominant modality—predicts emotional valence and arousal in Chinese, and whether such effects align with or diverge from English findings. Addressing these gaps will advance embodied accounts of Chinese lexical semantics and provide a cross-linguistic perspective on how perception and emotion jointly shape word meaning.

3 Methodology

3.1 Participants

A total of 160 native speakers of Mandarin Chinese participated in the study. All participants were recruited in mainland China, held at least a college-level education, and reported normal or corrected-to-normal vision and hearing. The sample was balanced for gender (80 male, 80 female) with a mean age of 24.7 years (SD = 3.2, range = 18–35). None reported a history of neurological or psychiatric disorders. Written informed consent was obtained prior to participation, and the study adhered to ethical guidelines for human subjects research.

3.2 Materials

The stimulus set was constructed from two principal sources to maximize lexical coverage and ensure representativeness across sensory modalities.

1. Existing norms – Sensory modality ratings for 133 monosyllabic Chinese adjectives were obtained from the database developed by Chen et al. (2019), which provides perceptual strength ratings across five modalities: visual, auditory, haptic, gustatory, and olfactory.

2. Newly identified items – To expand the coverage of perceptually salient adjectives, we adopted the selection by Peng et al. (2024), who identified 165 additional monosyllabic adjectives through expert linguistic judgment. These items were specifically chosen for their potential sensory relevance, based on semantic transparency, frequency, and morphological characteristics.

Affective ratings (valence and arousal) for all items were obtained from Peng et al. (2024), ensuring that both sensory and emotional dimensions were available for each word and no scale conversions were necessary.

3.3 Procedure

For the 165 newly identified adjectives, sensory ratings were collected in an online questionnaire format. Participants rated each word on five sensory dimensions (visual, auditory, haptic, gustatory, olfactory) in response to the prompt: "To what extent do you think the word *清* (clear) can be used to describe the following sensory experiences?" Ratings were made on a 6-point Likert scale from 0 ("no association") to 5 ("strong association").

The 165 adjectives were evenly distributed across eight questionnaire lists, each containing 21 or 22 items. Assignment of words to lists was pseudo-randomized to balance modality coverage across lists, and participants were randomly assigned to lists. All instructions were presented in Mandarin, and no time limit was imposed on responses.

3.4 Data Preparation

For each word, mean sensory ratings across participants were calculated for all five modalities. The dominant modality of each word was defined as the modality with the highest mean rating. Modality exclusivity was computed following Lynott and Connell (2009) as the range of the five modality ratings divided by their sum, yielding values from

0 (completely multimodal) to 1 (completely unimodal).

The final dataset comprised 298 adjectives (133 from Chen et al., 2019, ; 165 newly rated), each annotated with five sensory ratings and two affective ratings.

To visualize the overall distribution of ratings, Figure 1. presents histograms for all sensory and affective dimensions, with dashed lines marking the mean (red) and median (black) of each variable. These plots show that visual strength was rated highest on average, whereas gustatory and olfactory strengths were relatively low. Valence and arousal exhibited near-normal distributions, indicating a balanced coverage of positive–negative and low–high arousal adjectives.

3.5 Data Analysis

Data analysis proceeded in three stages. First, descriptive visualizations were generated to display the distribution of ratings across sensory and affective dimensions, enhancing transparency in the data.

Second, to examine the internal structure of sensory modality ratings, Spearman's rank correlation coefficients were computed among the five modalities (visual, auditory, haptic, gustatory, and olfactory). This approach allowed us to identify modality pairs that tend to co-occur in perceptual profiles, as well as those that remain relatively independent (Lynott and Connell, 2009). Significance levels were reported for each pairwise comparison, and the results were visualized using a correlation heatmap with significance markers.

Finally, two multiple linear regression models were constructed to assess how well affective evaluations could be predicted from sensory profiles. Valence and arousal served as the dependent variables, while the five sensory modalities and modality exclusivity were entered as predictors. To facilitate coefficient comparison, predictors were standardized prior to analysis. To ensure the validity and robustness of regression results, several standard diagnostic and correction procedures were employed. Multicollinearity was assessed using Variance Inflation Factors (VIF; Long and Ervin, 2000), which quantify how strongly predictors are linearly related. Heteroscedasticity was tested via the Breusch–Pagan test (Breusch and Pagan, 1979), and residual normality was examined using the Shapiro–Wilk test (Shapiro and Wilk, 1965). Influential observations were identified based on Cook's distance (Cook, 1977), which measures each data point's impact on model estimates. When heteroscedasticity was detected, HC3 heteroskedasticity-consistent standard errors (White, 1980; Long and Ervin, 2000) were applied to correct inferential tests. To further evaluate coefficient stability under potential outliers, robust regression using M-estimators (Huber, 1964) was conducted. These procedures provide a comprehensive evaluation of model reliability and ensure that the reported effects are not driven by violations of linear model assumptions. All analyses were

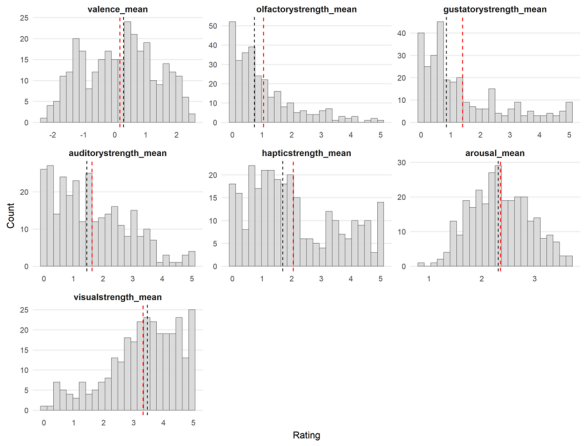


Figure 1: Distributions of sensory and affective ratings.

conducted in R (R Core Team, 2023) using the *stats*, *Hmisc* (Harrell, 2024), *car* (Fox and Weisberg, 2019), *lmtest* (Zeileis and Hothorn, 2002), sandwich (Zeileis, 2004), and MASS (Venables and Ripley, 2002) packages, with the significance level set at $\alpha = .05$.

4 Results

Figure 1 illustrates the distributions of sensory and affective ratings. Visual strength showed the highest overall ratings, gustatory and olfactory strengths were comparatively low, and both valence and arousal exhibited near-normal distributions, indicating balanced affective coverage and clear differentiation across sensory modalities.

4.1 Correlation analysis of sensory modality ratings

The correlation analysis among the five sensory modality ratings revealed both well-documented and novel patterns of interrelationships, offering a detailed picture of the sensory structure underlying Mandarin monosyllabic adjectives. As shown in Figure 2, which presents the full correlation matrix

using Spearman's rank coefficients (ρ), the most prominent positive association was found between gustatory and olfactory strength ($\rho = .81$, $p < .001$). This strong coupling is consistent with robust psycholinguistic and perceptual evidence that taste and smell are closely linked in both neural processing and conceptual representation (Auvray and Spence, 2008; Lynott and Connell, 2009). In practical terms, this means that adjectives highly associated with taste (e.g., 甜 "sweet") are also very likely to be strongly associated with smell, reflecting the multisensory integration that underlies flavor perception. The size of this correlation is comparable to, or even slightly higher than, that reported in English adjective norms (Lynott et al., 2009), suggesting that this gustatory–olfactory pairing may be a cross-linguistic and potentially universal feature of sensory semantics.

A weak but noteworthy positive correlation emerged between visual and haptic strength ($\rho = 0.11$, $p = .057$). Although this effect was marginally significant, its direction aligns with findings in English and other languages where visual attributes (e.g., *光滑* "smooth") often co-occur with tactile impressions. This pattern may reflect the fact that many surface properties—such as smoothness, roughness, or texture—are accessible both visually and through touch. The magnitude of this association in Mandarin appears weaker than in some English datasets (e.g., Speed and Brybaert, 2022), potentially reflecting differences in how adjectives encode surface descriptors in the two languages.

In contrast, visual strength exhibited significant negative correlations with both gustatory strength ($\rho = –0.30$, $p < .001$) and olfactory strength ($\rho = –0.29$, $p < .001$). This suggests a form of mutual exclusivity, where words highly tied to visual imagery tend to have weaker connections to taste and smell. Such negative associations may arise from the fact that visual descriptors in Mandarin often refer to properties of objects that are not inherently linked to flavor or scent (e.g., *亮* "bright," *暗* "dark"), and that taste- or smell-related words tend to evoke internal, bodily-oriented experiences rather than external, visually observable properties.

Auditory strength showed weak to moderate positive correlations with haptic ($\rho = .15$, $p = .011$) and olfactory strength ($\rho = .21$, $p < .001$), and a marginal association with gustatory ($\rho = .11$, $p = .053$). These patterns indicate occasional overlap between sound-related and other sensory imagery, perhaps in words depicting intensity or dynamic qualities (e.g., *响* "loud"). Nevertheless, the overall independence of auditory ratings from most other modalities supports the idea that auditory experience relies on distinct temporal and acoustic representations (Farmer et al., 2006; Lynott and Connell, 2009).

A weaker but statistically significant positive correlation was observed between haptic and olfactory strength ($\rho = 0.14$, $p = .017$). Although the effect size is small, it is intriguing because it suggests occasional overlap between tactile and olfactory imagery. One possible explanation is that some adjectives describe materials or substances (e.g., *腥* "fishy," *滑* "slippery") that are jointly characterized by both texture and smell.

Overall, the results reveal a perceptual organization in which gustatory–olfactory coupling forms a tightly integrated cluster, visual–haptic links remain marginal, auditory is largely distinct, and visual–taste/smell interactions are mutually exclusive. These findings establish a structural baseline for exploring how specific sensory channels contribute to affective meaning in subsequent analyses.

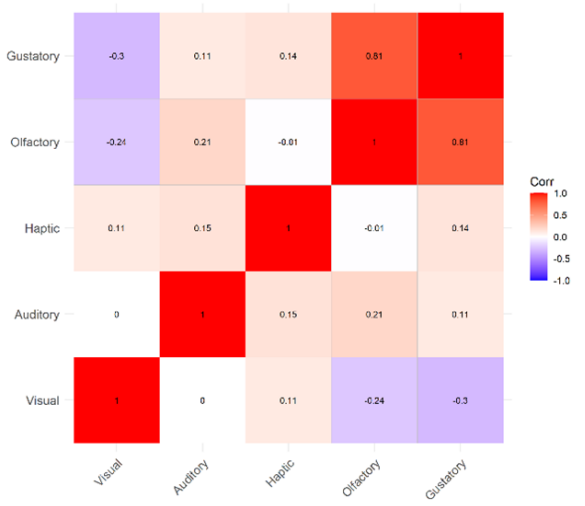


Figure 2: Correlation matrix of sensory modalities.

4.2 Regression analysis of sensory modality ratings and emotional dimensions

To examine how sensory modality profiles are associated with affective meaning, two multiple regression models were fitted, with valence and arousal as outcome variables, respectively. All predictors were standardized to facilitate coefficient comparison.

Comprehensive diagnostic checks were conducted to evaluate model validity. Variance In-

flation Factors (VIFs) were all below 3.3, suggesting no substantial multicollinearity among the predictors. The Breusch–Pagan test indicated heteroscedasticity in the valence model (BP = 37.80, $p < .001$), but not in the arousal model (BP = 4.62, $p = .59$). The Shapiro–Wilk test suggested slight deviations from normality for residuals (valence: W = 0.98, $p < .001$; arousal: W = 0.99, $p = .011$), though Q–Q plots confirmed that the distributions were approximately normal overall.

Inspection of Cook's distance identified 10 potential influence points in the valence model and 19 in the arousal model (Cook's D > 0.0135). Removing these points did not change the pattern or direction of the results; thus, all data points were retained.

To address potential heteroscedasticity and influential observations, we additionally computed HC3 robust standard errors and performed M-estimator regressions (MASS::rlm). The results were consistent across estimation methods, indicating that the observed relationships were statistically reliable and not driven by outliers.

The valence regression model was statistically significant ($F(6, 290) = 2.60$, $p = .018$; adjusted $R^2 = .031$), indicating that sensory modality ratings collectively explained about 3% of the variance in valence evaluations.

After applying HC3 robust standard errors, visual strength remained positively associated with valence ($\beta = 0.20$, $p = .005$), suggesting that adjectives evoking strong visual experiences tend to be perceived as more pleasant. In contrast, modality exclusivity showed a marginally negative association ($\beta = –0.19$, $p = .07$), indicating that adjectives tied to a single sensory channel may be evaluated as slightly less pleasant. Other modalities (auditory, haptic, gustatory, olfactory) did not show reliable associations (all $ps > .10$).

The arousal regression model accounted for a larger portion of variance ($F(6, 290) = 5.80$, $p < .001$; adjusted $R^2 = .089$). Under the HC3 correction, auditory strength ($\beta = 0.18$, $p < .001$), haptic strength ($\beta = 0.11$, $p = .007$), and modality exclusivity ($\beta = 0.18$, $p < .001$) each showed significant positive associations with arousal. This pattern suggests that adjectives evoking auditory and tactile experiences, or dominated by one sensory channel, are more likely to be judged as activating or intense.

The M-estimator regression confirmed the robustness of these findings: auditory ($\beta = 0.18$, $p < .001$), haptic ($\beta = 0.11$, $p = .007$), and modality exclusivity ($\beta = 0.19$, $p < .001$) remained significant.

Taken together, these analyses reveal that valence is modestly associated with visual and multimodal characteristics, whereas arousal is more strongly linked to auditory and haptic experiences, as well as modality dominance. Importantly, all interpretations are correlational in nature and should not be construed as evidence of direct causal influence between sensory and affective dimensions.

5 Discussion

The present study set out to examine how sensory and affective dimensions jointly shape the semantics of Mandarin monosyllabic adjectives, addressing three key questions: (1) the interrelationships among sensory modalities, (2) the extent to which valence and arousal can be predicted from sensory profiles, and (3) the role of modality exclusivity in affective evaluations. By integrating perceptual ratings across five modalities (visual, auditory, haptic, gustatory, olfactory) with affective ratings (valence, arousal) in a unified dataset, we provide a systematic account of how perceptual experience contributes to emotional word meaning in a lexical class that is both perceptually rich and affectively salient.

The correlation analysis revealed a perceptual organization broadly consistent with cross-linguistic findings, yet with certain language-specific nuances. As expected, gustatory and olfactory modalities formed a tightly coupled cluster, reflecting their shared experiential and neural basis in flavor perception (Lynott and Connell, 2009; Lynott et al., 2020). This strong coupling suggests that Mandarin adjectives related to taste often simultaneously evoke olfactory sensations, mirroring the multisensory integration that underlies flavor perception more generally.

A weak but noteworthy overlap was observed between visual and haptic modalities, implying that surface-related features—such as texture, smoothness, or glossiness—may be accessible both visually and through touch, though to a lesser degree than reported in English norms (Speed and Brybaert, 2022). This cross-linguistic difference suggests that Mandarin adjectives may encode surface properties more functionally or contextually rather than visually.

Acknowledgments

We are deeply grateful to Dr. Sudha Arunachalam for her genericity in sharing the video clips used in her studies.

Interestingly, visual imagery tended to exclude internal sensations such as taste and smell, implying a conceptual divide between outwardly observable properties and inward bodily experiences. This asymmetry aligns with the general observation that adjectives describing visual appearance (e.g., *亮* "bright", *暗* "dim") often capture spatially external, object-based qualities, while taste and smell adjectives (e.g., *酸* "sour", *臭* "smelly") pertain to embodied, interoceptive sensations. Meanwhile, auditory modality remained largely independent from the others, reinforcing the idea that auditory representation is grounded in temporally dynamic and acoustically distinct processing. Overall, Mandarin sensory adjectives exhibit a structured but asymmetrical organization—one dominated by taste–smell integration, with vision, touch, and sound occupying more specialized and partially independent roles.

Regression analyses revealed that different sensory channels contribute unequally to emotional meaning. Visual imagery emerged as a modest yet consistent predictor of pleasantness, suggesting that adjectives evoking vivid visual experiences tend to be evaluated as more pleasant. This tendency may reflect cultural and cognitive associations in Mandarin between brightness, clarity, and aesthetic harmony and positive emotional tone, consistent with the symbolic role of light and visual clarity in Chinese idioms, poetry, and moral metaphors.

In contrast, arousal was primarily linked to auditory and haptic modalities, consistent with the idea that sounds and physical sensations carry strong activation potential. This finding highlights how certain sensory experiences—particularly those involving abrupt or proximal stimuli—tend to evoke stronger physiological engagement. Adjectives describing auditory intensity (e.g., *吵* "noisy") or tactile extremity (e.g., *滑* "slippery") appear to embody dynamic or forceful qualities that naturally enhance arousal.

Notably, gustatory and olfactory modalities showed no reliable associations with either valence or arousal, diverging from English results where taste and smell often carry strong emotional valence (Winter, 2016). This discrepancy may be attributable to lexical and cultural factors: Mandarin adjectives describing taste and smell are fewer in number and often restricted to physical or contextual descriptions (e.g., describing food or environment) rather than abstract emotional evaluation, thereby limiting their contribution to affective meaning at the lexical level.

Modality exclusivity further clarified how the breadth of sensory experience shapes affective interpretation. Adjectives that evoke multiple sensory modalities were associated with higher pleasantness, whereas those tied to a single dominant modality were perceived as more activating. This dual pattern indicates that sensory focus intensifies emotional arousal—likely by engaging vivid and detailed mental simulations—but may narrow the hedonic scope, leading to less positive evaluations overall.

This finding complements embodied cognition accounts (Barsalou, 1999, 2008) by showing that the extent of sensory engagement, not merely the presence of sensory content, modulates affective meaning. Words restricted to one sensory channel may evoke more vivid and focused simulations, enhancing arousal; in contrast, multimodal words integrate diverse perceptual cues, producing richer, more balanced experiential representations that align with positive affect.

From an embodied cognition perspective, these findings reinforce that conceptual representation is grounded in perceptual and affective systems, with different modalities contributing unequally to distinct emotional dimensions. Vision appears more closely tied to hedonic evaluation (valence), while audition and touch contribute to physiological activation (arousal). This division may reflect underlying neurocognitive specializations: the visual system is linked to appraisal and aesthetic judgment, whereas auditory and tactile systems are connected to proximity, urgency, and survival-relevant reactions.

Furthermore, the role of modality exclusivity underscores the importance of sensory diversity and focus in shaping emotional meaning. Words evoking multiple modalities may reflect more integrative and experiential processing, while modality-specific words capture sharper but narrower affective tones. Taken together, these findings suggest that sensory profiles—rather than isolated modalities—form the foundation for affective meaning in lexical semantics.

6 Conclusion

The present study examined how sensory modalities and affective dimensions jointly shape the semantics of Mandarin monosyllabic adjectives. Integrating newly collected sensory ratings with affective evaluations, we constructed a dataset of 298 adjectives for the first systematic analysis of sensory–affective mappings in this lexical class. The results revealed an asymmetrical sensory organization: gustatory and olfactory modalities were strongly coupled, visual–haptic overlap was weak, and auditory imagery remained largely independent. Regression analyses showed that visual strength predicted more positive valence, whereas auditory and haptic strength predicted higher arousal. Modality exclusivity further influenced affective meaning—multimodal adjectives were more pleasant, while unimodal ones were more arousing. These findings support embodied cognition accounts, showing that sensory experience systematically contributes to affective semantics, with both universal and language-specific patterns in Mandarin.

7 Limitations and future directions

While the present study provides a comprehensive dataset and robust statistical analyses, several limitations should be noted. First, although the dataset covers a large number of monosyllabic adjectives, it does not include disyllabic adjectives or other word classes that may exhibit different modality–affect patterns. Future work could broaden the scope to these forms and compare results across lexical categories. Second, the study relies on explicit ratings, which capture conscious associations but may not fully reflect automatic or context-dependent processing. Incorporating time-sensitive methods such as EEG, eye-tracking, or priming paradigms could reveal how modality–affect mappings unfold during real-time comprehension. Third, cultural and contextual factors were not directly examined; cross-linguistic comparisons, for instance between Mandarin and English, could clarify universal versus language-specific effects.

Beyond theoretical contributions, the findings have implications for language education, computational linguistics, and affective computing. Understanding which sensory modalities align with positive or high-arousal meanings could inform vocabulary teaching and sentiment modeling, while the dataset may aid multimodal systems in interpreting and generating emotionally nuanced language.In sum, this study shows that sensory and affective dimensions are systematically linked in Mandarin adjectives, reflecting both universal and language-specific patterns. The integrated sensory–affective norms offer a theoretical contribution to embodied cognition and a practical resource for future psycholinguistic and cross-linguistic research.

References

Malika Auvray and Charles Spence. 2008. The multisensory perception of flavor. *Consciousness and Cognition*, 17(3):1016–1031.

Lawrence W. Barsalou. 1999. Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22(4):637–660.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

T. S. Breusch and A. R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

I-H Chen, Q Zhao, Y Long, Q Lu, and C-R Huang. 2019. Mandarin Chinese modality exclusivity norms. *PLOS ONE*, 14(2):e0211336.

Francesca M.M. Citron, Brendan S. Weekes, and Evelyn C. Ferstl. 2014. Arousal and emotional valence interact in written word recognition. *Language, Cognition and Neuroscience*, 29(10):1257–1267.

R. Dennis Cook. 1977. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Nikolay I. Dagaev and Yulia I. Terushkina. 2014. Conceptual knowledge of emotions includes somatosensory component: Evidence from modality-switch cost effect. *Journal of Cognitive Psychology*, 26(3):322–332.

A. R. Damasio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons, New York, NY.

Thomas A. Farmer, Morten H. Christiansen, and Padraic Monaghan. 2006. Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103(32):12203–12208.

John Fox and Sanford Weisberg. 2019. *An R Companion to Applied Regression*, third edition. Sage, Thousand Oaks, CA.

Curtis Hardin. 1993. The influence of language on thought. *Social Cognition*, 11(3):277–308.

Frank E. Harrell, Jr. 2024. *Hmisc: Harrell Miscellaneous*. R package version 5.1-2.

Peter J Huber. 1964. Robust estimation of a location parameter: Annals mathematics statistics, 35. *Ji, S., Xue, Y. and Carin, L.(2008),'Bayesian compressive sensing', IEEE Transactions on signal processing*, 56(6):2346–2356.

Boris Kogan, Enrique García-Marco, Agustina Birba, Camila Cortés, Margherita Melloni, Agustín Ibáñez, and Adolfo M. García. 2020. How words ripple through bilingual hands: Motor-language coupling during l1 and l2 writing. *Neuropsychologia*, 146:107563.

S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. Del Campo. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

J. Scott Long and Laurie H. Ervin. 2000. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Dermot Lynott and Louise Connell. 2009. Modality exclusivity norms for 423 object properties. *Behavior research methods*, 41(2):558–564.

Dermot Lynott and Louise Connell. 2013. Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior research methods*, 45(2):516–526.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior research methods*, 52(3):1271–1291.

Cheng Peng, Xu Xu, and Zhen Bao. 2024. Sentiment annotations for 3827 simplified chinese characters. *Behavior Research Methods*, 56(2):651–666.

Jesse J. Prinz. 2005. Passionate thoughts. In Rolf Zwaan and Diane Pecher, editors, *The Grounding of Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, pages 93–114. Cambridge University Press, Cambridge, MA.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. 1996. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2):131–141. Mental representations of motor acts.

Eleanor Rosch. 1974. Linguistic relativity. In Albert Silverstein, editor, *Human communication: Theoretical perspectives*, pages 95–121. Lawrence Erlbaum Associates, Hillsdale, NJ.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Laura J Speed and Marc Brybaert. 2022. Dutch sensory modality norms. *Behavior research methods*, 54(3):1306–1318.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*, fourth edition. Springer, New York. ISBN 0-387-95457-0.

Nicolas Vermeulen, Paula M Niedenthal, and Olivier Luminet. 2007. Switching between sensory and affective systems incurs processing costs. *Cognitive Science*, 31(1):183–192.

L. S. Vygotsky. 2000. *Thought and Language*. MIT Press, Cambridge, MA.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Halbert White. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Benjamin Lee Whorf. 1956. *Language, thought, and reality*. MIT Press, Cambridge, MA.

Bodo Winter. 2016. Taste and smell words form an affectively loaded and emotionally flexible part of the english lexicon. *Language, Cognition and Neuroscience*, 31(8):975–988.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 chinese words. *Behavior research methods*, 49(4):1374–1385.

Wei Yi, Haitao Xu, and Kaiwen Man. 2025. Perception of emotion across cultures: Norms of valence, arousal, and sensory experience for 4923 chinese words translated from english in warriner et al.(2013). *Behavior Research Methods*, 57(1):43.

Achim Zeileis. 2004. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17.

Achim Zeileis and Torsten Hothorn. 2002. Diagnostic checking in regression relationships. *R News*, 2(3):7–10.

Yin Zhong and Kathleen Ahrens. 2023. The emotion code in sensory modalities. In *Chinese Lexical Semantics*, pages 183–192, Cham. Springer Nature Switzerland.

Rolf A. Zwaan. 2003. The immersed experiencer: Toward an embodied theory of language comprehension. volume 44 of *Psychology of Learning and Motivation*, pages 35–62. Academic Press.

Rolf A Zwaan, Robert A Stanfield, and Richard H Yaxley. 2002. Language comprehenders mentally represent the shapes of objects. *Psychological science*, 13(2):168–171.

A Supplementary Material

The data for the experiment is available at https://osf.io/yzuhr/

KWordinaryVQA: A Keyword-Driven Generative Visual Question Answering System for Culinary Exploration

Huy Trieu[1,2], Thanh Thai Nguyen[1,2], Thanh Nghia Vo[1,2], Thinh Vuong Vo[1,2], Thanh Tu Dang[1,2], Tung Le[1,2,*]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam

[2]Vietnam National University, Ho Chi Minh city, Vietnam

{tghuy22,ntthai22,vtnghia22,vtvuong22,dttu22}@clc.fitus.edu.vn

[*]Corresponding author: lttung@fit.hcmus.edu.vn

Abstract

Visual Question Answering (VQA) has seen significant progress on general images, yet food imagery presents unique challenges requiring domain-specific understanding. This paper presents KWordinaryVQA[1], an end-to-end automated pipeline to construct large-scale food VQA datasets. Starting from raw images, we employ advanced Large Language Models (LLMs) to generate detailed descriptions and synthesize diverse question-answer pairs, followed by targeted manual validation to ensure high-quality evaluation data. We then benchmark multiple approaches–including LLMs under zero-shot and few-shot settings, a traditional retrieval baseline, and representative fine-tuned vision-language models–evaluating them on accuracy, human judgment, and inference efficiency. Our workflow mirrors a standard data science process of data collection, exploration, evaluation, and model building, providing a systematic framework for domain-specific VQA.

1 Introduction

Visual Question Answering (VQA) aims to develop systems that can answer natural language questions based on the visual content of an image. This inherently multimodal task requires a sophisticated integration of computer vision, natural language understanding, and often, commonsense reasoning. While significant strides have been made with general-purpose VQA benchmarks like VQA v2.0 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019), largely fueled by advances in transformer-based multimodal architectures, these models often struggle when applied to specialized domains. The food domain, in particular, presents unique challenges and opportunities, yet remains relatively underexplored despite its profound practical relevance.

Addressing VQA in the food domain extends beyond academic inquiry, unlocking numerous real-world applications. Such systems could revolutionize dietary tracking by identifying ingredients and portion sizes, enhance cooking education with interactive guidance, and offer crucial assistive technologies for individuals with visual impairments or dietary restrictions. For example, the ability to accurately detect allergens or forbidden food items is crucial for users with health concerns, such as food allergies and diabetes, or those following religious dietary laws, underscoring the need for ingredient-level understanding. Furthermore, visual cues like color and texture are vital indicators of food condition–whether an item is raw, perfectly cooked, or fresh–making an effective food VQA system invaluable for culinary evaluation, food safety, and quality control in both domestic and industrial settings such as restaurants and food processing plants. Spatial reasoning also plays a crucial role; understanding the arrangement of food components on a plate can inform analyses of food presentation, a key factor in professional culinary arts and brand consistency across F&B chains. AI systems could thereby assess presentation uniformity, portion control, and adherence to plating standards, directly impacting customer experience and brand perception.

Despite these compelling applications and the increasing capabilities of general VQA models, our preliminary analysis and a review of existing literature indicate that current state-of-the-art approaches, including large language models (LLMs) with visual grounding, classification-based methods, and statistical models, often falter on domain-specific food-related queries (as detailed in Section 2). This performance gap stems from several factors: the inherent visual complexity of food items (e.g., fine-grained differences between ingredients, varied cooking states), the need for specialized, implicit domain knowledge (e.g., culinary

[1]The final dataset is available on Kaggle.

techniques, cultural nuances), and critically, the scarcity of large-scale, high-quality, open-ended VQA datasets tailored specifically for the food domain.

To address the lack of domain-adapted benchmarks in visual question answering, we introduce KWordinaryVQA, an end-to-end automated pipeline for constructing VQA datasets from food imagery. Beginning with raw images from the public culinary platform Allrecipes (Allrecipes contributors, 2025), the pipeline first employs Gemini 2.0 Flash (Google DeepMind, 2025) to perform image captioning and then extract key phrases from these captions. Subsequently, question-answer pairs are generated from the captions and key phrases using DeepSeek-V3 (Liu et al., 2024). This automated workflow enables the construction of scalable datasets with minimal human effort, promoting deeper reasoning through open-ended formats rather than restrictive multiple-choice or classification tasks.

Our contributions are fourfold. First, we design and implement a fully automated pipeline for generating domain-specific VQA data from food imagery. Second, we construct the KWordinaryVQA dataset, consisting of 43,455 QA pairs across 8,693 images, including manually validated test and validation splits. Third, we benchmark a diverse set of VQA approaches–ranging from zero-shot LLMs and TF-IDF retrieval to fine-tuned vision-language transformers. Finally, we conduct in-depth dataset analysis, covering question types, linguistic features, and performance breakdowns, offering insight into domain-specific VQA challenges.

To facilitate further research, we publicly release the KWordinaryVQA dataset, generation pipeline, and evaluation code, aiming to advance domain-adapted VQA in food and other specialized domains.

2 Related Works

The growing interest in AI for the food domain has led to the development of several datasets for food-related tasks. These datasets, while valuable, can be broadly categorized by their focus on either cultural depth or task-specific evaluation, each with inherent limitations.

Several benchmarks offer deep insights into specific culinary traditions. For instance, FoodieQA (Li et al., 2024) provides a manually annotated, multimodal benchmark for Chinese cuisine, while IndiFoodVQA (Agarwal et al., 2024) uses a knowledge-graph-enhanced pipeline to assess reasoning in the Indian food domain. Although these datasets are rich in specialized knowledge, their cultural specificity and, in the case of FoodieQA, reliance on manual annotation, can limit their scalability and general applicability.

In parallel, other large-scale efforts often concentrate on more constrained task formulations. WorldCuisines (Winata et al., 2024), despite its impressive scale and multilingual support, primarily targets dish and origin identification rather than complex reasoning about visual attributes. Similarly, Food-VQA-Benchmark (Cheng et al., 2024) evaluates a suite of tasks but largely relies on closed-set formats or structured outputs, which may not fully capture the complexities of truly open-ended VQA where models must generate free-form answers.

Collectively, while these datasets have advanced the field, a clear gap persists. They are often constrained by the laborious nature of manual annotation, which impacts scale and diversity, or their task formulations favor identification and structured information over fostering a broad spectrum of open-ended inquiries that demand nuanced visual reasoning. This highlights a pressing need for a large-scale, open-ended VQA dataset for general food imagery, developed through a scalable and adaptable pipeline–a need that our work, KWordinaryVQA, directly aims to address.

3 Dataset Acquisition and Preprocessing

To construct the KWordinaryVQA dataset, we initially crawled 49,332 structured data entries from the public recipe source Allrecipes (Allrecipes contributors, 2025), covering a broad diversity of cuisines, food types, and presentations, encompassing both professionally staged studio shots and user-submitted home-cooking photographs. Each entry included a food image accompanied by metadata such as food names, descriptions, and ingredients. Given the scale of this raw dataset and computational constraints, we designed a multi-step preprocessing pipeline to curate a high-quality, representative, and manageable subset suitable for robust model training and evaluation.

3.1 Initial Data Filtering

The initial preprocessing involved two main stages. First, to normalize food names, a fuzzy string matching technique using the RapidFuzz (Bach-

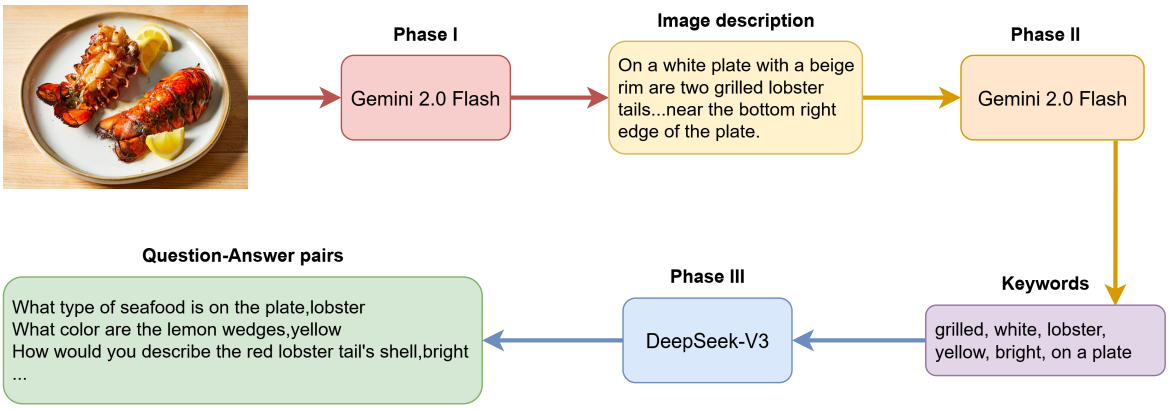


Figure 1: The dataset creation pipeline consists of three stages: (I) generating image descriptions, (II) extracting relevant keywords, and (III) creating question-answer pairs.

mann, 2024) library was applied with a token set ratio scorer and an 80% similarity threshold. Within each group of near-duplicate names, the shortest variant was selected as the canonical form to improve naming consistency. This step resulted in 27,270 entries. Subsequently, we removed all samples with incomplete metadata to ensure data integrity for downstream tasks. Specifically, 498 entries lacked calorie data and 9 were missing image dimensions. By retaining only complete entries, we obtained a final dataset of 21,370 samples.

3.2 Outlier Analysis and Retention

Outliers were analyzed using the interquartile range (IQR) method, identifying caloric values outside an 800-calorie range as potential outliers. In the context of KWordinaryVQA, high-calorie dishes (e.g., energy-rich foods) were deemed valuable for calorie-related questions. Thus, we retained these outliers to preserve informative samples, ensuring the dataset remained representative of diverse nutritional profiles.

3.3 Undersampling for Class Balance

The dataset exhibited significant class imbalances. To address this, we implemented a two-tiered undersampling strategy. First, to mitigate the dominance of common dishes (e.g., chicken, cake), we capped their sample count at 100-150 per dish. Conversely, extremely rare dishes (1-4 samples) were removed, as their low representation was insufficient for effective model learning. This process reduced the dataset to 16,817 entries.

Second, to address the severe imbalance between vegan and non-vegan dishes, we capped the number of samples for non-vegan dishes with more than 20 instances at 20, while all vegan and rare non-vegan dishes were kept unchanged. This strategy yielded a more reasonable class ratio (approximately 1:4) while preserving dataset diversity, resulting in a raw set of 9,191 unique food images.

3.4 Scope-based Filtering

Finally, a manual review was conducted to ensure linguistic and cultural consistency. To create a robust benchmark, we narrowed the dataset's scope to focus on food items commonly understood within general English culinary discourse. Entries for dishes requiring specific, non-English cultural context for identification (e.g., "Cao Lau," a Vietnamese specialty) were excluded. This deliberate choice, while limiting cultural breadth, was crucial for enhancing the dataset's internal consistency and suitability for our defined VQA task. This step resulted in the final pool of 8,693 images used for generation.

4 Dataset Creation

Following the preprocessing pipeline, a final collection of 8,693 unique food images served as the visual foundation for the KWordinaryVQA dataset. This curated set captures a wide variety of cuisines, food types, and presentation styles. The subsequent dataset creation process involved three main automated stages: description generation, keyword extraction, and question-answer synthesis.

4.1 Description Generation

For each image, we first generated a detailed textual description using Gemini 2.0 Flash (Google DeepMind, 2025). Given a food image, the model was prompted (see Appendix A.1) to produce a concise, standalone paragraph of up to 200 words, describ-

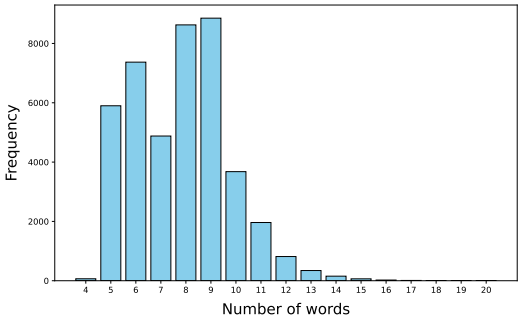


Figure 2: Distribution of question lengths.

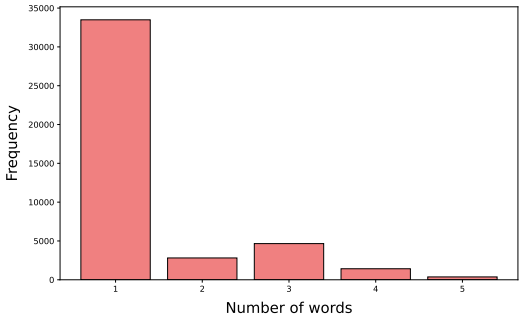


Figure 3: Distribution of answer lengths.

ing its salient contents, including ingredients, dish type, and serving context. For example, for an image of ramen, the model might output: "A bowl of noodle soup with sliced pork, half an egg, green onions, and seaweed on top, on a wooden table." A detailed description is foundational, as it provides the rich, factual grounding required for generating diverse and meaningful questions. While the process was largely automated, minor manual corrections were occasionally applied to address clear factual inaccuracies (e.g., misidentifying an ingredient).

4.2 Keyword Extraction

To guide the question synthesis process and establish ground-truth answers, we then extracted key terms from each image description. This step, again utilizing Gemini 2.0 Flash, was designed to produce a set of target answers for the subsequent QA generation. A specific prompt (see Appendix A.2) instructed the model to identify six diverse keywords–including at least one verb, adjective, noun, color, and preposition–to ensure a variety of question types. For instance, from the ramen description above, the extracted keywords, which would later serve as target answers, might be: *slice, fresh, pork, green, on the table, soup*.

4.3 Question-Answer Synthesis

Using the generated descriptions and extracted keywords, we synthesized QA pairs with DeepSeek-V3 (Liu et al., 2024). For each image, the model received its description and keywords, tasked with generating one question per keyword. The prompt (Appendix A.3) enforced several critical constraints: questions had to be grounded in the description, the answer for each question had to be the corresponding keyword, and "Where" questions were specifically generated for prepositional keywords. This automated workflow allowed for the rapid, large-scale synthesis of question-answer pairs, averaging approximately five pairs per image.

4.4 Data Splitting and Validation

The dataset was partitioned at the image level to prevent content leakage between splits: 90% of the images and their corresponding QA pairs were allocated to the training set (39,051 pairs), and the remaining 10% were reserved for the test set (4,342 pairs). A subset of the training data was subsequently held out for validation.

To ensure the high quality of our evaluation benchmark, the initial test set underwent a rigorous manual validation process for factual accuracy, visual relevance, and clarity. This involved removing or rephrasing samples that: (1) referenced details not visually apparent (i.e., hallucinations); (2) were vague or inadequately specified; or (3) required subjective inference or external knowledge. For example, speculative queries about non-visible attributes (e.g., spiciness) were discarded. This meticulous process resulted in a high-confidence test set of 3,699 QA pairs (a 15.0% reduction). The training and validation sets were not subjected to this manual filtering. Illustrative examples from the resulting dataset can be found in Appendix D.

5 Dataset Analysis

To characterize the newly constructed KWordinaryVQA dataset and assess the output of our generation pipeline, we conducted an exploratory data analysis. This analysis provided insights into properties such as question diversity and potential biases, informing our understanding of the dataset's characteristics and suitability for evaluating VQA models.

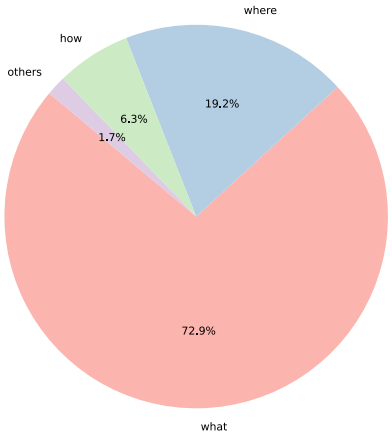


Figure 4: Question types distribution.

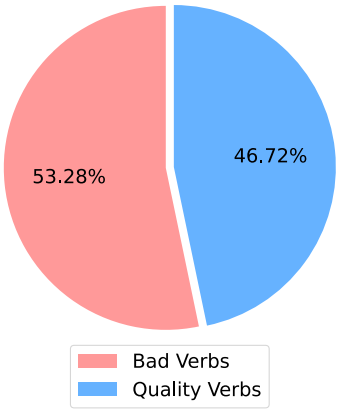


Figure 5: Analysis of verb quality in test set answers. The chart illustrates the proportion of quality verbs among all verb-based answers. Of the answers identified as verbs, 46.72% were deemed to be meaningful (quality verbs).

5.1 Length Distributions

We first analyzed the length of questions and answers in terms of word count. Questions in KWordinaryVQA tend to be concise, with a mean length of 7.79 and a median of 8 words. As shown in Figure 2, the distribution peaks around 8-9 words, with a long tail extending to about 15 words, corresponding to more complex inquiries. Answers are typically even shorter; over 80% are single or two-word responses. The average answer length is 1.41 words, with a heavy concentration on one-word answers (Figure 3). This finding has direct implications for evaluation design, suggesting that metrics should accommodate brief answers and that exact-match criteria may be overly stringent for some phrasal responses.

5.2 Question Types

We categorized the questions by their starting words to identify common patterns. As illustrated in Figure 4, the dataset is overwhelmingly dominated by “What” questions, which typically inquire about ingredients, objects, or dish names. “Where” questions focusing on spatial relations are the next most frequent category, followed by “How” questions, which encompass both counting and descriptive inquiries.

Notably, other question types such as polar and causal “Why” questions are extremely rare. This distribution highlights a potential bias in our automated generation pipeline, which favors descriptive inquiries grounded directly in visual evidence over more abstract or inferential reasoning.

5.3 Quality Estimation

While the training sets were not manually examined, we aimed to quantitatively estimate the incidence of semantically poor labels (i.e., label noise). To this end, we conducted a comparative analysis of verb quality[2] between the unfiltered test set and its manually validated counterpart. We argue that this analysis is representative of the entire dataset, given that all splits were generated from the same pipeline and exhibit consistent linguistic distributions, as detailed in Appendix E and F.

Our methodology leveraged the pre-existing filtered test set as ground truth. First, we employed a BERT-uncased part-of-speech tagging model (Blagojevic, 2024) to programmatically identify all verb-based answers in the initial, unfiltered test set, resulting in a total of 137 such answers. We then applied the same process to the manually validated test set, which yielded only 64 verb-based answers.

By this definition, the 64 verbs that survived the manual validation process were considered "high-quality" (e.g., "eat", "contain"), while the 73 verbs that were filtered out were considered "low-quality" or generic (e.g., “do”, “be”). As illustrated in Figure 5, this comparison yields a verb quality rate of 46.72%. However, given that verb-based answers constitute a small fraction of the dataset, accounting for just 3.2% of all questions in the unfiltered test set. Therefore, while this analysis highlights

[2]For this analysis, a “verb” is defined as a word tagged as VERB, from which we excluded words ending in “-ed” to filter out potential passive voice forms.

a qualitative weakness in our pipeline, its overall quantitative impact on the integrity of the automatically generated training and validation sets is likely limited.

5.4 Data Summary

Statistic	Number
Size of dataset	42,750
Unique questions	29,822
Unique answers	4,413
Number of images	8,693
Average question length	7.79
Average answer length	1.42

Table 1: Dataset statistics.

Table 1 summarizes the key properties of the KWordinaryVQA dataset. Crucially, our analysis confirms that key linguistic characteristics are consistently maintained across the training, validation, and test splits (see Appendix E and F). This consistency ensures that our benchmark provides a fair and representative basis for evaluating model performance.

6 Experimental Setup

6.1 Evaluation Metrics

To provide a holistic assessment of model performance on the KWordinaryVQA test set, we employed a comprehensive suite of metrics targeting various dimensions of answer quality. We began by measuring Accuracy, defined as the proportion of predictions that exactly matched the ground-truth answers. This was followed by the computation of standard token-level Precision, Recall, and F1-score (Goutte and Gaussier, 2005) using normalized text. For lexical overlap, we used BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004). Lastly, to evaluate semantic relevance beyond surface-level matching, we computed GPTScore (Fu et al., 2024) using the GPT-4o model via the OpenAI API.

6.2 Traditional Baselines

We first established performance using two traditional methods that treat VQA as a retrieval or classification task over a fixed answer set.

6.2.1 Lexical Retrieval

As a non-parametric benchmark, we adopt a TF-IDF retrieval method (Sparck Jones, 1972) to capture lexical patterns independent of visual input. In this setup, each question is represented as a TF-IDF weighted bag-of-words vector after stopword removal and stemming. For any given test question, the answer of the most lexically similar training question–determined by cosine similarity–is adopted as the prediction. This approach, while computationally efficient, entirely disregards visual input.

6.2.2 Fine-Tuned Classifiers

To evaluate supervised, classification-based VQA, we fine-tuned two models representing distinct architectural paradigms: BEiT-3 (Wang et al., 2023), a state-of-the-art unified model with an end-to-end fused architecture, and LXMERT (Tan and Bansal, 2019), a canonical two-stream model that operates on pre-extracted visual region features.

Both models were fine-tuned on the KWordinaryVQA training set with a classification objective, using a cross-entropy loss over a predefined answer set. To prepare the data for this task, answer labels in the training and validation sets underwent a consistent normalization process, including lowercasing and punctuation removal. Reflecting its two-stream design, LXMERT also required a separate feature extraction step, for which we employed a pretrained Faster R-CNN model (Ren et al., 2015) with a ResNet-50 backbone (He et al., 2016) and Feature Pyramid Network (Lin et al., 2017). Further details on the fine-tuning hyperparameters are provided in Appendix C.

6.3 Generative Model Baselines

Next, we evaluated the performance of several state-of-the-art multimodal Large Language Models (LLMs) under different prompting conditions.

6.3.1 Zero-Shot Evaluation

In the primary evaluation setting, four models were tested under strict zero-shot setting: Llama 3.2-Vision 11B Instruct (Meta AI, 2024), MiniCPM-o 2.6 (Yao et al., 2024), Qwen2.5-VL 7B Instruct (Bai et al., 2025), and [3]Gemini 2.0 Flash (Google DeepMind, 2025). Each model received only the test image (resized to 480×480 pixels) and a question, without any in-context examples.

Initial tests with both 'Raw' (unconstrained) and 'Instructed' (concise format) prompting yielded low

[3]Questions were synthesized by DeepSeek-V3, mitigating potential self-enhancement bias in Gemini's evaluation.

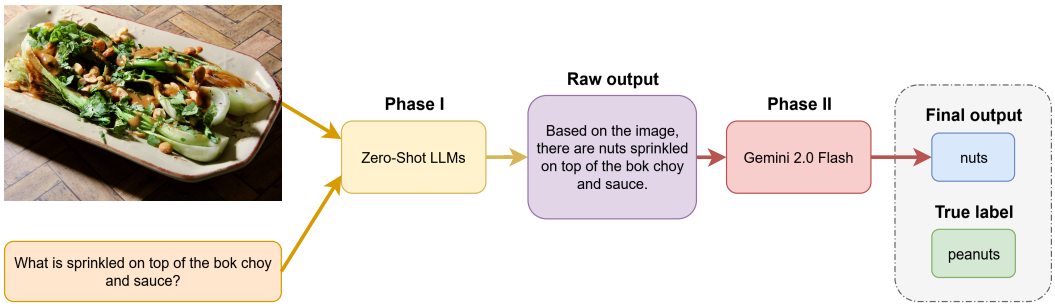


Figure 6: The two-phase post-processing pipeline applied to our Zero-Shot VLM baselines. (I) A LLM initially generates a raw, often verbose, answer (e.g., 'The top of the enchiladas is golden brown.') to the visual question. (II) This raw output is then post-processed by Gemini 2.0 Flash to extract a concise final answer (e.g., 'golden brown'), aligning it with the desired concise format.

Model	Accuracy	Precision	Recall	F1-score	BLEU	ROUGE-L	GPTScore
TF-IDF	0.2163	0.2756	0.2762	0.2728	0.2209	0.2981	0.4285
BEiT-3 (fine-tuned)	**0.4804**	0.5543	**0.5436**	**0.5455**	**0.4806**	**0.5780**	**0.6864**
LXMERT (fine-tuned)	0.3574	**0.6203**	0.3574	0.3017	0.3549	0.4457	0.5686

Table 2: Evaluation of classification-based VQA models. Fine-tuned BEiT-3 achieved the best overall performance among non-generative models, while LXMERT showed high precision but lower recall. The TF-IDF baseline, though simple, performed reasonably well in repetitive query scenarios.

scores due to the models' tendency to generate verbose answers. Consequently, we implemented a crucial post-processing step, where an LLM was used to extract a concise, relevant answer from the raw output (see Appendix B). This refinement was applied to all zero-shot results to ensure fair comparison.

6.3.2 Few-Shot Evaluation

To further assess in-context learning capabilities, we conducted few-shot evaluations on two of the models: Llama 3.2-Vision 11B Instruct and MiniCPM-o 2.6. The evaluation was performed under two distinct conditions, utilizing both 5 and 10 demonstration exemplars. These exemplars, carefully curated from the training set, consisted of image-question-answer triplets. This contextual prefix was prepended to each test instance to prime the models towards the appropriate response format. Unlike the zero-shot setting, no post-processing was applied to the few-shot outputs.

7 Results and Analysis

The empirical results of our benchmarking experiments are presented below. These evaluations use a range of current models to probe the difficulty and characteristics of the KWordinaryVQA dataset.

7.1 Traditional and Fine-Tuned Baselines

The performance of the fine-tuned models and the lexical retrieval baseline is presented in Table 2. A substantial performance gap was observed between the fine-tuned models and the TF-IDF retrieval baseline. BEiT-3, the strongest performer in this category, achieved a modest accuracy of 0.4804. This result indicates that the dataset is not easily solved even by powerful, domain-adapted vision-language models, suggesting that successful task completion requires a level of reasoning beyond simple pattern recognition.

7.2 Generative Model Baselines in Zero-Shot Settings

The performance of large generative models in a zero-shot setting further highlights the challenge of KWordinaryVQA (Table 3). Without intervention, all evaluated LLMs performed poorly, primarily due to a systemic failure to produce answers in the required concise format. This outcome demonstrates that the benchmark effectively tests a model's ability to adhere to precise output constraints in addition to its content understanding. While a post-processing pipeline substantially improved the scores–with Gemini 2.0 Flash achieving the highest GPTScore–the necessity of this external step underscores a fundamental difficulty that the dataset exposes in current generative models.

Model	Version	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	GPTScore
LLaMA 3.2 Vision	Raw	0.0224	0.1045	**0.6439**	0.159	0.0419	0.1591	0.3476
	Instructed	0.2160	0.3489	0.5723	0.3923	0.2437	0.4193	0.6245
	Post-Processed	**0.2958**	**0.4355**	0.5346	**0.4620**	**0.3151**	**0.5066**	**0.6782**
MiniCPM-o 2.6	Raw	*negligible*	0.0699	**0.6665**	0.1160	0.0172	0.1148	0.3787
	Instructed	0.2690	0.4143	0.5533	0.4502	0.2978	0.4908	0.6599
	Post-Processed	**0.3160**	**0.4568**	0.5287	**0.4724**	**0.3275**	**0.5228**	**0.6936**
Qwen2.5-VL	Raw	*negligible*	0.0539	**0.6494**	0.0953	0.0119	0.0910	0.3503
	Instructed	0.0657	0.2443	0.5890	0.3163	0.1062	0.3442	0.6599
	Post-Processed	**0.2609**	**0.4111**	0.5083	**0.4356**	**0.2756**	**0.4832**	**0.6668**
Gemini 2.0 Flash	Raw	0.0370	0.1476	**0.6633**	0.2150	0.0100	0.2490	0.5607
	Instructed	0.1703	0.3153	0.3690	0.3239	0.1815	0.3710	0.6113
	Post-Processed	**0.3533**	**0.4974**	0.5694	**0.5130**	**0.3627**	**0.5597**	**0.7239**

Table 3: Zero-shot performance of generative vision-language models on the KWordinaryVQA benchmark under three evaluation settings: **Raw** denotes direct model output from image-question input without prompt engineering; **Instructed** augments input with manually designed prompts to guide answer generation; and **Post-Processed** applies a secondary model to refine raw outputs for improved alignment with reference answers. Bold values indicate the best scores within each model.

Model	Version	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	GPTScore
LLaMA 3.2 Vision	5 samples	0.3060	0.4196	0.4866	0.4360	0.3198	0.4755	0.6108
	10 samples	**0.3466**	**0.4559**	**0.5226**	**0.4694**	**0.3568**	**0.5096**	**0.6541**
MiniCPM-o 2.6	5 samples	0.2374	0.3971	0.5472	0.4363	0.2613	0.4810	0.6528
	10 samples	**0.3225**	**0.4546**	**0.5394**	**0.4761**	**0.3400**	**0.5178**	**0.6702**

Table 4: Performance of generative models with few-shot setting. Both LLaMA 3.2 Vision and MiniCPM-o 2.6 struggled to match the concise answer format of the dataset, yielding moderate scores across all metrics. This highlights the importance of output refinement for domain-specific VQA.

7.3 Few-Shot vs. Post-Processed Zero-Shot

Our final set of experiments confirmed the dataset's robustness against common learning strategies. As shown in Table 4, the effectiveness of providing in-context examples was inconsistent. While a 10-shot configuration surpassed the F1-score of the post-processed zero-shot baseline, a 5-shot configuration proved insufficient to achieve the same, indicating that adapting to the dataset's diversity necessitates a substantial number of exemplars. Crucially, a persistent trade-off was observed across both settings: the few-shot approach improved exact-match accuracy but resulted in lower semantic relevance (GPTScore) compared to the post-processed counterpart. This finding demonstrates that the core challenges of KWordinaryVQA, particularly the dual demand for fine-grained reasoning and strict output formatting, are not easily circumvented by simple prompting strategies.

7.4 Inference Cost

Our evaluation reveals critical trade-offs between performance, cost, and computational requirements, as detailed in Table 7. Fine-tuned models, particularly BEiT-3, offer the highest efficiency, delivering moderate accuracy with minimal inference time and no financial cost.

In contrast, large generative models present a more complex cost-benefit profile. Notably, a zero-shot approach combined with our optional post-processing step achieves superior performance to 10-shot prompting but at a fraction of the computational cost and time. This result suggests that for tasks like KWordinaryVQA, refining the output of a cost-effective zero-shot model can be a more pragmatic and effective strategy than computationally expensive few-shot prompting.

7.5 Overall Evaluation

Our collective results reveal that no single modeling paradigm excels across all dimensions of performance, efficiency, and cost on the KWordinaryVQA benchmark. Instead, the findings highlight a series of critical trade-offs that present a nuanced decision for practical applications.

Fine-tuned models, exemplified by BEiT-3,

achieve the highest exact-match accuracy and inference efficiency, but require a significant upfront investment in training. Conversely, large generative models offer flexibility and eliminate training costs, with a post-processed zero-shot approach–using Gemini 2.0 Flash–delivering the best semantic relevance. However, this approach's reliance on an external refinement step and the general failure of expensive few-shot prompting underscore the inherent challenges these models face with the dataset's specific constraints.

Ultimately, these complex trade-offs solidify KWordinaryVQA's value as a multifaceted benchmark. It effectively probes models on distinct capabilities–from precise classification to semantic understanding and adherence to formatting–demonstrating that a truly robust system for food-domain VQA must balance these competing demands.

8 Conclusion

In this paper, we introduced KWordinaryVQA, a large-scale, automatically generated dataset for visual question answering in the food domain. Our primary contribution is a novel, challenging benchmark designed to probe the reasoning capabilities of modern vision-language models. Our comprehensive evaluations demonstrate that while no single modeling paradigm excels across all metrics, a series of critical trade-offs exist between accuracy, semantic relevance, and computational efficiency.

The empirical results underscore the dataset's difficulty. We found that even powerful, fine-tuned models like BEiT-3 achieve only modest accuracy, while large generative models struggle with the dataset's concise formatting requirements, necessitating external post-processing steps. Furthermore, our experiments revealed that for deployment, refining the output of a cost-effective zero-shot model can be a more pragmatic and effective strategy than computationally expensive few-shot prompting.

Our work has two main limitations that open avenues for future research. First, the automated generation pipeline introduces a degree of semantic noise. Future work should focus on developing methods for automated noise filtering to further enhance dataset integrity. Second, our post-processing technique relies on a static answer-length threshold, limiting its applicability in dynamic, real-world systems. We believe that developing adaptive output refinement strategies is a crucial next step.

By publicly releasing the KWordinaryVQA dataset, our generation pipeline, and evaluation code, we aim to facilitate further research into developing more robust and accurate food-centric AI systems and to provide a valuable resource for benchmarking in this specialized domain.

Limitations

A foundational limitation of our study is the integrity of the dataset itself, which is constrained by semantic noise from the automated generation process. This noise manifests as factually incorrect ground-truth answers, where the generated text fails to align with the visual evidence (detailed in Section 5.3). The implications of this label noise are twofold. First, it corrupts the learning signal during training, potentially forcing the model to form erroneous associations rather than robust, generalizable knowledge. Second, it complicates evaluation, as a model providing a visually faithful answer may be marked incorrect, leading to an underestimation of its actual reasoning abilities. Consequently, our reported results should be viewed as a conservative baseline, acknowledging that model performance is likely suppressed by these data artifacts.

Beyond the data itself, a second limitation lies in the practical applicability of our post-processing technique. The method relies on a pre-determined threshold for answer length–for instance, instructing a model to generate 'no more than 5 words'. Setting an optimal threshold requires analyzing the entire dataset in advance to understand its global statistics. This assumption of having full, prior knowledge of the corpus is feasible for static, offline benchmarks but is unrealistic for real-world, dynamic systems where data arrives sequentially. This dependency thus restricts the direct deployment of this specific method and underscores the need for more adaptive post-processing strategies in future work.

Acknowledgments

Huy Trieu is supported by the research funding from the Faculty of Information Technology, University of Science, VNU-HCM, Vietnam. This work was also supported in part by the Air Force Office of Scientific Research under award number FA2386-24-1-4034 granted to Tung Le.

References

Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. 2024. Indifoodvqa: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176.

Allrecipes contributors. 2025. Allrecipes. Accessed: 2025-03-20.

Max Bachmann. 2024. Rapidfuzz. Version 3.12.1.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Vladimir Blagojevic. 2024. Bert english uncased finetuned pos. Accessed: 2025-04-15.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 2.0 flash. Accessed: 2025-03-25.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meta AI. 2024. Llama 3.2 11b vision instruct. Released under the Llama 3.2 Community License. Accessed: 2025-04-20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong

Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *CoRR*, abs/2410.12705.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Appendices

A Prompts for Data Generation

A.1 Image Description Generation

```
The following food items are present in
    this image: {food_name}.
Describe the color and relative
    location of each food item in
    details.

Instruction:
- Only print the final description, do
    not print anything else like
    headers or human-like response!
- The summary has maximum 200 words
- Do not break the line!
```

A.2 Keyword Extraction

```
Given a summary of an image for the VQA
    task, extract the top 6 important
    keywords, ensuring diversity in
    word types, including at least one
    verb, one adjective, one noun, one
    color, and one preposition (if
    present).

Summary: {summary}
Instruction:
- Only print the extracted keywords as
    a comma-separated list.
- If there is a preposition, include
    the full phrase containing it
    (e.g., 'on the table' instead of
    just 'on').
- Do not print anything else like
    headers or human-like responses!
```

A.3 Question-Answer Generation

```
Based on these keywords: {keywords} And
    the food image description:
    {summary}
Please generate one simple question per
    keyword where:
1. Each question is based on the food
    image description.
2. The answer to each question must
    exactly match its corresponding
    keyword
3. The number of questions must be
    equal to the number of keywords
4. For keywords that are prepositions
    (e.g., on the table), ask a Where
    question.
5. The order of the questions must
    match the order of the keywords.
Instruction:
- Generate only the questions as a
    comma-separated list.
- Do not include headers, explanations,
    or human-like responses.
```

B Prompt for Post-Processing

```
You are a helpful VQA assistant.
Your task is to extract a single, most
    relevant answer to the question,
    based on the given prediction text.
The answer must:
- Directly address the question's intent
- Contain no more than 5 words
- Be written entirely in lowercase
    letters
- Not include any commas, lists, or
    explanations
- Be concise and natural, like a
    typical VQA answer (e.g., 'red
    shirt', 'top left', 'enchiladas',
    'golden-brown', etc.)
- If multiple candidates appear in the
    prediction, select the one most
    relevant to the question
- Respond with only the final answer
    and nothing else

Question: {question}
Predict: {prediction}
```

C Training Configuration

Hyperparameter	Value
Optimizer	Adam
Learning rate	5×10^{-5}
Batch size	8
GPU	NVIDIA Tesla P100

Table 5: Training configuration for BEiT-3 and LXMERT.

For the BEiT-3 model, we fine-tuned only the classification head and the text embedding layers over 4 epochs, which took approximately 6 hours, while keeping the remaining parameters frozen. The LXMERT model was fine-tuned for 6 epochs, requiring about 1 hour.

D Illustrative Examples

Image	Question & Answer
	**Question:** What color are the corn kernels in the stew? **Answer:** yellow **Question:** Where is the dollop of sour cream located? **Answer:** atop the center
	**Question:** What is the state of the cheese layer on top of the pie? **Answer:** melted **Question:** What type of food is primarily located on the right side of the image? **Answer:** pie **Question:** What color is the crust of the pie? **Answer:** tan **Question:** What color is the baking tin the pie is sitting in? **Answer:** silver
	**Question:** What color are the string beans? **Answer:** green **Question:** How are the string beans arranged in the image? **Answer:** tangled **Question:** How are the string beans arranged in the image? **Answer:** tangled **Question:** Do the string beans overlap any other food in the image? **Answer:** yes **Question:** Where are the string beans located in the image? **Answer:** in the bottom center

Image	Question & Answer
	**Question:** What color are the falafels? **Answer:** golden-brown **Question:** What color is the lemon? **Answer:** yellow **Question:** What type of food is shown in the image? **Answer:** falafels
	**Question:** What is the predominant color of the lasagna noodles? **Answer:** yellow **Question:** What is the name of the dish described? **Answer:** lasagna **Question:** Where are the fresh green parsley leaves placed? **Answer:** on top
	**Question:** What type of pasta is in the center of the dish? **Answer:** linguine **Question:** What color is the shredded cheese? **Answer:** white

Table 6: Illustrative examples from the KWordinaryVQA dataset.

E Length Distribution

E.1 Training Set

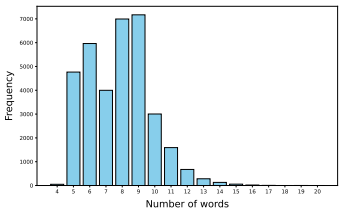


Figure 7: Question lengths distribution in training set.

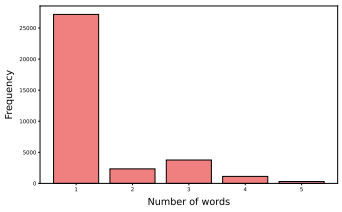


Figure 8: Answer lengths distribution in training set.

E.2 Validation Set

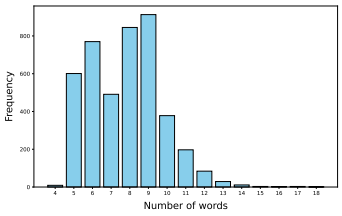


Figure 9: Question lengths distribution in validation set.

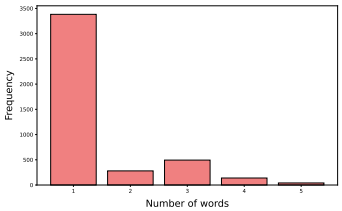


Figure 10: Answer lengths distribution in validation set.

E.3 Test Set

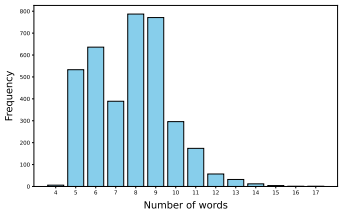


Figure 11: Question lengths distribution in test set.

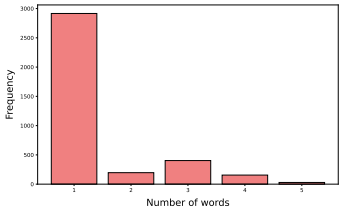


Figure 12: Answer lengths distribution in test set.

F Question Types Distribution

F.1 Training Set

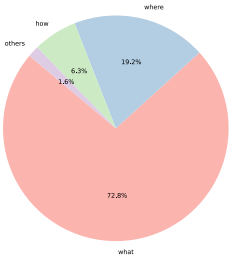


Figure 13: Distribution of question types in training set.

F.2 Validation Set

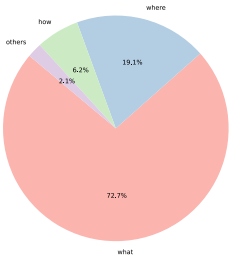


Figure 14: Distribution of question types in validation set.

F.3 Test Set

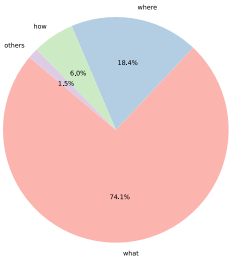


Figure 15: Distribution of question types in test set.

Model	Type	Time (hours)	Cost (VND)	GPU	Overall Evaluation
TF-IDF (Sparck Jones, 1972)	computed	**< 0.01**	**Free**	**None (CPU)**	Very Low Accuracy - Free
BEiT-3 (Wang et al., 2023)	fine-tuned	**0.06**	**Free**	P100	**Moderate Accuracy - Free**
LXMERT (Tan and Bansal, 2019)	fine-tuned	0.134	**Free**	P100	Low Accuracy - Free
Llama 3.2 Vision (Meta AI, 2024)	zero-shot	1.5	8,000	P40	Low Accuracy - Moderate Cost
Llama 3.2 Vision (Meta AI, 2024)	few-shot	4.5	100,000	A100 PCIE	Low Accuracy - Very High Cost
MiniCPM-o 2.6 (Yao et al., 2024)	zero-shot	1.5	8,000	P40	Low Accuracy - Moderate Cost
MiniCPM-o 2.6 (Yao et al., 2024)	few-shot	2.725	30,000	RTX 5090	Low Accuracy - Very High Cost
Qwen2.5-VL (Bai et al., 2025)	zero-shot	1.5	8,000	RTX 3090	Low Accuracy - Moderate Cost
Gemini 2.0 Flash (Google DeepMind, 2025)	zero-shot	1.5	**Free**	**None (API)**	Very Low Accuracy - Free

Table 7: Inference time, estimated cost (VND), GPU type, and qualitative evaluation for each model on the KWordinaryVQA test set. Zero-shot results are reported without including any post-processing time; optionally applying a post-processing step would add approximately 1.5 hours to the total inference time. For few-shot models, we only report results for the setting with 10 in-context examples. Reported inference time includes only model execution and excludes any additional refinement steps.

MoE4EDiReF: Mixture of Experts for Emotion Discovery and Reasoning its Flip in Conversation

Katarzyna Szczepaniak
Warsaw University of Technology
katarzyna.szczepaniak2.stud@pw.edu.pl

Piotr Andruszkiewicz
Warsaw University of Technology
IDEAS Research Institute
piotr.andruszkiewicz@pw.edu.pl

Abstract

Modern artificial intelligence systems are increasingly tasked with solving the problem of emotion recognition in conversation, which is a key element in improving the quality of human-computer interaction. However, recognizing emotions in dynamically conducted conversations remains a challenge. This paper proposes a new approach to the problem based on the Mixture of Experts technique, which allows for the simultaneous execution of both emotion recognition and the identification of utterances that cause emotion flips in conversations. The paper presents the results obtained for current approaches and the proposed method. The experiments were conducted on real-world datasets consisting of transcriptions of conversations. The experimental results indicate a significant improvement compared to other solutions.

1 Introduction

The ability to recognize emotions in dynamic conversations is an emerging challenge for modern AI systems and plays a vital role in advancing human-computer interaction. Traditional emotion classification solutions (Abdul-Mageed and Ungar, 2017; Akhtar et al., 2022) rely on static approaches, overlooking the complexity of emotional dynamics, including emotion changes triggered by specific utterances in conversations.

Integrating emotion recognition and identifying utterances that cause emotion changes into a single task supports the development of natural language processing (Kumar et al., 2024). This task involves two main aspects: emotion recognition in conversation (ERC) (Ghosal et al., 2019; Jiao et al., 2019) and emotion-flip reasoning (EFR) (Kumar et al., 2022), which aims to identify the utterances responsible for changing the emotional state of one of the speakers.

The goal of this paper is to develop a model for solving the tasks of emotion recognition and reasoning its flip in conversation. The proposed approach is based on the Mixture of Experts (MoE) technique (Eigen et al., 2013), which enables the simultaneous modelling of both the ERC and EFR tasks. The paper aims to develop a model that achieves results comparable to or better than currently available solutions.

The conducted research focused on four experimental areas:

1. Identifying the impact of the number and type of gating networks, the number of experts, and the type of experts used on the quality of the model,

2. Investigating the effectiveness of activating only the top k experts during model training,

3. Assessing the impact of the learning rate and the number of epochs on the results using the best models selected in the previous stages of the experiments,

4. Examining the effectiveness of translating the dataset from Hindi to English prior to model training.

2 Related Work

This study builds upon a comprehensive review of existing literature, focusing primarily on the task of Emotion Recognition and Reasoning its Flip in Conversation (EDiReF), recently introduced by Kumar et al. (Kumar et al., 2024). This interdisciplinary task combines emotion classification with the identification of utterances that trigger emotional shifts within conversations. While this area remains relatively new in the research community, early work highlights significant challenges, including the modeling of temporal dependencies, participant interactions, and the dynamics of emotional change over time.

Due to the limited number of publications directly addressing EDiReF, the literature review also encompasses broader research on emotion recognition (ER) and, more specifically, emotion recognition in conversation (ERC). ERC methods commonly fall into three categories:

1. **Knowledge-based techniques:** These use lexicons and rule-based systems such as WordNet (Miller, 1994), SenticNet (Cambria et al., 2016), or ConceptNet (Speer et al., 2016). They are interpretable but limited by linguistic ambiguity and poor scalability.

2. **Statistical methods:** Traditional machine learning models, including Support Vector Machines (SVM) (Chavhan et al., 2010) and Naive Bayes classifiers (Sun et al., 2017), perform well on well-structured datasets. More recently, deep learning approaches like RNNs (Li et al., 2021) and transformer-based models such as BERT (Devlin et al., 2019; Bhat, 2024) have become standard due to their context-aware representations and scalability.

3. **Hybrid methods:** These integrate structured knowledge with data-driven models, enhancing performance and interpretability. Notable examples include (Gievska et al., 2015), which combines affective lexicons with deep learning.

Research in ERC typically adopts either supervised or unsupervised approaches. Supervised methods include DialogueRNN (Majumder et al., 2018), DialogueGCN (Ghosal et al., 2019), and newer transformer-based systems like BERT-ERC (Qin et al., 2023) and ERC-DP (Wang et al., 2024). These methods require large annotated corpora but demonstrate state-of-the-art results.

Conversely, unsupervised methods such as clustering techniques (e.g., SCCL (Yang et al., 2023) and DeepEmoCluster (Lin and Busso, 2024)) and probabilistic models like Hidden Markov Models (HMMs) (Nwe et al., 2003; Schuller et al., 2003) are useful for low-resource settings, though generally less accurate.

The EDiReF task itself was first modeled by Kumar et al. (Kumar et al., 2022), who proposed a two-stage architecture combining emotion recognition and flip reasoning. They used a Masked Memory Network (MMN) for utterance-level ERC and a Transformer-based model (TX) for instance-level EFR. MMN hierarchically encodes contextual information from past utterances, while TX models cause-effect relationships across the dialogue. Their system outperformed several baselines, including CMN (Hazarika et al., 2018b), ICON (Hazarika et al., 2018a), DialogueGCN (Ghosal et al., 2019), and AGHMN (Jiao et al., 2019). Additionally, they released MELD-FR, a benchmark dataset derived from MELD, annotated for flip reasoning tasks.

The EDiReF task was introduced as a shared task at SemEval 2024 (Kumar et al., 2024) (Semantic Evaluation is a series of research workshops focused on natural language processing, aimed at advancing knowledge in the field of semantic analysis) with three subtasks: the ERC task on a Hindi dataset (referred to as task A), the EFR task on a Hindi dataset (referred to as task B), and the EFR task on an English dataset (referred to as task C).

This setup allowed for multilingual evaluation and benchmarking. The top systems employed large language models (LLMs) such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2019), and Zephyr (Tunstall et al., 2023), often fine-tuned or instruction-tuned for the task. Approaches like prompting and few-shot learning were also explored. The best-performing systems achieved F1 scores of 0.70, 0.79, and 0.76 for tasks A, B, and C, respectively, indicating their high effectiveness in the context of the task being solved. In the case of task C, this also represents an improvement over the results obtained in the original paper (0.33 for MMN and 0.45 for TX).

3 Methodology

In this section, the idea behind the EDiReF task is explained, followed by a description of the datasets used for this task. Next, the applied method is discussed, and the evaluation methods used to evaluate the performance of the model, which make it possible to accurately measure the effectiveness of the proposed solution and compare it with existing solutions, is presented.

3.1 Problem Definition

In the context of the SemEval 2024 workshop (Kumar et al., 2024), the problem of emotion recognition and the identification of utterances causing emotional changes in conversation was defined as the ERC task and the EFR task on a dataset composed of the phonetic transcriptions of conversations in the Hindi language and on a dataset in

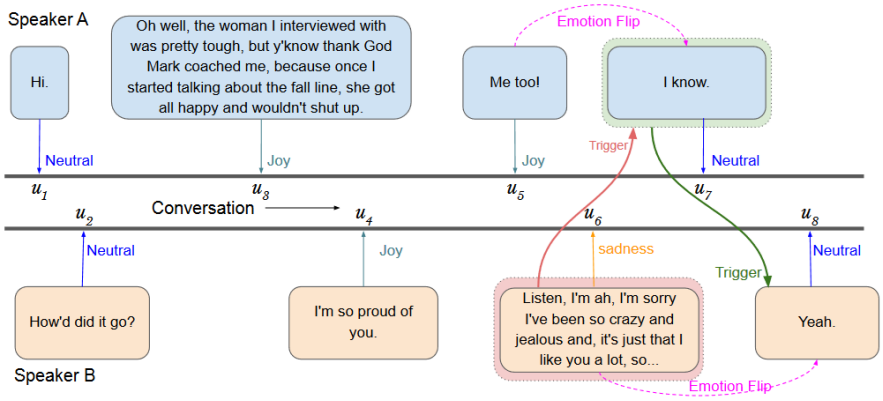


Figure 1: An example of a conversation with assigned emotion labels and utterances marked as contributing to emotional flips (Kumar et al., 2024).

English.

Figure 1 presents an example conversation and the key idea behind the EDiReF task. The figure shows a conversation between two speakers – Speaker A and Speaker B. For each statement (labeled consecutively from u_1 to u_8), an emotion is assigned. Additionally, the statements u_6 and u_7 are highlighted as those causing emotional changes.

The model solving the EDiReF task should assign to each utterance in the provided conversation an emotion label from the set of emotions, as well as a label indicating whether the statement causes an emotional change in the conversation partner.

3.1.1 Emotion Recognition in Conversation

In the ERC task, the model receives textual utterances as input, and its goal is to predict the emotion label for each utterance. Each conversation is represented as a list of tuples $D_{ERC} = \{(s_1, u_1), (s_2, u_2), \ldots, (s_n, u_n)\}$, where s_i denotes the speaker of the utterance u_i. The goal of the model is to predict the emotion e_i for each utterance u_i.

3.1.2 Emotion-Flip Reasoning

In the EFR task, the model analyzes the utterances presented in the form of a list of tuples $D_{EFR} = \{(s_1, u_1, e_1), (s_2, u_2, e_2), \ldots, (s_n, u_n, e_n)\}$, where e_i denotes the emotion present in the utterance u_i spoken by the speaker s_i. The goal is to identify the utterances t_i that trigger an emotion change in the conversation partner and label them as triggers with a value of 1. Otherwise, a value of 0 is assigned.

3.2 Datasets

Two datasets were used to train and verify the models. Both datasets were extended with labels needed for the EFR task. The annotations were prepared by professionals in the field of dataset annotation.

3.2.1 MELD

The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018) is a dataset in English that is used for the ERC task. In the SemEval 2024 workshop and for the purposes of this work, a modified version of the MELD dataset, i.e. the MELD-FR dataset, was used.

There are seven emotion labels in the dataset – disgust, joy, surprise, anger, fear, neutral and sadness. The neutral label is assigned to those utterances that do not show the emotionality typical of the other emotions.

3.2.2 MaSaC

MaSaC (Bedi et al., 2021) is a set of conversations in Hindi using the phonetic transcription of the language. The conversations are from the TV series *Sarabhai vs Sarabhai* and for the purposes of the SemEval Workshop have been annotated with labels covering eight emotions and labels indicating triggers in the conversations. Seven of the eight labels overlap with labels from the MELD-FR dataset. An additional label in the MaSaC dataset is *contempt*.

3.3 Mixture of Experts Technique

Mixture of Experts (MoE) is a machine learning technique that uses multiple networks, called experts, to divide the problem space into smaller problems (Eigen et al., 2013; Du et al., 2021). During training, each expert is trained, but during testing, depending on the input, only a subset of experts is activated, allowing the trained model to generate answers faster.

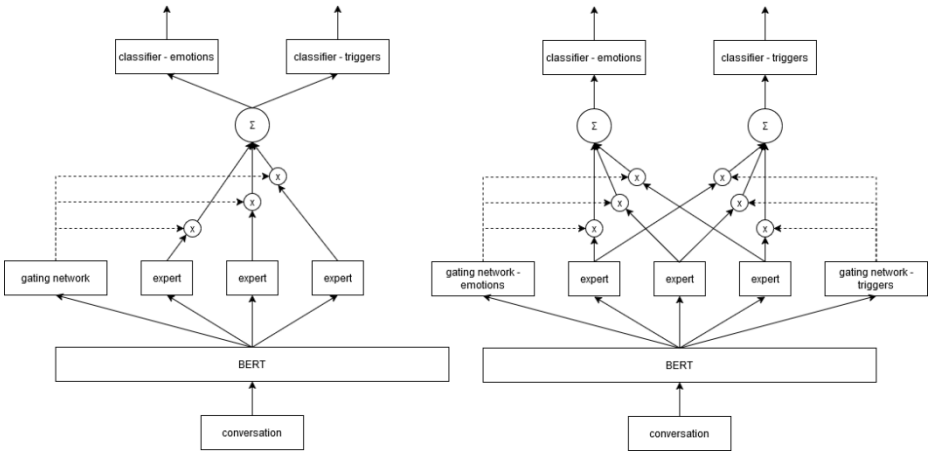


Figure 2: Model architectures with a single (left) and dual (right) gating network used for solving the EDiReF task.

3.3.1 Expert

An expert is a single, specialized model or module that is designed to solve a specific task. Experts operate independently and are responsible for generating results for their part of the problem.

3.3.2 Gating Network

A gating network is a mechanism that decides which experts should be activated for a given input. It works by assigning weights to experts, indicating which of them are best suited to solve a given task.

3.4 Methods of Model Evaluation

The F1 Score was used to assess the quality of the model. The F1 Score is the harmonic mean of precision and recall, balancing these two measures. It is a metric often used in natural language processing to compare solutions.

4 Experimental Setup

This section discusses the model architecture and the set of hyperparameters used in the experiments and presents the details of the experiments conducted to evaluate the effectiveness of the MoE technique in the tasks of recognizing emotions and utterances causing emotion change in conversations.

4.1 Model Architecture

Two model architectures employing the Mixture of Experts technique were designed for the EDiReF task. The first architecture uses a single gating network, while the second incorporates two separate gating networks, each dedicated to a different task: emotion recognition in conversation and identification of utterances that trigger emotional changes. Introducing a second gating network enables a more precise alignment of experts with the specific requirements of each task, potentially improving both emotion recognition and the detection of emotion triggers. Figure 2 (left) illustrates the architecture with a single gating network. In this version, conversations are passed to the model input, where a BERT model generates embedding vectors. These vectors are then sent to both the experts and the gating network, which computes a weight vector used to aggregate the responses from individual experts. The weighted outputs are then combined and forwarded to the classifiers responsible for emotion and trigger classification. Figure 2 (right) shows the dual-gate architecture. Each gating network is assigned to a different task: one to emotion recognition (ERC) and the other to emotion trigger recognition (EFR). This approach allows each gating network to be trained independently in the context of its specific task, enabling a more tailored selection of experts to meet distinct task requirements.

In both datasets, conversations are stored as lists of strings. Since BERT requires a single string as input, all utterances in a conversation are concatenated, separated by the special `[SEP]` token. This enables the model to better capture the dynamics of the conversation.

The preprocessed conversation is then passed to BERT, which produces a vector representation in the embedding space. For the MELD dataset and the translated MaSaC dataset, the `bert-base-cased`[1] model is used. For the original Hindi MaSaC dataset, the

[1] https://huggingface.co/google-bert/bert-base-cased

	Gate type	N. of gates	Expert type	Number of experts	ERC	EFR
M1	Linear	1	Linear	2	88.9	**79.8**
M2	Linear	1	MLP	2	83.8	79.0
M3	Linear	1	LSTM	2	74.2	78.5
M4	MLP	2	Linear	2	**89.6**	79.2
M5	MLP	1	Linear	4	89.2	79.7
M6	MLP	1	LSTM	4	72.8	79.3
M7	MLP	2	Linear	4	89.4	79.4
M8	MLP	2	MLP	4	83.2	79.2

Table 1: Comparison of the results obtained for the MELD dataset and the different setups of gating networks and experts.

	Gate type	Number of gates	Expert type	Number of experts	ERC	EFR
S1	Linear	1	Linear	2	**91.8**	89.5
S2	Linear	1	LSTM	2	65.5	88.2
S3	MLP	1	Linear	2	90.2	89.5
S4	MLP	1	LSTM	4	64.6	**89.8**
S5	MLP	1	Linear	8	90.4	89.2
S6	MLP	1	MLP	8	83.0	89.3
S7	Linear	2	Linear	8	90.6	**89.8**
S8	MLP	2	MLP	8	81.3	89.7

Table 2: Comparison of the results obtained for the MaSaC dataset and the different setups of gating networks and experts.

`bert-base-multilingual-cased`[2] model is employed.

The gating networks and experts may take the form of a linear layer or a multilayer perceptron. Experts can also be implemented using long short-term memory (LSTM) networks.

4.2 Hyperparameters

In all experiments we used the same set of the following hyperparameters, which ensures consistency of results and enables direct comparison of performance of different model configurations. The batch size was set to 32. To reduce overfitting, dropout regularization with a value of 0.1 was applied. The loss function for the emotion recognition task was CrossEntropyLoss. For the trigger identification task, BCEWithLogitsLoss was used.

4.3 Conducted Experiments

In order to investigate the effectiveness of the Mixture of Experts technique, three stages of experiments were designed. Each stage examines different aspects of the MoE technique. Additionally, it was also assessed whether translating the MaSaC dataset into English before training improves the results for the final model. Furthermore, the performance of the best models with that of other existing approaches was compared. The F1 measure in the form of a percentage value was used as a criterion for evaluation and comparison.

4.3.1 Impact of Type and Number of Gating Networks and Experts

During the first stage of experiments, it was examined how the quality of the model is affected by the use of either one or two gating networks with different architectures. A single linear layer and a multilayer perceptron were investigated.

The influence of different numbers of experts with different architectures was also examined. A single linear layer, a multilayer perceptron and a long short-term memory network were investigated as possible expert architectures. The choice of the number of experts was motivated as follows: with two experts, each could specialize in one task (ERC or EFR). In the case of four experts, one could handle EFR, while the rest of the experts would handle ERC divided into positive, negative and neutral emotions. With eight experts, one could be assigned to EFR, and the rest to individual emotions

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased

	Gate type	Number of gates	Expert type	Number of experts	Top k	ERC	EFR
M9	MLP	1	Linear	4	1	86.8	**80.1**
M10	MLP	1	Linear	4	2	87.9	79.6
M5	MLP	1	Linear	4	4	**89.2**	79.7
M11	MLP	2	Linear	4	1	88.2	**79.6**
M12	MLP	2	Linear	4	2	88.4	79.1
M7	MLP	2	Linear	4	4	**89.4**	79.4

Table 3: Comparison of the results obtained for the MELD dataset and the different number of experts activated during training.

	No.G.	No.E.	Top k	ERC	EFR
S9	1	2	1	90.8	89.0
S1	1	2	2	**91.8**	**89.5**
S10	2	8	1	88.5	89.5
S11	2	8	2	90.4	89.4
S12	2	8	4	88.4	89.4
S7	2	8	8	**90.6**	**89.8**

Table 4: Comparison of the results obtained for the MaSaC dataset and the different number of experts activated during training. All configurations with linear gate type and linear expert type. No.G. - Number of Gates, No.E. - Number of Experts.

in ERC. In this set of experiments, the models were trained for 5 epochs with a learning rate of $2{\cdot}10^{-5}$.

Selected results obtained within the first set of experiments investigating the impact of type and number of gating networks and experts on the MELD and MaSaC datasets are presented in Table 1 and 2, respectively. The best results for ERC and EFR are marked in bold.

Models M4 and M1, which achieved the highest F1 scores for ERC and ERF, respectively, utilized a double (M4) and a single (M1) gating network, implemented as a multilayer perceptron and four experts implemented as individual linear layers. The remaining four best-performing model configurations, two per task, also used linear layers as experts, suggesting that for the characteristics of the MELD dataset, using simpler experts was beneficial. No significant difference was observed between approaches employing dual gating networks and those using a single gate, nor was there an improvement in performance when using a more complex gating architecture such as a multilayer perceptron.

The results on the MaSaC dataset support the hypothesis that when using the mixture of experts technique, a very simple architecture—such as a linear layer—is sufficient to achieve strong performance (F1 score of 91.8 in the case of S1 for ERC and 89.5 for EFR). An interesting difference between the results for the MELD and MaSaC datasets is the relationship between the number of experts used and the achieved performance. In the case of MaSaC, a significant number of high-performing models employed 8 experts, whereas for the MELD dataset, a smaller number of experts was preferred (with 4, and in some cases even just 2, experts yielding better results).

4.3.2 Impact of Activating Top k Experts During Training

To compare the efficiency of activating the top k experts, i.e. limiting the number of activated experts, two models, which achieved the best result in the first set of experiments, were selected for each dataset.

The summary of the results on the MELD dataset is presented in Table 3. The results for the MaSaC dataset are presented in Table 4.

Regardless of the model configuration, those that activated all available experts during training achieved better results than models that activated only the top-k experts. Additionally, no improvement in training speed was observed when using top-k activation, and thus this approach is not recommended.

Similarly, for the MaSaC dataset, the use of top-k expert activation did not yield any benefits in terms of higher F1 scores or shorter training times.

4.3.3 Impact of Learning Rate and Number of Epochs

In the final stage of the experiments, three model configurations that had so far achieved the best results were used and the influence of learning rates of $2{\cdot}10^{-5}$, $3{\cdot}10^{-5}$, $4{\cdot}10^{-5}$ and $5{\cdot}10^{-5}$ and the number of epochs of 3, 5 and 7 was investigated. These parameters were selected based on the recommendations from the article on the BERT model (Devlin

	L. rate	N. of epochs	ERC	EFR
M13	$2{\cdot}10^{-5}$	3	74.9	78.8
M14	$5{\cdot}10^{-5}$	3	87.9	79.4
M5	$2{\cdot}10^{-5}$	5	89.2	**79.7**
M15	$5{\cdot}10^{-5}$	5	93.3	78.1
M16	$2{\cdot}10^{-5}$	7	92.1	79.4
M17	$3{\cdot}10^{-5}$	7	**93.9**	79.4
M18	$5{\cdot}10^{-5}$	7	**93.9**	78.9

Table 5: Comparison of the results obtained for the MELD dataset and the different learning rates and number of epochs.

	L. rate	N. of epochs	ERC	EFR
S13	$2{\cdot}10^{-5}$	3	64.3	**90.0**
S14	$5{\cdot}10^{-5}$	3	80.3	89.3
S1	$2{\cdot}10^{-5}$	5	91.8	89.5
S15	$5{\cdot}10^{-5}$	5	96.6	89.7
S16	$2{\cdot}10^{-5}$	7	95.8	**90.0**
S17	$4{\cdot}10^{-5}$	7	**97.7**	89.4
S18	$5{\cdot}10^{-5}$	7	97.0	89.5

Table 6: Comparison of the results obtained for the MaSaC dataset and the different learning rates and number of epochs.

et al., 2019), indicating them as some of the optimal ones.

The summary of the results on the MELD dataset is given in Table 5. The model used in the experiments had a single MLP-type gating network and 4 linear experts, which all were activated during training.

The results obtained on the MaSaC dataset are shown in Table 6. The model used in the experiments had a single linear gating network and 2 linear experts, both of which were activated during training.

Models M15 and M16, which were trained for 3 epochs, achieved the worst results among all considered configurations. In particular, model M15, trained with a learning rate of $2 \cdot 10^{-5}$, had the lowest performance, reaching an F1 score of 76.85. Models M17 and M18, trained for 7 epochs with learning rates of $3 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ respectively, achieved similar results – the average F1 score for the ERC task was 93.9 for both models, and the average F1 score for the EFR task differed by only 0.5 percentage points.

It can therefore be concluded that increasing the number of epochs from 5 to 7, as well as raising the learning rate from $2 \cdot 10^{-5}$ to $3 \cdot 10^{-5}$ or even $5 \cdot 10^{-5}$, has a positive effect on the final results.

For the MaSaC dataset, as well as for the MELD dataset, models trained for 7 epochs generally achieved better results than those trained for only 3 or 5 epochs, regardless of the learning rate value.

4.3.4 Impact of Translating MaSaC Dataset into English

The pyhinavrophonetic[3] library and a variant of the NLLB-200 model developed by Facebook were used to translate the MaSaC dataset into English.

[3]Available at: https://pypi.org/project/pyhinavrophonetic/, Last accessed: 08.01.2025.

	No.G.	No.E.	Top k	ERC	EFR
MaSaC					
T1	1	2	2	97.7	89.4
T2	2	8	2	97.6	89.4
T3	2	8	2	**98.2**	88.8
T4	2	8	8	97.3	**89.7**
Translation of MaSaC					
T1	1	2	2	96.7	87.9
T2	2	8	2	96.7	88.5
T3	2	8	2	96.7	88.5
T4	2	8	8	96.7	87.6

Table 7: Comparison of the results obtained for the MaSaC dataset and its translation. All configurations with linear gate type and linear expert type. No.G. - Number of Gates, No.E. - Number of Experts.

The script iterated through the MaSaC dataset, converting each conversation from the phonetic script of the Hindi language to the official alphabet, and then passed it to the NLLB-200 model, which can translate text between multiple languages.

Four best model configurations from the previous experiments were used. Each model was trained for 7 epochs. Models T1, T3, and T4 with a learning rate of $4{\cdot}10^{-5}$, and model T2 with a learning rate of $3{\cdot}10^{-5}$. The results are summarized in Table 7.

For both datasets, the models achieved very good results. However, when comparing paired configurations (i.e., T1 on the MaSaC dataset and T1 on the translated version of MaSaC), it becomes evident that the preliminary translation of the Hindi dataset into English did not yield significant benefits. Each model trained on the translated MaSaC dataset achieved lower average F1 scores for the ERC task, the F1 score for the EFR task, or the average F1 score across both tasks.

	ERC								EFR
	Dg	Jy	Sr	An	Fr	Ne	Sa	Avg	Trigger
MMN	20,2	48,7	50,4	42,9	9,80	71,9	29,6	55,7	33,4
TX	0,00	4,00	5,00	1,90	0,00	61,2	0,00	29,5	44,8
GAVx	-	-	-	-	-	-	-	-	76,0
MoE-1	84,9	93,4	93,6	**91,8**	90,4	95,0	**94,4**	93,9	79,4
MoE-2	**86,1**	**93,8**	**94,0**	91,4	**92,8**	**95,2**	92,8	**94,0**	**79,7**

Table 8: Comparison of the results obtained for the MELD dataset (Dg: disgust, Jy: joy, Sr: surprise, An: anger, Fr: fear, Ne: neutral, Sa: sadness).

	ERC									EFR
	Dg	Jy	Sr	An	Fr	Ne	Sa	Co	Avg	Trigger
TW-NLP	-	-	-	-	-	-	-	-	46.0	79.0
FeedForward	-	-	-	-	-	-	-	-	51.0	77.0
UCSC NLP	-	-	-	-	-	-	-	-	45.0	79.0
MoE-3	**98.0**	97.7	**96.8**	97.2	96.7	98.4	96.0	96.2	97.7	**89.4**
MoE-4	96.5	**98.4**	96.5	**97.4**	**97.9**	**98.8**	**97.2**	**97.2**	**98.2**	88.8

Table 9: Comparison of the results obtained for the MaSaC dataset (Dg: disgust, Jy: joy, Sr: surprise, An: anger, Fr: fear, Ne: neutral, Sa: sadness, Co: contempt).

4.3.5 Comparison of the Results for the MELD Dataset

The results obtained by Kumar et al. (Kumar et al., 2022), the GAVx team (Nguyen and Zhang, 2024) that achieved the best results in the SemEval 2024 workshop for task C, and our own results (models MoE-1 and MoE-2) are summarized in Table 8.

MoE-1 uses a single MLP-type gating network and 2 linear experts, and MoE-2 uses 2 MLP-type gating networks and 4 linear experts. Both models were trained for 7 epochs with a learning rate of $3{\cdot}10^{-5}$.

The comparison shows that, for the ERC task, each of the proposed models outperformed the models introduced by Kumar et al. Unfortunately, it was not possible to compare these results with those of the GAVx teams due to the unavailability of their scores.

However, comparison is possible for the EFR task. In this case, the proposed models also achieved better results than those reported in the original paper introducing the EDiREF task. Furthermore, they outperformed the top participants of the SemEval 2024 workshop, exceeding the first-place result by at least 3.1 percentage points.

4.3.6 Comparison of the Results for the MaSaC Dataset

The results obtained by the TW-NLP (Tian et al., 2024), FeedForward (Shaik et al., 2024) and UCSC NLP (Wan et al., 2024) teams, which achieved the best results in the SemEval 2024 workshop for tasks A and B, and our own results (models MoE-3 and MoE-4) are presented in Table 9.

MoE-3 uses a single linear gating network and 2 linear experts, and MoE-4 uses 2 linear gating networks and 8 linear experts, of which the best 2 were activated during training. Both models were trained for 7 epochs with a learning rate of $4{\cdot}10^{-5}$.

For the MaSaC dataset, it is not possible to compare classification results for individual emotions; however, it is possible to compare the average F1 scores for the ERC and EFR tasks. In both cases, each of the proposed models achieved better results than the best-performing model presented at the SemEval 2024 workshop. For ERC, the difference is at least 46.3 percentage points, while for EFR, the difference amounts to 9.8 percentage points.

5 Conclusions

The presented experimental results show that the use of the Mixture of Experts technique is an effective solution for the task of recognizing emotions and reasoning its flip in conversation. The solution achieved higher results than the current solutions, which means that the intended goal of the work was achieved.

It was observed that using a smaller number of experts with a simpler architecture, such as a linear layer or a simple multilayer perceptron, and train-

ing the model for a larger number of epochs with a lower learning rate has a positive effect on the final performance of the model. It was not observed that changing the number or type of gating network or increasing the number of experts significantly improved the results. However, using the activation of top k experts during training worsened the result, so it is not recommended to use this solution.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2022. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, 13:285–297.

Manjot Bedi, Shivani Kumar, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14:1363–1375.

Siddhanth Bhat. 2024. Emotion classification in short english texts using deep learning techniques. *ArXiv*, abs/2402.16034.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Bjoern Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, Osaka, Japan. The COLING 2016 Organizing Committee.

Yashpalsing Chavhan, Manikrao Dhore, and Yesaware Pallavi. 2010. Speech emotion recognition using support vector machines. *International Journal of Computer Applications*, 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2021. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *CoRR*, abs/1312.4314.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sonja Gievska, Kiril Koroveshovski, and Tatjana Chavdarova. 2015. A hybrid approach for emotion detection in support of affective interaction. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2015:352–359.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2019. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI Conference on Artificial Intelligence*.

Shivani Kumar, Md. Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. SemEval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1933–1946, Mexico City, Mexico. Association for Computational Linguistics.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang. 2021. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683.

Wei-Cheng Lin and Carlos Busso. 2024. Deep temporal clustering features for speech emotion recognition. *Speech Communication*, 157:103027.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and E. Cambria. 2018. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI Conference on Artificial Intelligence*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Vy Nguyen and Xiuzhen Zhang. 2024. GAVx at SemEval-2024 task 10: Emotion flip reasoning via stacked instruction finetuning of LLMs. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 326–336, Mexico City, Mexico. Association for Computational Linguistics.

Tin Nwe, S.W. Foo, and Liyanage De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ArXiv*, abs/1810.02508.

Xiangyu Qin, Zhiyu Wu, Jinshi Cui, Ting Zhang, Yanran Li, Jian Luan, Bin Wang, and L. xilinx Wang. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. *ArXiv*, abs/2301.06745.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden markov model-based speech emotion recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2:401–404.

Zuhair Hasan Shaik, Dhivya Prasanna, Enduri Jahnavi, Rishi Thippireddy, Vamsi Madhav, Sunil Saumya, and Shankar Biradar. 2024. FeedForward at SemEval-2024 task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 745–756, Mexico City, Mexico. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

Wei Tian, Peiyu Ji, Lei Zhang, and Yue Jian. 2024. TW-NLP at SemEval-2024 task10: Emotion recognition and emotion reversal inference in multi-party dialogues. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 311–315, Mexico City, Mexico. Association for Computational Linguistics.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944.

Neng Wan, Steven Au, Esha Ubale, and Decker Krogh. 2024. UCSC NLP at SemEval-2024 task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1492–1497, Mexico City, Mexico. Association for Computational Linguistics.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14:3269–3280.

Diagnosing the Cultural Gap in Machine Translation of Classical Chinese: A Computational Analysis of *The Analects*

Menghan Dong* **and Youzhi Wu and Siqi Wang and Zhen Wu**
The Hong Kong Polytechnic University
24072087g@connect.polyu.hk,
24067312g@connect.polyu.hk,
24104191g@connect.polyu.hk,
24138226g@connect.polyu.hk

Abstract

This paper presents a computational diagnostic of how current AI systems translate classical Chinese, using *The Analects* as a case study. We formalize a multi-view evaluation framework that integrates lexical overlap (Jaccard similarity), distributional semantics (cosine similarity), keyword salience (TF–IDF), topic structure (LDA), semantic networks, and sentiment analysis. Applying the framework to a representative human translation (Legge) and an AI translation (ChatGPT), we find high distributional consistency (cosine $\approx$0.91) but low lexical overlap (Jaccard $\approx$0.33), fragmented thematic structure in AI outputs (7 topics vs. 2 in the human translation), and attenuated affective intensity. A qualitative error taxonomy further reveals challenges unique to classical Chinese: homophonic loan characters, polysemy and sense selection, ancient–modern semantic drift, ellipsis/inversion, and value-laden terms. Our study complements standard MT evaluation by moving beyond surface-form metrics (e.g., n-gram overlap or learned estimators) to operationalize cultural and thematic fidelity for classical Chinese. We discuss how this multi-view diagnostic can guide sense-aware MT evaluation and modeling for low-resource, pre-modern language varieties.

1 Introduction

In recent years, artificial intelligence (AI) has improved translation, including literary translation, by increasing speed and efficiency and raising accuracy in specific settings. This progress has broadened access to literary works and provided technical support for translators, while also enabling wider access to Chinese materials for non-Chinese readers. At the same time, evaluating translation quality for pre-modern language varieties such as classical Chinese remains difficult: surface n-gram metrics and even learned estimators often miss sense selection, value-laden terminology, and culturally anchored themes. To examine these issues concretely, we focus on the English translation of *The Analects* as a case study. Our goal is to assess how effectively AI captures thematic depth and cultural subtleties in a classical text. We hypothesize that AI translation can capture major themes in *The Analects*, and we test this through a comparative analysis of AI generated translations and those produced by human scholars, with attention to the accuracy and depth of thematic and ethical representation. Our objective is to propose and validate a model-independent diagnostic framework for assessing systems' understanding of Classical Chinese, rather than to rank the translation performance of specific models. Accordingly, we adopt a widely accessible general-purpose large language model (ChatGPT) as the reference baseline and compare its output with a human translation. Domain-specific models for Classical Chinese (e.g., "XunziLLM") may involve uncertainties in version availability, interface stability, and licensing, making it difficult in the short term to meet the level of reproducibility required for our experiments. To preserve comparability under identical preprocessing and evaluation pipelines, we therefore prioritize publicly accessible models with stable access. We use James Legge's translation as the human reference, and we generated AI translations with ChatGPT and Tongyi Qianwen; after comparing the two, we retain ChatGPT's output for detailed analysis. Legge's translation adopts smaller units that follow original order and is widely treated as a standard version, and in our study it serves as a culturally grounded reference rather than a sen-

*Corresponding author

tence level gold standard. Our research questions are as follows: RQ1 (Semantic consistency): How closely-aligned are the AI and human translations at the distributional semantic level? RQ2 (Lexicon and themes): Do the AI translations exhibit systematic deviations in lexicalization and thematic structure (e.g., term generalization, topic fragmentation)? RQ3 (Affect and value terms): Do the AI translations attenuate affective intensity or mishandle value-laden terminology central to Confucian ethics?

2 Literature Review

Artificial intelligence (AI) translation tools have advanced rapidly and expanded assistive use in literary translation, yet they do not replace human expertise for culturally dense texts (Massion, 2017). From an NLP perspective, evaluation remains difficult: standard automatic metrics (BLEU, TER, chrF; learned metrics such as BERTScore, COMET) emphasize surface overlap or sentence level adequacy/fluency but under represent sense selection, value-laden terms, allusions, and culturally anchored themes typical of classical texts. Empirical comparisons reflect this gap. Ding (2024) contrasts human and AI renderings of Li Bai's "Farewell to a Friend," noting human advantages in stylistic nuance and lexical choice, with AI useful for research/learning but uneven on poetic effects. Zaid and Bennoudi (2023) compare human, ChatGPT, and Google Translate on Arabic religious texts, finding automatic systems fairly accurate yet weaker on depth, cultural relevance, and nuanced understanding—dimensions only partially reflected by automatic scores. Al Sawi and Allam (2024) analyze Arabic subtitles for Birdman and show that humans better handle allusions and cultural cues, while AI exhibits limitations on culturally complex content. In NLP evaluation, this motivates complements to surface metrics—distributional similarity, topic modeling, co-occurrence networks, and sentiment/affect profiling—which better probe lexical, thematic, and affective fidelity in classical material. Our study follows this line, using *The Analects* as a testbed to examine where AI aligns with a human reference and where it drifts in lexicon, themes, and affect.

3 Data and Methodology

3.1 Data Sources

We selected James Legge's public domain translation as the human reference and used Tongyi Qianwen and ChatGPT to translate *The Analects*. After evaluating and comparing the two AI translations, we retained ChatGPT's output as the representative machine translation for this study. In the observations, Qwen frequently (i) mixes bilingual translation (e.g., pinyin with parenthetical glosses explanation), (ii) introduces interpretive or modernized paraphrasing, and (iii) alternates near-synonyms or explanatory term pairs for single concepts, which will interfere the accuracy and affect the stability of the following measurement.

3.2 Collection Methods

We extracted Legge's translation from Ctext and submitted the original classical Chinese text of *The Analects* to the AI systems on a chapter-aligned basis to obtain machine translations. All translations were saved as UTF 8 plain text (.txt) files.

3.3 Processing Techniques

We used R and Python to preprocess the .txt files. Steps included Unicode normalization, normalization of quotation marks and hyphens, lowercasing, punctuation removal (except apostrophes), whitespace cleanup (removing redundant spaces), and regex-based English tokenization (RegexpTokenizer). Stopwords were removed for specific analyses; for LDA (and overlap) we used an expanded list including domain-dominant terms (e.g., confucius, master, gentleman, said, zi), while for TF–IDF we used only general English stopwords. Lemmatization was applied for LDA.

3.4 Data Analysis

Using Python and R, we conducted the following analyses: word frequency, word cloud, cosine similarity, Jaccard similarity, TF–IDF, LDA topic modeling, semantic co-occurrence network analysis, and sentiment/affect analysis. Section 4 details each component (4.1–4.7).

3.5 Task Formulation and Metrics

Task. Given a classical Chinese source segmented by canonical chapters, a human translation (H) and an AI translation (M) in the same target language, we quantify the cultural/thematic fidelity of M relative to H across complementary views.

Metrics. Lexical overlap: Jaccard similarity over content word vocabularies after stopword removal (optional lemmatization). Distributional semantics: cosine similarity between TF–IDF vectors at chapter and document levels. Keyword salience: compare TF–IDF top-100 terms and weight concentrations across H and M. Thematic structure: LDA topic modeling; select optimal K via coherence and compare K(H) vs. K(M) and topic separability. Semantic networks: co-occurrence graphs; compare core node centrality and community structure (e.g., modularity). Affective profile: lexicon based polarity/intensity distributions at the chapter level. Qualitative checks: classical specific phenomena (ellipsis, inversion, ancient–modern semantic drift, polysemy, value-laden terms). Stopwords are removed for overlap and LDA with an expanded domain-specific list, while TF–IDF uses only general stopwords; lemmatization is applied for LDA (optional elsewhere). Random seeds are fixed for vectorization and LDA, and chapter level metrics are aggregated to document level summaries.

4 Multi-View Diagnostic Analysis

4.1 Word Frequency and Word Cloud

By conducting a word frequency computing and word cloud analysis, we can observe the similarities and differences in lexical usage between the two translations. The statistical results show that there are certain similarities in the distribution of high-frequency words between these two versions. For example, the words "Confucius" and "Master" both refer to Confucius, and "virtue" refers to personal morality. "Superior" and "gentleman" refer to the concept of "junzi," "government" and "ruler" refer to the sovereign, and these words all appear frequently. This shows that AI translation can capture the core concepts and main ideas of *The Analects*. However, further analysis reveals differences in translation tendencies and vocabulary selection between the two. In the human translation, high-frequency words are closely related to the themes of the superior man and social governance, such as "propriety" and "virtuous." These words reflect the high emphasis on moral norms and ideal personality in *The Analects* and show the translator's profound understanding of the original text. In contrast, the high-frequency words in AI translation incorporate more modernized words, such as "gentleman," "benevolence," and "rites," which are highly abstract and general words that are easy to understand. These words are related to the core concepts of the text, but the choice of words is more inclined toward general expressions and fails to fully reflect the cultural context and philosophical depth of *The Analects*.

Legge's version	
master	524
virtue	107
superior	95
people	93
gong	79
confucius	74
replied	57
government	56
propriety	52
heard	41
prince	41
virtuous	39
love	36
zhong	35
rules	34
duke	33
xia	33
called	32
learning	32
conduct	31
practice	31
heaven	29
music	29
zhang	28
principles	27
day	26
perfect	26
qiu	24
disciples	21
learn	21

AI's version	
confucius	457
master	128
people	104
gentleman	99
replied	83
benevolence	64
person	57
gong	51
rites	51
virtue	47
benevolent	40
ruler	40
duke	38
love	33
zhong	32
zhang	31
understand	30
heard	29
called	28
xia	28
speak	27
yan	26
follow	24
petty	24
heaven	23
learning	23
minister	23
zigong	22
respect	22
qiu	20

Figure 1: Top-word Frequency Comparison

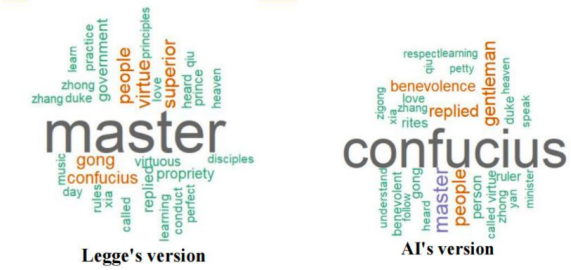


Figure 2: Word Cloud

4.2 Cosine Similarity

Next, we conducted semantic consistency analysis to statistically measure the degree of similarity between the two translations. We used two indicators: cosine similarity and Jaccard similarity. Cosine similarity is a similarity measurement tool for vectorized text that evaluates the consistency of text in its overall semantic structure. In our study, the cosine similarity between AI and human translation was 0.906, indicating high consistency in document-level lexical distribution and topical signal (TF–IDF vectors), rather than sentence-level structure. This means that AI translation can translate accurately in terms of sentence meaning, semantic logic, and content expression.

4.3 Jaccard Similarity

Jaccard similarity places more emphasis on the degree of overlap in vocabulary selection. However,

the Jaccard similarity was only 0.325, demonstrating significant differences in specific vocabulary and phrases between the two translations. It was observed that human translation tends to choose words that fit the context of ancient Chinese and pays more attention to the cultural connotations, such as explaining and interpreting the meaning of a core concept rather than summarizing it directly. For instance, "junzi" is rendered as "a man of complete virtue." AI translation tends to choose direct or generalized expressions, especially using modern or more universally applicable words to express a core concept with a single word. For instance, "junzi" is rendered as "gentleman." This is because AI translation lacks a deep understanding of the historical context and cultural background of the text, and its word choice tends to be more uniform and general. Human translation, on the other hand, can combine the context of the text and use flexible word choice and more detailed explanations to more accurately convey the core ideas of the text. Through a combined analysis of cosine similarity and Jaccard similarity, it can be seen that AI translation has good semantic consistency and can convey the logical structure and main content of *The Analects* to a large extent. However, its limitations in vocabulary selection are obvious, resulting in low word overlap.

4.4 TF–IDF Analysis

Through TF–IDF analysis, the differences in keyword extraction and the distribution of vocabulary weight between the two translations can be clearly identified. We compute document-level TF–IDF with scikit-learn TfidfVectorizer on a two-document corpus (human vs. AI), after lowercasing, punctuation and digit removal, and stopword filtering; weights are L2-normalized per document, and we report top-k terms (k = 100). TF–IDF is a method for reflecting the relative importance of a word by calculating its frequency of occurrence in a single document and its distribution in the entire corpus. The high TF–IDF words in the human translation include "master" (0.7897), "virtue" (0.1613), and "superior" (0.1432), which reflect the core ethical and philosophical concepts of Confucius in *The Analects*, such as the concept of the "junzi" and "morality." The high TF–IDF words in the AI translation include "Confucius" (0.737), "gentleman" (0.2425), and "benevolence" (0.1145). It can be seen that the key words in both translations are highly consistent in content theme, reflecting a focus on the core concepts of *The Analects*, but with different specific choices of words. Focusing on the weights, although the AI translation can capture the keywords and translate them in a way that is close to the human translation, its allocation of weight to these keywords is more balanced and even, lacking prominence and emphasis, indicating that it still has a problem in theme identification.

Legge's version		AI's version	
master	0.7897	confucius	0.737
zi	0.2758	zi	0.279
virtue	0.1613	gentleman	0.2425
superior	0.1432	master	0.2161
people	0.1402	people	0.1725
gong	0.1191	replied	0.1338
confucius	0.1115	benevolence	0.1145
lu	0.104	person	0.0968
replied	0.0859	lu	0.0951
government	0.0844	rites	0.0839
propriety	0.0784	gong	0.0822
rules	0.072	virtue	0.079
heard	0.0618	benevolent	0.0693
prince	0.0618	ruler	0.0693
virtuous	0.0588	duke	0.0613
perfect	0.0551	ji	0.0613
love	0.0543	if	0.0612
		petty	0.0544
		love	0.0532

Figure 3: TF-IDF

4.5 LDA Topic Modeling

LDA topic modeling reveals the difference in theme extraction. For LDA, we segment the translation into sentence-like units by punctuation and retain segments longer than 50 characters, yielding 907 document units (7,395 tokens) for the AI translation. We select the number of topics via c_v coherence under a lightweight screening configuration (fixed seed), and then retrain the final model with a stronger configuration. Under this setting, the selected number of topics for the human translation is two, and the themes are concentrated on two core areas of "social governance" (state, officer, minister) and "personal morality" (virtue, superior, love). The boundaries between these two themes are very clear, reflecting the translator's deep understanding of the main themes of *The Analects*. On the other hand, the themes extracted by AI translation are more dispersed, with a total of seven themes identified. Some of the topic words do not have clear practical meanings, and there are also problems of semantic overlap or blurred boundaries between the themes. This phenomenon of semantic and

theme fragmentation indicates that AI translation has weak theme coherence in extracting deep-level semantics. Therefore, AI translation lacks proficiency in grasping the overall semantic meaning, has limited understanding of the deep cultural connotations, and has a weaker ability to maintain semantic coherence when dealing with high-difficulty texts such as ancient Chinese literature.

Figure 4: LDA Topic Modeling

4.6 Semantic Networks

Semantic network analysis visualizes conceptual co-occurrence and helps examine community structure and the separation of conceptual clusters. Core concepts (such as "truth," "benevolence," "wisdom") play a central role in both translations, indicating that both of them can capture important semantic nodes in the text. The semantic network structure of the human-translated text is clearer, with nodes distributed more evenly, and core concepts such as "truthfulness" and "benevolence" being more independent from other nodes. The connections between positive and negative words are more distinct, presenting a clearer semantic logic that reflects the accurate grasp of ethical reasoning in *The Analects*. In contrast, the AI network appears denser with lower apparent modular separation among conceptual clusters, suggesting weaker separation of thematic communities. The boundaries between positive and negative emotional words are blurred, and the connections between core concepts are high in density but lack logical coherence. This demonstrates that AI translation tends to handle affect in a generalized manner and tends to summarize relationships between words, failing to fully reflect the clarity of ethical oppositions in *The Analects*.

4.7 Sentiment Analysis

Sentiment analysis results show that the indices are close between these two versions, but the values for AI are generally lower than the human version. AI translation exhibits a general phenomenon of "intensity decay," while maintaining the overall emotional distribution pattern. This "intensity decay" may stem from AI translation's tendency to opt for "safer" lexical and semantic choices. The intensity may be gradually reduced due to probability smoothing, which shows more "conservatism" and "uncertainty" in the translation. While this conservatism ensures stability and reliability, it also weakens the expression of emotional intensity and ethical opposition.

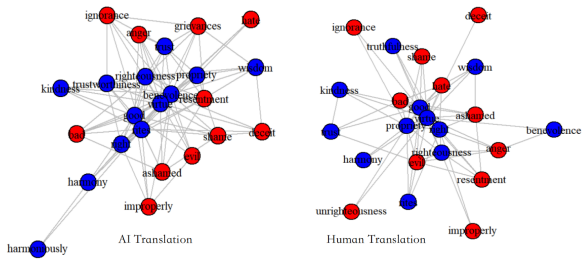


Figure 5: Semantic Network

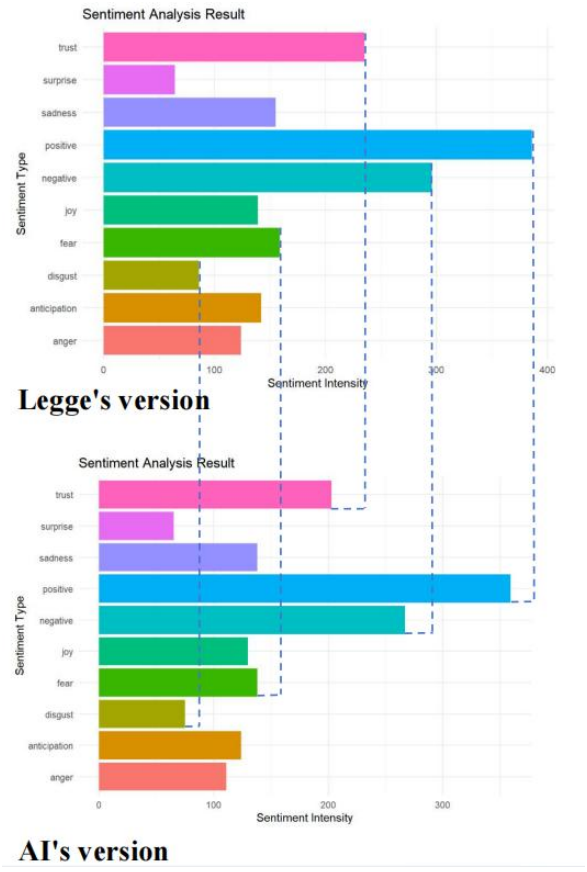


Figure 6: Sentiment Analysis

5 Targeted Case Analyses on Classical Chinese Phenomena

Case analysis primarily explores the potential details and problems that AI translation may encounter when processing *The Analects* text, specifically including loan characters (substitute characters with similar pronunciation), polysemy, special sentence structures, and value embodiment. This part is not suitable for quantitative research, so we adopt qualitative analysis. Loan Characters (substitute characters with similar pronunciation): When dealing with homophones, AI translation tends to give literal translations without fully understanding the context of the text. For example, "樊迟问知" (Fan Chi asks about what is wisdom), "知" (knowledge) is a homophone of "智" (wisdom). However, it is translated as "knowledge" by AI translation, while human translation would more accurately choose "wisdom" to convey the deep meaning in the ethical context. Characters with Different Meanings in Ancient and Modern Chinese For the translation of words with different meanings in ancient and modern times, AI translation performs well. For example, in the sentence "人皆有兄弟，我独亡。" (All men have brothers, but I have none), the word "亡" is translated accurately by AI as "have none," which means "without." Polysemy: One word with multiple meanings in classical Chinese is a difficult aspect in translation, such as "dao," which can represent "road," "principle," or "method" in different contexts. For example, in the sentence "君子所贵乎道者三," the human translator handled "dao" as "principles of conduct," accurately conveying its meaning, indicating the principles that the morally exemplary person should adhere to and keep doing, while the AI translation tends to translate it directly as "way," which cannot fully convey its deep meaning in the specific context. Special Sentence Structure: There are often elliptical sentences and inverted sentences in *The Analects*, which are not easy to understand or translate. For example, "君子上达，小人下达。" Human translators can complete the implicit elements to fully present the logical relationship of the sentence, while AI translation directly renders it as "Gentlemen reach high positions, while petty people reach low positions," translating it as a difference in social status rather than moral levels, thus failing to accurately convey the core ideology of Confucius. Values: The core theme of *The Analects* lies in the conveying of ethics and values, such as the phrase "有教无类," which conveys the idea that education can narrow the moral gap between different people. AI translation translates it as "Education should have no class distinction," which is a literal translation, assuming that people from any class—whether high or low—can receive equal education, which does not take into account the actual historical background at that time, when only nobles could receive a good education.

6 Limitations

Single reference translation: The study mainly relies on James Legge's English version. One reference can't cover the range of translator styles and term systems, which may shift keyword weights, topic distributions, and affect strength, and in turn change the robustness of the findings. Metrics focuses mainly on lexical and phrase distributions: The current framework doesn't fully capture sense allusion, rhetoric, and discourse structure. Future work will add consistency checks based on sentence and paragraph embeddings and run small expert evaluations with annotated samples. Corpus and genre scope: The analysis uses only the Analects, which is aphoristic in form. Other classics and genres may change topic granularity, discourse organization, and terminology expression. It can be extended across texts and genres to other ancient classics.

7 Conclusion

Our analyses indicate strong alignment in the overall topical signal between the AI translation and the human reference (document-level TF–IDF cosine = 0.906), yet systematic divergences in lexicalization and thematic organization remain: lexical overlap is limited (Jaccard = 0.325), keyword salience is flatter in the AI output, LDA reveals fragmented and less separable topics (H: K = 2 with coherent "social governance" and "personal morality"; M: K = 7 with overlap), semantic networks are denser and less modular, and affective intensity is attenuated relative to the human translation. Taken together, these findings suggest that contemporary systems can capture major themes of *The Analects* while diluting cultural anchoring and ethical nuance, addressing our research questions by demonstrating high distributional alignment alongside inaccuracies in concept realization, thematic cohesion, and affect. Methodologically, the study contributes a reproducible evaluation pipeline that inte-

grates distributional similarity, lexical overlap and keyness, topic modeling, semantic networks, sentiment profiling, and targeted diagnostics tailored to classical Chinese, underscoring that surface or distributional scores alone can overestimate adequacy for culturally dense material. Limitations include reliance on a single primary human reference, one retained AI system, English-side analyses, and lexicon-based sentiment. Future work will incorporate learned metrics, sense- and allusion-aware representations, graph-based coherence measures, expert human evaluation with multi-reference targets, and extensions to other pre-modern genres, with code and processed data released for reproducibility.

References

I. Al Sawi and R. Allam. 2024. Exploring challenges in audiovisual translation: A comparative analysis of human- and ai-generated arabic subtitles in Birdman. *PLOS ONE*, 19(10):e0311020.

M. Ding. 2024. Comparative analysis of classical chinese poetry translation using artificial intelligence: A case study of different english versions of li bai's "farewell to a friend". *Education Journal*.

James Legge. 1861. *The Chinese Classics: Confucian Analects*. Trübner & Co.

F. Massion. 2017. Artificial intelligence, smart assistants and the role of language professionals. *Lebende Sprachen*, 62(2).

D. Wang. 2008. Comparative study of english translations of the "analects". Master's thesis, Shandong University, Jinan, China.

A. Zaid and H. Bennoudi. 2023. Ai vs. human translators: Navigating the complex world of religious texts and cultural sensitivity. *International Journal of Linguistics, Literature and Translation*, 6(11):173–182.

From Span Extraction to Classification: A Multi-step Framework for Cognitive Distortion Analysis

Manh-Cuong Phan[1], **Thi-Ngoc-Phuong Nguyen**[1], **Huu-Loi Le**[2], **Huy-The Vu**[1], **Hajime Hotta**[3], and **Minh-Tien Nguyen**[1]*
[1] Hung Yen University of Technology and Education, Hung Yen, Vietnam.
cuongpm@spkt.edu.vn; {nguyenngocphuong, thevh, tiennm}@utehy.edu.vn
[2] AI Academy Vietnam, Vietnam.
loilh@aiacademy.edu.vn
[3] Hajime Institute, Kuala Lumpur, Malaysia.
hotta@hajime.institute

Abstract

Cognitive distortions (CDs) are biased thought patterns linked to conditions like depression and anxiety. Identifying CDs in therapy conversations is crucial for mental health support. Different from prior work that used rule-based and supervised methods but struggled with contextual understanding and data sparsity, we introduce a multi-step deep learning framework that first detects distortions, then leverages a machine reading comprehension module to extract distortion spans, and finally classifies their types. The task uses a machine reading comprehension module to extract distortion spans for noisy reduction, followed by a classifier to identify their types. Experimental results on a publicly available benchmark dataset, which has been widely adopted in prior studies, show that our framework achieves superior performance compared to strong baselines in distortion detection, span extraction, and classification.

1 Introduction

Cognitive distortions (CDs) refer to biased or irrational thought patterns that are often associated with various psychological disorders, such as depression, anxiety, and post-traumatic stress disorder (Beck, 2020). These distortions manifest in everyday conversations, e.g., patients and therapists, and can significantly hinder the therapeutic process (Beutel et al., 2019). Detecting and classifying cognitive distortions accurately within such interactions are crucial for multiple purposes, including practical applications such as enhancing therapeutic interventions (Chen et al., 2023b) and personalized mental health care (Shreevastava and Foltz, 2021), as well as research directions such as probing the reasoning ability of Large Language Models (LLMs) (Chen et al., 2023b; Wang et al., 2024; Lim et al., 2024).

*Corresponding Author.

Traditional methods for detecting cognitive distortions rely heavily on manual annotation and rule-based systems, which are limited in their scalability and ability to capture the subtle nuances of natural language (Shreevastava and Foltz, 2021). For example, Shreevastava and Foltz 2021 introduced a supervised learning framework for detecting CDs in patient-therapist interactions, leveraging feature engineering. Despite its success, this approach faces a major challenge relating to the use of the entire input (full speech text) that may add noisy information to CD models.

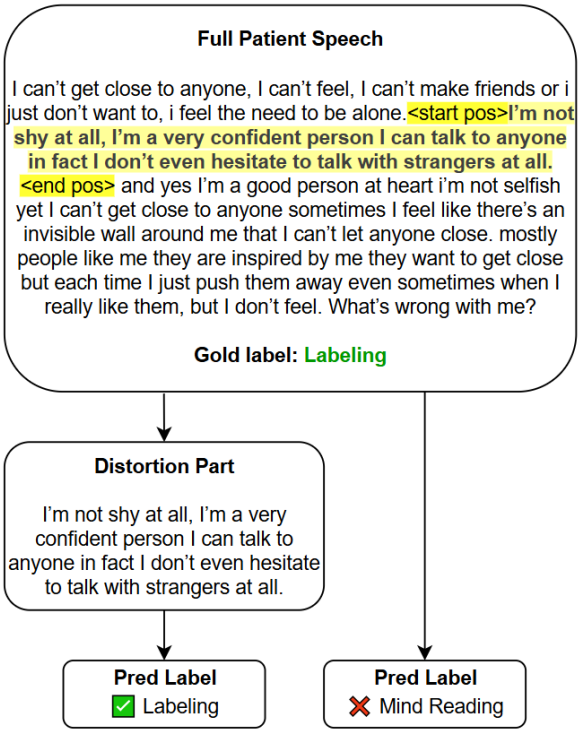


Figure 1: Example of cognitive distortion span extraction and classification. The full patient speech includes a distortion span (<start pos> and <end pos>). While our model correctly predicts the distortion type based on the span, Shreevastava and Foltz (2021) using the full speech incorrectly classifies it as "Mind Reading".

Let us take Figure 1 as an example. The patient's

speech contains a distortion span marked by start and end positions. When the model makes a prediction based on the full patient's speech, it incorrectly classifies the distortion as *Mind Reading* (Shreevastava and Foltz, 2021), likely due to the presence of unrelated or misleading context. However, when focusing solely on the distortion span, our model correctly identifies the distortion type as *Labeling*. This demonstrates the importance of isolating the distorted segment for accurate classification. Based on this observation, we argue that the performance of CD models can be improved by using distortion information extracted from the patient's speech.

This paper introduces a multi-step framework that takes advantage of span extraction for distortion detection. The framework consists of three main steps: (1) cognitive distortion detection, (2) distortion span extraction, and (3) distortion classification. The core hypothesis is that essential signals for distortion detection are often localized within a small portion of the input context. To address this, the framework first pinpoints these critical spans via a span extraction module, followed by a classification module that categorizes the type of distortions. The span extraction module uses pre-trained language models (PLMs) (e.g., BERT) to capture fine-grained semantic dependencies and contextual cues, enhancing detection performance. The framework is evaludated on the benchmark dataset introduced by Shreevastava and Foltz (2021), and experimental results demonstrate significant improvements over previous methods across multiple metrics. This paper makes two main contributions as follows.

- It introduces a framework for detecting and classifying cognitive distortions from patient-therapist interactions. The framework includes three steps: distortion detection, distortion span extraction, and final distortion classification in a single pipeline.
- It conducts a comprehensive evaluation on a publicly available dataset, demonstrating clear improvements across multiple metrics such as accuracy, precision, recall, and F1-score.

2 Related Work

Cognitive distortions are closely associated with mental health disorders such as depression and anxiety (Beck, 2020). The automatic detection and classification of such distortions has garnered increasing attention, particularly with advancements in natural language processing (NLP) and LLMs.

Shreevastava and Foltz (2021) introduced one of the first publicly available datasets focusing on cognitive distortions in patient speech. They annotated 2,531 utterances with both binary (distorted or not) and multi-class (distortion type) labels. Their semi-supervised approach leveraged labeled and unlabeled data but faced challenges in generalizability and interpretability due to limited context and data scale. To enhance interpretability, Chen et al. (2023c) proposed the Diagnosis of Thought (DoT) prompting framework, which utilizes LLMs such as ChatGPT and GPT-4. The DoT approach follows a structured reasoning process that consists of three steps: subjectivity assessment, contrastive reasoning, and schema analysis. Their results show that this method outperforms zero-shot baselines and provides clinically meaningful explanations, as confirmed by evaluations from licensed therapists.

Building on this foundation, Lim et al. (2024) introduced the Extraction, Reasoning, and Debate (ERD) framework, which employs multi-agent reasoning among LLMs to reduce diagnostic bias and improve performance in multi-class classification. Singh et al. (2024) investigated multimodal LLMs that integrate textual, auditory, and visual signals to improve zero-shot detection of distortions in patient–doctor interactions. Lin et al. (2024) developed a Mandarin-language dataset containing parallel reframing examples, enabling models to both detect distortions and suggest positive alternatives grounded in psychological theory.

In terms of scalability, Kim and Kim (2025) introduced KoACD, a large-scale dataset in Korean focusing on adolescent populations, and Babacan et al. (2025) leveraged GPT-4 to generate synthetic training data for cognitive distortion classification. While these datasets contribute to broader coverage across languages and domains, they are less applicable to our setting, which targets English patient–therapist interactions. Therefore, in this paper we evaluate our framework on the benchmark dataset introduced by Shreevastava and Foltz (2021).

While sharing the goal of cognitive distortion detection with Shreevastava and Foltz (2021) and Chen et al. (2023c), our work differs by adding a span extraction stage formulated as a machine reading comprehension (MRC) task. This step isolates distortion-relevant text, reducing noise and improving classification accuracy.

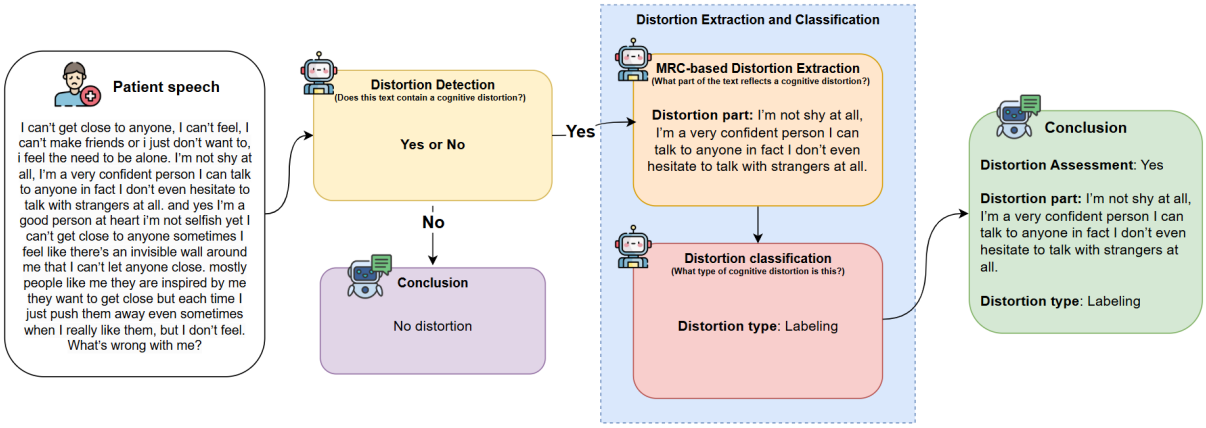


Figure 2: The proposed framework using machine learning techniques.

3 Methodology

3.1 Problem Statement

Cognitive distortions are irrational thought patterns that negatively impact mental health. Given an input text x (n tokens) (a conversation between a patient and a therapist), our goal is to (1) assess whether x contains any CD, (2) extract the span(s) responsible for the distortion, and (3) classify each extracted span into one of K predefined categories. We formalize the overall problem of cognitive distortion analysis in patient–therapist conversations as three learning subtasks: cognitive distortion detection (Section 3.3), distortion span extraction (Section 3.4), and distortion type classification (Section 3.5). Each task is trained independently with its objective function.

3.2 Overview of the Framework

Figure 2 shows the proposed framework for distortion detection and classification. The proposed framework consists of three primary components: distortion detection, the extraction of cognitive distortions using span extraction, and the classification of these distortions into specific types. The system processes patient–therapist interaction texts through a multi-stage pipeline designed to identify and classify cognitive distortions.

3.3 Cognitive Distortion Detection

Detection is performed first to filter out distortion-free utterances. This reduces noise and search space, making span extraction and type classification more accurate and efficient.

The first step detects whether each patient utterance is distortion-free or not. To do that, we formulate the detection as a binary classification problem that involves learning a classifier $f_d : \mathcal{X} \rightarrow \{0, 1\}$ which determines whether an utterance x_i contains a cognitive distortion. The prediction is denoted as:

$$y_i = f_d(x_i; \theta)$$

The detection is done in three steps: pre-processing, feature representation, and classification.

Pre-processing The dataset used for training the model is first preprocessed. This involves cleaning the text, tokenizing it into smaller units (tokens), and converting it into a format suitable for input into the MRC model. Text normalization techniques, such as removing stop words and punctuation, are applied to improve the quality of the input data. We employ the `nltk` and `re` libraries for basic text preprocessing, such as lowercasing, punctuation removal, and stop-word filtering. For PLM-based models (e.g., BERT, RoBERTa, DeBERTa), we apply only minimal preprocessing. Specifically, we rely on the model-specific subword tokenization and encoding provided by the Hugging Face `AutoTokenizer`[1], with lowercasing applied only when using uncased variants. We avoid aggressive transformations such as stop-word removal or punctuation stripping, since these may discard semantically informative tokens and harm PLM performance.

Class conversion The original dataset contains multiple classes representing different types of cognitive distortions, such as *Labeling*, *Mind Reading*, *Catastrophizing*, etc. To make the dataset suitable for the task of **distortion detection**, we convert these original classes into two binary categories:

[1] https://huggingface.co/docs/transformers

Distortion and **No-distortion**. All samples that contain any specific type of cognitive distortion are grouped under the *Distortion* class, while the rest (including normal or undefined responses) are labeled as *No-distortion*. This conversion enables the use of binary classification models while preserving the key distinction between distorted and non-distorted content. Tables 1 and 2 shows the statistics of labels of original and converted classes.

Table 1: Original class distribution.

Original Class	Sample Count
No-distortion	933
Mind Reading	239
Overgeneralization	239
Magnification	195
Labeling	165
Personalization	153
Fortune-telling	143
Emotional Reasoning	134
Mental filter	122
Should statements	107
All-or-nothing thinking	100

Table 2: Binary class distribution.

Binary Class	Sample Count
Distortion	1597
No-distortion	933

Feature representation. In this work, we adopt three different methods to represent an input conversation x as contextual vectors. The first method is Bag-of-Word (BoW) that creates a dictionary on the whole corpus and then maps each input x_i to a fixed-size vector (Joachims, 1998). The second method is TF-IDF that focuses more on the importance of words in the corpus (Ramos, 2003).

The third feature representation method leverages PLMs such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or DeBERTa (He et al., 2023).

This is because these PLMs were trained with a huge amount of data to capture contextual information of input tokens. Given an input x_i, the framework adds a new [CLS] token to x_i to form a new sequence: [CLS] x_i. This sequence is fed into PLMs, and we take the hidden representation of the [CLS] token from the final encoder layer to form $\mathbf{H}$. The vector $\mathbf{H}$ is then passed to a classification layer for prediction.

Classification Given contextual vectors from feature representation, classification uses two types of methods: traditional and PLMs. Traditional methods use BoW and TF-IDF features, and PLMs use the hidden vectors $\mathbf{H}$ for classification. Results are shown in Tables 4 and 5.

3.4 Distortion Span Extraction

Span extraction aims to identify one or more spans within the input x_i that correspond to distorted expressions. Each span is defined as $s_i = (\text{start}_i, \text{end}_i) \subset x_i$, and the set of all extracted spans (m spans) is represented as follows.

$$P = f_{se}(x; \theta), \quad \text{where } P = \{p_1, p_2, \ldots, p_m\}.$$

We observed that distortion cues are localized to small segments rather than the entire utterance (Shreevastava and Foltz, 2021). Different from prior work that fed the whole utterance x_i for the final classification, we argue that using the full utterance may introduce noise for classifiers. To mitigate this, the framework extracts distortion spans from each utterance x_i for the final classification.

The extraction is formulated as a MRC (Machine Reading Comprehension) problem due to two reasons. First, MRC models can be utilized to pinpoint specific segments within a patient's narrative that indicate distorted thinking. By framing the detection task as a question-answering problem, the model can be prompted with questions such as, "What part of the text reflects a cognitive distortion?" This formulation allows the framework to focus on extracting evidence-based segments that signify distorted thoughts (Nguyen et al., 2023; Chen et al., 2023a). Second, recent studies have shown the efficacy of MRC frameworks in clinical concept extraction, highlighting their potential in identifying nuanced psychological patterns, including cognitive distortions (Chen et al., 2023a).

Given an input utterance $x_i = \{w_1, w_2, ..., w_n\}$ consisting of n tokens, we obtain its contextualized representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ from the PLM encoder. The span extraction module then predicts a set of distortion spans $P = \{p_1, p_2, ..., p_m\}$, where each span s_i is represented by its start and end positions ($S, E \in \mathbb{R}^c$) following the standard BERT QA formulation (Devlin et al., 2019).

The prediction uses a softmax over all hidden features $\mathbf{H}$ to predict the start or end positions of a

token j^{th} for the answer span p_i as follows.

$$p_{si}^{j} = \frac{\exp(h_i^{j\top} S)}{\sum_{i'} \exp(h_{i'}^{j\top} S)}; \quad p_{ei}^{j} = \frac{\exp(h_i^{j\top} E)}{\sum_{i'} \exp(h_{i'}^{j\top} E)}$$

The start and end positions of answer span p_i were calculated as follows.

$$s^i = argmax_i(p_{si}^j); \qquad e^i = argmax_i(p_{ei}^j)$$

For training, we utilize annotated datasets in which each distortion instance is labeled with its corresponding text span. These gold spans serve as supervision signals for the model to learn start and end token positions using cross-entropy loss. Table 6 shows results of the extraction.

3.5 Cognitive Distortion Classification

The final step is to predict distortion types of each input utterance x_i using the extracted spans in Section 3.4. For each patient utterance x_i, the span s_i identified in the previous step is classified by a model f_{cl} into one of K distortion categories:

$$c_i = f_{cl}(s_i; \theta), \quad c_i \in \{1, 2, \ldots, K\}.$$

Followed by Section 3.3, the framework uses three types of feature representation: BoW, TF-IDF, and contextual vectors from PLMs (BERT, RoBERTa and DeBERTa). The final classification also follows methods in Section 3.3.

To improve classification accuracy, we fine-tune pre-trained language models (BERT, RoBERTa, DeBERTa) on labeled spans. Each span s_i is passed through a classification head (a linear layer and softmax) to predict distortion types. The models are optimized using cross-entropy loss and AdamW optimizer (Loshchilov and Hutter, 2019), with early stopping applied based on validation performance.

4 Experimental Settings

4.1 Dataset

Experiments were conducted on the annotated dataset introduced by Shreevastava and Foltz (2021). It contains 2,531 utterances extracted from real-world patient-therapist sessions. Each utterance is annotated with binary labels (distortion or non-distortion). If distorted, it is further classified into one of ten predefined cognitive distortion types (Table 3). The dataset also includes start and end positions of distortion spans. It was split into training and test sets using an 80/20 ratio, ensuring consistent distribution across distortion categories.

4.2 Settings

Our proposed framework consists of three core components, each trained under the following settings.

Distortion detection. We treat distortion assessment as a binary classification task. Following Shreevastava and Foltz (2021), we experiment with five machine learning models: **Logistic Regression** (Hosmer et al., 2013), **Support Vector Machines (SVM)** (Cortes and Vapnik, 1995), **Decision Tree** (Safavian and Landgrebe, 1991), **k-NN** (Cover and Hart, 1967), and **MLP** (Rumelhart et al., 1986). Each classifier is trained with two types of feature representations: Bag-of-Words (BoW) and TF-IDF. In addition, we fine-tune **BERT**, **RoBERTa**, and DeBERTa (Devlin et al., 2019; Liu et al., 2019; He et al., 2023) for binary classification using Hugging Face's `Transformers` library (Wolf et al., 2020). For fine-tuned PLMs, we use a batch size of 16, learning rate of 2e-5, AdamW optimizer, and train for 5 epochs.

Distortion span extraction. This task is formulated as a span-based question answering problem, following the MRC paradigm. Given an utterance and a query such as "Which part of the text reflects a cognitive distortion?", the model outputs the relevant span. We fine-tune **BERT**, **RoBERTa**, and **DeBERTa** using the Hugging Face `QA pipeline`[2]. Training is performed with a batch size of 12, learning rate of 3e-5, AdamW optimizer, and 3 epochs.

Cognitive distortion type classification. After extracting distortion spans, the framework classifies them into specific distortion categories. We employ the same feature representations and classifiers as distortion detection. For fine-tuned PLMs, we use a batch size of 16, learning rate of 2e-5, and 5 epochs.

4.3 Evaluation Metrics

Each subtask is evaluated using task-appropriate metrics as follows. **Distortion detection** uses F1 score for the positive class (i.e., distorted utterances) (Shreevastava and Foltz, 2021). **Span extraction** uses Exact Match (EM) and token-level F1 score, following common practice in span-based QA tasks (Rajpurkar et al., 2016). **Cognitive distortion classification** uses accuracy and weighted

[2] https://huggingface.co/docs/transformers/en/main_classes/pipelines#transformers.QuestionAnsweringPipeline

Table 3: Common cognitive distortion types and example speech (Beck, 2020; Shreevastava and Foltz, 2021).

Cognitive Distortion Type	Interpretation	Example Distorted Speech
Personalization	Personalizing or taking up the blame for a situation that involved many factors outside the person's control.	My son is pretty quiet today. I wonder what I did to upset him.
Mind Reading	Suspecting what others are thinking or the motivations behind their actions.	My house was dirty when my friends came over; they must think I'm a slob!
Overgeneralization	Drawing major conclusions based on limited information.	Last time I was in the pool I almost drowned; I am a terrible swimmer and should not go into the water again.
All-or-nothing thinking	Seeing a situation as either black or white with no middle ground.	If I cannot get my Ph.D., then I am a total failure.
Emotional reasoning	Letting feelings override factual evidence.	Even though Steve is here at work late every day, I know I work harder than anyone else at my job.
Labeling	Assigning a fixed label to oneself or others without deeper examination.	My daughter would never do anything I disapproved of.
Magnification	Emphasizing the negative or minimizing the positive aspects of a situation.	My professor said he made some corrections on my paper, so I know I'll probably fail the class.
Mental filter	Focusing only on the negatives of a situation.	My husband says he wishes I was better at housekeeping, so I must be a lousy wife.
Should statements	Creating rigid rules about how one or others should behave.	I should get all A's to be a good student.
Fortune-telling	Predicting that things will turn out badly without evidence.	I was afraid of job interviews so I decided to start my own thing.

F1 score to account for class imbalance across ten categories.

5 Results and Discussion

5.1 Performance Comparison

This section reports the comparison of the proposed framework to baselines for three problems: distortion detection, span extraction, and distortion classification. All reported metrics are computed on the held-out test set.

5.1.1 Distortion detection

Performance with traditional methods To evaluate the effectiveness of different feature representations for cognitive distortion detection, we experimented with various linguistic and semantic feature sets. These include Sentence Embeddings using SIF (Arora et al., 2017), BERT-based embeddings (Reimers and Gurevych, 2019), psycholinguistic features from LIWC (Pennebaker et al., 2001), Part-of-Speech (POS) tags (Toutanova et al., 2003), and combinations thereof. The results of these representations are derived from original papers (Shreevastava and Foltz, 2021) for fair comparison. Additionally, we compared these against our features: BoW and TF-IDF (Section 3.3).

As shown in Table 4, simple lexical representations such as BoW and TF-IDF consistently outperform more complex embedding-based features in this task. TF-IDF achieves the strongest overall results across most classifiers, while BoW also performs competitively, particularly with certain tree-based methods. Among the learned embedding approaches, combinations incorporating psycholinguistic features (e.g., LIWC) yield moderate improvements, highlighting that such features still hold value for cognitive distortion assessment.

Table 4: Evaluation of traditional methods. LR: Logistic Regression, DT: Decision Tree. Reported metric is F1. † Results copied from Shreevastava and Foltz (2021).

Feature	LR	SVM	DT	k-NN	MLP
SIF †	0.75	0.77	0.65	0.74	0.73
BERT †	0.74	0.79	0.67	0.75	0.70
LIWC †	0.77	0.78	0.67	0.76	0.77
POS †	0.73	0.77	0.66	0.75	0.72
BERT+LIWC †	0.74	0.76	0.64	0.75	0.74
BoW (our)	0.77	0.80	**0.71**	0.71	0.74
TF-IDF (our)	**0.81**	**0.81**	0.70	**0.79**	**0.78**

These findings suggest that, for this domain, sparse and interpretable lexical features can be more effective than dense contextual embeddings, possibly due to the distinct and recurring linguistic markers associated with distorted thinking.

Comparison of PLMs Table 5 shows the detection results using BERT, RoBERTa and DeBERTa. Compared to traditional methods in Table 4, the detection using PLMs obtains better performance. This is because PLMs were trained with a huge amount of data. When fine-tuning for downstream tasks, they usually produce better accuracy than traditional methods, which have much smaller model sizes.

Table 5: Distortion detection with PLMs.

Models	Precision	Recall	F1
BERT-base	0.76	0.89	0.82
BERT-large	0.76	0.91	0.83
RoBERTa-base	0.76	**0.93**	**0.84**
RoBERTa-large	0.77	**0.93**	**0.84**
DeBERTa-v3-base	**0.78**	0.91	**0.84**
DeBERTa-v3-large	**0.78**	0.91	**0.84**

5.1.2 Span extraction

Table 6 shows that model capacity and architectural design both play an important role in distortion span extraction. Larger variants of BERT and RoBERTa consistently outperform their base counterparts, suggesting that increased parameterization enhances the ability to capture fine-grained cues. Moreover, DeBERTa-v3 achieves the strongest overall performance, indicating that architectural refinements such as disentangled attention further improve span identification beyond mere scaling.

Table 6: Evaluation of distortion extraction.

Models	EM	F1
BERT-base	72.92	84.11
BERT-large	77.27	87.46
RoBERTa-base	75.89	85.23
RoBERTa-large	77.08	85.77
DeBERTa-v3-base	78.26	88.05
DeBERTa-v3-large	**78.54**	**88.47**

These findings support the feasibility of formulating cognitive distortion extraction as a machine reading comprehension problem and demonstrate that contextualized language models are effective in capturing fine-grained psychological patterns.

5.1.3 Cognitive distortion classification

This section shows the performance of distortion classification in two settings: using two steps (span extraction and classification) and the full pipeline. This design isolates classifier performance from upstream errors and reveals how detection mistakes propagate to final classification.

Step-wise evaluation with traditional methods The classification task was performed on two settings: the full speech text and the extracted distorted spans (distorted parts). The full speech text uses the whole utterance for classification. The extracted distorted spans use extracted spans from Section 3.4 for classification. We also include gold labels to observe the gaps between extracted spans and gold-labeled spans.

Table 7 shows consistent trends across traditional and transformer-based models. Using entire utterances leads to lower performance due to noise and irrelevant content, whereas gold-standard spans yield clear improvements. Automatically extracted spans perform close to the gold setting, indicating that the extraction step effectively reduces noise while retaining task-relevant information.

Table 7: Results for cognitive distortion classification with step-wise and full pipeline. † Results copied from Shreevastava and Foltz (2021).

Speech Part	Methods	BoW		TF-IDF	
		Acc	F1	Acc	F1
Full speech	K-NN †	–	–	–	0.24
	Logistic Reg.	0.23	0.23	0.24	0.19
	SVM	0.19	0.19	0.28	0.24
	Decision Tree	0.13	0.13	0.16	0.16
	K-NN	0.16	0.16	0.19	0.20
	MLP	0.23	0.23	0.25	0.24
		Acc		F1	
	BERT-base	0.27		0.25	
	BERT-large	0.29		0.28	
	RoBERTa-base	0.28		0.28	
	RoBERTa-large	0.30		0.33	
	DeBERTa-base	0.25		0.25	
	DeBERTa-large	0.27		0.28	
Distortion span extraction (gold)	Logistic Reg.	0.32	0.32	0.35	0.31
	SVM	0.32	0.32	0.35	0.35
	Decision Tree	0.20	0.20	0.17	0.18
	K-NN	0.16	0.15	0.20	0.20
	MLP	0.32	0.32	0.32	0.32
		Acc		F1	
	BERT-base	0.41		0.42	
	BERT-large	0.47		0.47	
	RoBERTa-base	0.43		0.44	
	RoBERTa-large	0.47		0.48	
	DeBERTa-base	0.43		0.43	
	DeBERTa-large	0.45		0.47	
Distortion span extraction (ours)	Logistic Reg.	0.31	0.31	0.32	0.28
	SVM	0.31	0.31	0.31	0.30
	Decision Tree	0.18	0.19	0.19	0.19
	K-NN	0.16	0.15	0.23	0.23
	MLP	0.30	0.29	0.29	0.28
		Acc		F1	
	BERT-base	0.38		0.39	
	BERT-large	0.41		0.42	
	RoBERTa-base	0.43		0.44	
	RoBERTa-large	**0.45**		**0.46**	
	DeBERTa-base	0.42		0.42	
	DeBERTa-large	0.41		0.43	
Full pipeline	Logistic Reg.	0.28	0.29	0.29	0.28
	SVM	0.25	0.24	0.26	0.23
	Decision Tree	0.14	0.17	0.16	0.19
	K-NN	0.11	0.13	0.21	0.21
	MLP	0.25	0.28	0.26	0.28
		Acc		F1	
	BERT-base	0.33		0.36	
	BERT-large	0.36		0.40	
	RoBERTa-base	0.36		0.39	
	RoBERTa-large	0.38		0.42	
	DeBERTa-base	0.38		0.41	
	DeBERTa-large	0.40		0.42	

Evaluation with PLMs For transformer-based models, a similar trend is observed. Full-speech inputs result in the lowest performance, gold-standard spans lead to the largest gains, and automatically extracted spans maintain performance levels close to the gold spans. These results reinforce the hypothesis that focusing on distorted segments is an effective strategy for improving cognitive distortion classification.

Full pipeline evaluation We further evaluate the full pipeline performance, where span extraction and classification are executed in sequence. The lower part of Table 7 shows that transformer-based models still outperform traditional ones. Compared to classification on gold spans, these results of the full pipeline show competitive performance, which is still better than directly using full text. The scores of traditional classifiers using the full pipeline are better than those of using full speech. This confirms the contribution of distortion span extraction. However, the performance of methods using distortion span extraction is sill better than that of using full pipeline due to error accumulation.

5.1.4 Discussion with LLM-based models

Here we compare our framework, which relies on relatively small models, with LLMs that contain orders of magnitude more parameters.

Table 8: LLM-based models results. ($\star$) results copied from Chen et al. (2023c). Numbers in subscript denote the standard deviation over five runs.

Methods	Distortion Detection (F1)	Distortion Classification (Weighted F1)
Full training*	75.00	24.00
Vicuna*	$73.81_{0.95}$	$11.23_{0.78}$
ChatGPT*	$73.47_{0.58}$	$19.24_{1.00}$
ChatGPT + ZCoT*	$77.10_{1.21}$	$20.21_{1.02}$
ChatGPT + DoT*	$81.19_{0.11}$	$22.25_{0.70}$
GPT-4*	$83.04_{0.51}$	$33.86_{0.83}$
GPT-4 + ZCoT*	$81.97_{1.21}$	$33.22_{1.36}$
GPT-4 + DoT*	$82.77_{0.81}$	$34.64_{1.40}$
RoBERTa-base (our)	**84.00**	**44.00**

The higher scores of our method in Table 8 do not imply a direct superiority over the GPT-based approaches, but rather reflect the advantage of task-specific fine-tuning. This is because while the GPT-based methods in Table 8 operate in a zero-shot setting, our RoBERTa-base model is fine-tuned on the target dataset. This table is included to provide an additional observation on the performance gap between large, general-purpose LLMs without training on domain data and smaller, fine-tuned transformer models.

5.2 Error Analysis

Table 9 presents two representative cases. In Case 1, the full-speech model is misled by **mind-reading cues** ("people like me, they are inspired by me"), while our span-based model focuses on **self-labeling expressions** ("i'm ... a very confident person", "i'm not shy at all"), leading to the correct prediction of `Labeling`. In Case 2, the utterance contains **threat amplification and consequence escalation** ("every little thing ... terrified", "physically can't sleep", "so afraid"), which indicate `Magnification`. However, the full-speech model is influenced by **generalization cues**, predicting `Overgeneralization`, and our span-based model is biased by **first-person pronouns** and self-reference, predicting `Personalization`.

Table 9: Examples of two representative cases in error analysis.

Case 1: Full-speech misclassified, span prediction correct	**Case 2: Both full-speech and span prediction misclassified**
Patient speech (shortened): "... mostly people like me, they are inspired by me ... but each time I just push them away ..."	*Patient speech (shortened):* "... every little thing is making me terrified ... unless I take benedryl I physically can't sleep ... minor hallucinations ..."
Extracted span: "i'm not shy at all ...i'm a very confident person ... i don't even hesitate to talk with strangers ..."	*Extracted span:* "I see the light flickering ... someone's there ... every little thing is making me terrified ... I physically can't sleep ..."
Gold: `Labeling`	*Gold:* `Magnification`
Full-speech prediction: `Mind Reading`	*Full-speech prediction:* `Overgeneralization`
Extracted span prediction: `Labeling`	*Extracted span prediction:* `Personalization`

6 Conclusions

This paper presented a multi-step framework for cognitive distortion analysis in patient–therapist dialogues. Unlike prior studies that operate on full utterances, the proposed method explicitly models distortions at the span level by formulating span extraction as a machine reading comprehension task. The framework first detects whether a cognitive dis-

tortion is present, extracts the corresponding text spans, and then classifies them into fine-grained distortion categories. By isolating distortion-relevant spans, the approach reduces irrelevant context and enhances both interpretability and classification accuracy.

Comprehensive experiments on a benchmark dataset demonstrated consistent improvements over strong baselines, including traditional feature-based classifiers and pre-trained language models applied directly to full utterances. Span-level modeling was shown to yield notable gains in multi-class classification, with transformer-based models such as RoBERTa and DeBERTa achieving state-of-the-art performance. These findings indicate that sub-utterance representations are more effective than coarse-grained utterance-level inputs, particularly in domains requiring nuanced semantic distinctions.

The study also contributes to the growing body of computational methods for supporting cognitive-behavioral therapy (CBT). By aligning the modeling pipeline with the way distortions are annotated in clinical practice, the proposed framework provides results that are both more accurate and more interpretable, making it a promising step toward real-world mental health applications.

Future research directions include extending the framework to multi-span extraction and label-aware span selection, validating the approach on multilingual and cross-domain datasets such as KoACD and synthetic corpora, and incorporating multimodal signals (e.g., prosody, facial expressions, dialogue context) to enhance robustness in naturalistic settings.

Overall, the findings establish span-based cognitive distortion detection as a promising research direction, bridging the gap between automated NLP techniques and clinically meaningful psychological constructs. This framework lays the foundation for more accurate, interpretable, and clinically applicable systems in computational mental health.

Limitations

While the proposed framework performs well on the benchmark dataset, it relies on supervised span annotations, which may be unavailable in some domains. Evaluation is limited to English patient–therapist interactions, leaving its applicability to other languages, cultures, and settings untested. The current approach also focuses solely on textual data, excluding multimodal cues such as prosody or facial expressions.

Ethics Statement

This work uses publicly available, anonymized datasets for research purposes. The system is intended to support, not replace, mental health professionals, and should not be used as a stand-alone diagnostic tool. Potential biases in predictions must be monitored to avoid misclassification and possible harm in clinical contexts.

Acknowledgements

The authors thank the anonymous reviewers for their valuable feedback and suggestions. Support from colleagues and institutions during this research is gratefully acknowledged.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations (ICLR)*.

Erdem Babacan, Melis Kaya, and Tolga Ozdemir. 2025. Synthetic cognitive distortion data generation with gpt-4 for safer model training. *Fırat University Medical Bulletin*, 30(1).

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

M. E. Beutel et al. 2019. Cognitive distortions and their role in the development of mental disorders. *European Psychiatry*, 56:48–54.

Danqi Chen et al. 2023a. Clinical concept extraction with machine reading comprehension models. *Journal of Artificial Intelligence in Medicine*, 118:103329.

Zhiyu Chen, Yujie Lu, and William Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Zhiyu Chen, Yujie Lu, and William Wang. 2023c. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Jiwon Kim and Seunghyun Kim. 2025. Koacd: A large-scale korean dataset for adolescent cognitive distortion detection. arXiv preprint arXiv:2505.00367.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. Erd: A framework for improving llm reasoning for cognitive distortion classification. *arXiv preprint arXiv:2403.14255*.

Xiaohui Lin, Jun Zhou, and Yuwei Wang. 2024. Positive reframing of distorted thoughts: A mandarin dataset and llm evaluation. In *Findings of ACL 2024*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1833–1844.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Minh-Tien Nguyen, Nguyen Hong Son, et al. 2023. Gain more with less: Extracting information from business documents with small data. *Expert Systems with Applications*, 215:119274.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

S.R. Safavian and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Arjun Singh, Neha Patel, and Rui Zhang. 2024. Multimodal detection of cognitive distortions in patient–doctor conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*, pages 173–180.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. Patient-: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing.

Is CCGbank Semantically Valid? Insights from Negation Scope Analysis

Kentaro Kojima[1], **Yoshihide Kato**[2], **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University
[2]Information & Communications, Nagoya University
kojima.kentaro.a0@s.mail.nagoya-u.ac.jp

Abstract

Combinatory Categorial Grammar (CCG) is a grammatical theory that has been widely used in various semantic analyses. To analyze sentences with CCG, it is essential to construct their derivation trees using a CCG parser. Since many of these parsers are typically trained on CCGbank, evaluating the validity of CCGbank is crucial for ensuring the accuracy of the resulting semantic analyses. In this study, we investigate the validity of CCGbank from a semantic perspective, focusing specifically on the negation scope. Our investigation is based on the assumption that if CCGbank is semantically valid, it must correctly capture negation scopes. We conducted experiments comparing the negation scopes derived from CCGbank with those used in a negation scope resolution task, and confirmed that the scope of the quantifier "no" does not align well. The experimental results show that CCGbank does not capture the semantics of quantifiers correctly.

1 Introduction

Combinatory Categorial Grammar (CCG) (Steedman, 2000) is a grammatical theory that establishes a one-to-one correspondence between syntactic and semantic composition. It has been widely used in various semantic analyses (Mineshima et al., 2015; Abzianidze, 2015; Martínez-Gómez et al., 2017; Beschke and Menzel, 2018). To analyze sentences with CCG, it is essential to construct their derivation trees using a CCG parser, and several such parsers have been made publicly available (Clark and Curran, 2004; Lewis et al., 2016; Yoshikawa et al., 2017; Yamaki et al., 2023) . These parsers are typically trained on the CCGbank (Hockenmaier and Steedman, 2007), and consequently, they inherit and reproduce its characteristics.

However, CCGbank has been criticized for being semantically invalid in certain aspects (Boxwell and White, 2008). A key reason for its invalidity is that it is automatically generated from the Penn Treebank (Marcus et al., 1993). To mitigate the issues, some researchers have modified derivation trees to ensure semantic validity, particularly in the interpretation of noun phrases (Honnibal et al., 2010). As an alternative approach, other studies (Hu and Moss, 2018; Hu et al., 2019) have introduced extra semantic rules, which undermine the advantage of CCG having a transparent relation between syntax and semantics.

The aim of this study is to examine CCGbank from a semantic perspective that has not been previously explored: namely negation scope. Our investigation is based on the following assumption:

- If CCGbank is semantically valid, then it must capture negation scopes correctly.

We conducted experiments comparing the negation scopes derived from CCGbank's derivation trees with those used in a negation scope resolution task. The results show that while the scope of "not" generally aligns well, the scope of the quantifier "no" does not. These findings suggest that CCGbank lacks semantic validity in its annotation of the quantifier "no".

The main contributions of this study are summarized as follows:

- We demonstrate that CCGbank is semantically invalid from the negation scope perspective.
- We conduct a linguistic analysis of its causes and find that it lies not in CCG but in the CCGbank.

The remainder of this paper is organized as follows. Section 2 presents the preliminary definitions necessary for understanding this study. Section 3 outlines the methodology for evaluating the validity of CCGbank. Section 4 reports the experiment for evaluating the validity of CCGbank. Section 5 conducts an additional experiment. Finally, Section 6 presents the conclusion.

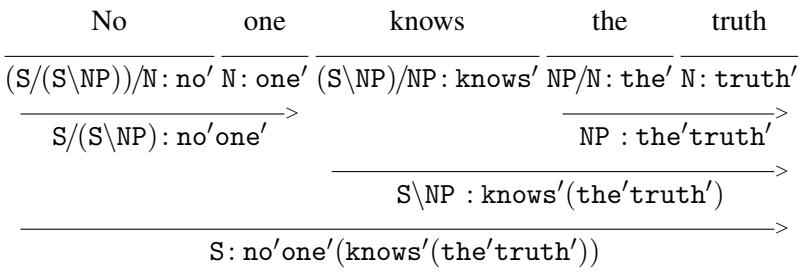


Figure 1: CCG derivation tree of sentence (1)

2 Preliminaries

This section presents preliminary definitions and essential background for understanding this study.

2.1 Negation Scope

Negation is an important phenomenon that frequently appears in natural language. Negation is caused by *negation cue*, such as prefixes (e.g. im-, un-), single words (e.g. not, no), or multiple words (e.g. no more than) and a *negation scope* is a part of a sentence affected by it. For example, in the following sentence (1), "No" functions as the negation cue, whereas "one knows the truth" constitutes the negation scope:

(1) **No** one knows the truth.

Additionally, in the following sentence (2), "not" functions as the negation cue, whereas "I am" and "a student" constitutes the negation scope:

(2) I am **not** a student.

In what follows, negation cues will be indicated in bold, and negation scopes will be marked with underlining. Several datasets annotated with negation cues and scopes have been released, including BioScope (Szarvas et al., 2008) for biomedical texts, SFU Review Corpus (Konstantinova et al., 2012) for product reviews, and the ConanDoyle-neg (Morante and Daelemans, 2012) based on Conan-Doyle's novels. The ConanDoyle-neg is well-known for its ability to capture more complex linguistic phenomena, such as long-range dependencies and discontinuous scope, compared to the other two datasets (Fancellu et al., 2017). Therefore, we adopt the definition of negation scope of the ConanDoyle-neg.

2.2 CCG

CCG is a lexicalized grammatical theory in which each word is assigned a syntactic category. These categories are classified into two types: basic categories (e.g., S for sentence, NP for noun phrase) and complex categories, which are formed by combining categories using the operators / and \. A category of the form X/Y indicates that it expects an expression of category Y to its right in order to form an expression of category X, whereas $X\backslash Y$ expects Y to its left.

In CCG, combining syntactic categories corresponds to one-to-one with that of semantic representations, which can be formalized using lambda calculus. Figure 1 shows the CCG derivation tree for the sentence (1). The syntactic combination illustrated in Figure 1 is based on function application, whose semantic representations are obtained as follows where f and a are λ-terms:

- $X/Y{:}f \quad Y{:}a \Rightarrow X{:}fa$
- $Y{:}a \quad X\backslash Y{:}f \Rightarrow X{:}fa$

CCG also includes another rules[1]:

generalized function composition

- $X/Y{:}f\ Y|_1Z_1\cdots|_dZ_d{:}g \Rightarrow$
  $X|_1Z_1\cdots|_dZ_d{:}\lambda z_d\cdots z_1.f(gz_d\cdots z_1)$
- $Y|_1Z_1\cdots|_dZ_d{:}g\ X\backslash Y{:}f \Rightarrow$
  $X|_1Z_1\cdots|_dZ_d{:}\lambda z_d\cdots z_1.f(gz_d\cdots z_1)$

type raising

- $X{:}a \Rightarrow T/(T\backslash X){:}\lambda f.fa$
- $X{:}a \Rightarrow T\backslash(T/X){:}\lambda f.fa$

Applying the above rules, the CCG derivation tree for sentence (2) is shown in Figure 2.

Here, the notation used in the following sections is defined as follows. For an expression of the form $Y|_1Z_1\cdots|_dZ_d$, the sequence $|_1Z_1\cdots|_dZ_d$ is referred to as the argument stack and is denoted

[1]Here, $|_i \in \{/, \backslash\}$ and Z_i is a category $(1 \leq i \leq d)$.

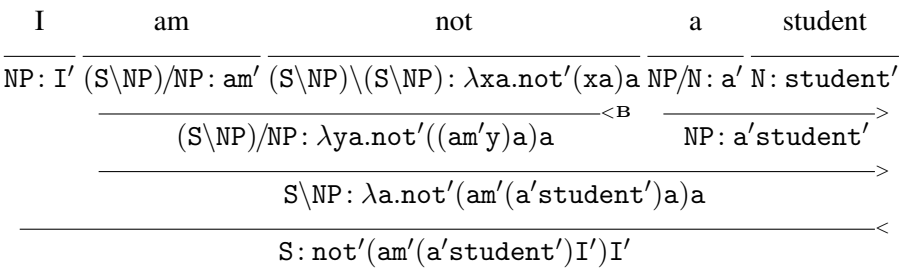


Figure 2: CCG derivation tree of sentence (2)

by a Greek letter, typically α. Specifically, we write $Y|_1 Z_1 \cdots |_d Z_d = Y\alpha$ and define $|\alpha| = d$. When a category X is expressed in the form $X = Y\alpha$, where Y is a basic category, the arity of X is defined as follows:

$$arity(X) = |\alpha|$$

3 Negation Scope on CCGbank

This section outlines the methodology for evaluating the semantic validity of CCGbank. We base our analysis on the following assumption:

- If CCGbank provides semantically valid derivation trees, then the negation scopes derived from them should correspond to those annotated in a general corpus (as described in Section 2.1).

In our analysis, we derive the negation scope from CCGbank as follows[2]:

1. Assign a λ-term to each leaf node in the derivation tree
2. Construct a λ-term according to the CCG rules and obtain its β-normal form M
3. Obtain the negation scope from M

In the following sections, we provide a detailed explanation of 1. and 3., which represent the key components of this evaluation.

3.1 λ-term as Semantic Representation

We do not rely on any specific semantic theory; instead, we use λ-terms that encode only the functional relationships. More specifically, for each word w, a corresponding symbol $\mathtt{w}'$ is introduced and assigned as its semantic representation. By adopting such a primitive semantic representation, it is possible to directly evaluate whether CCGbank correctly captures negation scope without being influenced by any particular semantic theory. In the following sections, we explain special treatment for handling the annotations unique to CCGbank.

3.1.1 Adjunct

Adjuncts are represented by categories of the form $X\alpha|X\alpha$, where X is a basic category. Both instances of $X\alpha$ share the same argument stack α in such cases. For example, in the category $(\mathtt{S\backslash NP})\backslash(\mathtt{S\backslash NP})$, the information about the NP argument expected by the left-hand $\mathtt{S\backslash NP}$ must also be passed to the right-hand $\mathtt{S\backslash NP}$. Accordingly, for a word w with the category $(X\alpha|X\alpha)\beta$, the following λ-term is assigned:

$$\lambda b_1 \cdots \lambda b_{|\beta|} \lambda x \lambda a_1 \cdots \lambda a_{|\alpha|}.M$$
$$M \equiv \mathtt{w}' b_1 \cdots b_{|\beta|}(x\ a_1\ \cdots\ a_{|\alpha|}) a_1 \cdots a_{|\alpha|}$$

The λ-terms corresponding to the categories in the argument stack β are sequentially assigned to $b_1, \ldots, b_{|\beta|}$ in the above λ-term. Subsequently, the λ-term associated with the right-hand $X\alpha$ is assigned to x. The variables $a_1, \ldots, a_{|\alpha|}$ are then substituted into this expression, ensuring that the argument stack α on both sides of the adjunction shares the same semantic information.

3.1.2 Coordination

Coordinating conjunctions are represented by a special category called conj, which serves to connect constituents of the same category. It is necessary to distribute the information received by the entire coordinated phrase to each conjunct in coordination structures where the conjuncts are complex categories. For instance, in a coordination structure of verb phrases, the category is $\mathtt{S\backslash NP}$ and the conjuncts must receive information about the subject NP. As with adjuncts, this information sharing

[2] McKenna and Steedman (2020) proposes a method that resolves negation scope using CCGbank derivation trees; however, it is not suitable for our purpose because it uses CCG categories as features and does not necessarily conform to the semantic compositionality of CCG.

must be properly represented in the corresponding λ-term. When the category of the coordination structure is X with $arity(X) = n$, the λ-term for the coordinating conjunction is defined as follows:

$$\lambda x_1 \lambda x_2 \lambda y_1 \cdots \lambda y_n . \mathtt{w}'(x_1\ y_1\ \cdots\ y_n)(x_2\ y_1\ \cdots\ y_n)$$

The λ-term corresponding to the right conjunct is assigned to x_1. This term receives the semantic representation passed to the entire coordination structure via the variables $y_1, \ldots, y_n$. Similarly, the variable x_2 is assigned the λ-term corresponding to the left conjunct, which also receives the semantic representation passed to the entire coordination structure via $y_1, \ldots, y_n$.

3.1.3 Type Changing Rule

CCGbank includes type-changing rules that convert a category X into another category Y. We interpret such cases as involving an implicit lexical item with the category Y/X, which applies to the category X via function application to yield the category Y.

3.1.4 Non-local Dependencies

Category information is sometimes shared even outside of adjunct constructions. For example, when the word "which" in the phrase "food which John likes" is assigned the category $(\mathtt{NP}\backslash\mathtt{NP})/(\mathtt{S}\backslash\mathtt{NP})$, the $\mathtt{NP}$ of $\mathtt{S}\backslash\mathtt{NP}$ and the right-hand $\mathtt{NP}$ of $\mathtt{NP}\backslash\mathtt{NP}$ refer to the same entity (food). To handle such non-local dependencies, we follow the treatment outlined in the CCGbank User's Manual (Hockenmaier and Steedman, 2005) and handle them analogously to adjuncts. Accordingly, we assign the following λ-term to the word "which":

$$\lambda x \lambda y . \mathtt{which}'(xy)y$$

3.2 Negation Scope of CCGbank

Based on the β-normal form λ-term L, the negation scope for the negation cue c is obtained as follows:

- For any subterm of the form fa in L, if $\mathtt{func}(f)$ is the λ-term associated with the negation cue c, and the symbol $\mathtt{w}'$ corresponding to the word w is an element of $\mathtt{Sym}(a)$, then w is included in the scope of c.

$\mathtt{func}$ and $\mathtt{Sym}$ are defined as follows where M and N are λ-terms:

$$\mathtt{func}(f) = \begin{cases} f & (f \text{ is a symbol}) \\ \mathtt{func}(M) & (f = MN) \\ \text{undefined} & (f = \lambda x.M) \end{cases}$$

	# cues	# sentences
no	618	590
not	4505	4206

Table 1: Statistics of negation cues

	Precision	Recall	F-Score
no	83.35	27.54	41.40
not	84.08	92.50	88.09

Table 2: Results of experiment 1

$$\mathtt{Sym}(a) = \begin{cases} \{a\} & (a \text{ is a symbol}) \\ \mathtt{Sym}(M) \cup \mathtt{Sym}(N) & (a = MN) \\ \mathtt{Sym}(M) & (a = \lambda x.M) \end{cases}$$

4 Experiment 1: Validity of CCGbank

To evaluate the validity of CCGbank from a negation scope perspective, we conducted an experiment using the semantic representation described in Section 3. We used Sections 02-21 of the CCGbank, which are traditionally used as training data.

4.1 Negation Cue

In this experiment, we focus on the negation cues "no" and "not", as they are frequently used in English. Other negation cues, including multiword constructions such as "by no means" and "no longer", as well as instances where "no" and "not" appear as part of such expressions, are excluded. Table 1 presents the total number of negation cues and sentences.

4.2 Negation Scope

In this experiment, the gold standard negation scopes are provided by NegBERT[3](Khandelwal and Sawant, 2020). This model is selected due to its strong performance, having achieved an F-score of 92.94% on the ConanDoyle-neg. A preliminary experiment confirms that NegBERT demonstrates comparable performance within the CCGbank domain. Specifically, Section 00 of CCGbank is manually annotated for negation scope following the annotation guidelines of the ConanDoyle-neg (Morante et al., 2011), and NegBERT is evaluated against these annotations. The result yields a token-level F-score of 90.90%, supporting the model's suitability for this domain.

[3]https://github.com/adityak6798/Transformers-For-Negation-and-Speculation

negation cue	transformation	Precision	Recall	F-Score
no	None	83.35	27.54	41.40
no	quantifier	71.87	62.06	66.61
no	restrictive post-nominal modification	90.04	49.34	63.75
no	quantifier & restrictive post-nominal modification	77.53	83.80	80.54
not	None	84.08	92.50	88.09
not	quantifier	84.08	92.50	88.09
not	restrictive post-nominal modification	84.97	92.11	88.40
not	quantifier & restrictive post-nominal modification	84.97	92.10	88.39

Table 3: Results for each configuration

4.3 Results of Experiment 1

We evaluated the degree of scope agreement using the token-level F-score. The results are presented in Table 2. As evident from these results, the negation scopes derived from the CCGbank align well with the gold-standard for the negation cue "not", but less so for "no". This means the following:

- The high F-Score with respect to "not" demonstrates that the primitive semantic representation described in Section 3 work well. This is due to the identical treatment of "not" in CCGbank and in CCG.
- Nevertheless, the F-Score with respect to "no" is low. This suggests one possibility: CCGbank does not capture the negation scope of "no".

5 Experiment 2: Linguistic Analysis

Two primary factors may account for the low F-Score for "no". First, the CCGbank does not validly capture the semantics of quantifiers. In Steedman's (2000) analysis, the valid CCG derivation tree for sentence (1) is that shown in Figure 1; however, CCGbank represents it as in Figure 3. Second, in the CCGbank, all post-nominal modifications are uniformly treated as non-restrictive[4] (Hockenmaier and Steedman, 2007; Honnibal et al., 2010). For example, the derivation tree for sentence (3) is shown in Figure 4.

(3) He made **no** remark as to the contents.

These CCGbank's invalid treatment of quantifiers and post-nominal modifications are likely to be the cause of disagreements in negation scope.

[4]In CCG, restrictive and non-restrictive post-nominal modifiers are represented as $\mathtt{N}\backslash\mathtt{N}$ and $\mathtt{NP}\backslash\mathtt{NP}$, respectively.

To investigate the effect of such treatment on CCGbank's negation scope, we modify the CCGbank derivation trees as described in Sections 5.1 and 5.2 and recalculate the F-Score.

5.1 Quantifier

To adopt Steedman's CCG analysis for quantifiers, we transform the category of "no" according to the following:

- When a NP beginning with "no" is an argument of a category of the form $X|_1\mathtt{NP}$ (where $X \neq \mathtt{NP}$), we replace all occurrences of this NP with the category $X|_2(X|_1\mathtt{NP})$ (where $|_2$ denotes the inverse slash of $|_1$).

5.2 Post-nominal Modification

To treat post-nominal modifications as restrictive, we convert CCGbank derivation trees using the conversion rule shown in Figure 5. This rule is identical to the one proposed by Honnibal et al. (2010). Figure 6 shows the modified version of the derivation tree in Figure 4.

5.3 Results of Experiment 2

The experimental results corresponding to each configuration are presented in Table 3. The results show that the negation scope on the modified derivation trees align well with the gold-standards. This indicates the following:

- The current CCGbank (rather than CCG itself) is semantically invalid from the viewpoint of negation scope.

6 Conclusion

This study investigated the validity of CCGbank from the negation scope perspective. Specifically, we compared the negation scopes derived from

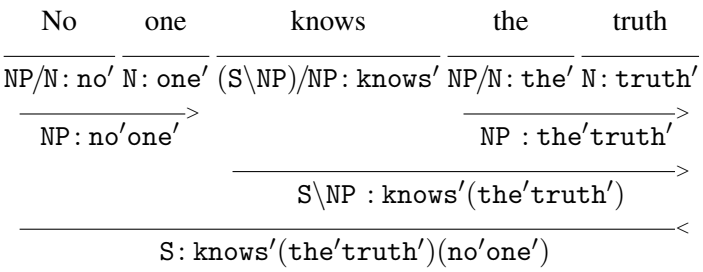


Figure 3: CCGbank derivation tree of sentence (1)

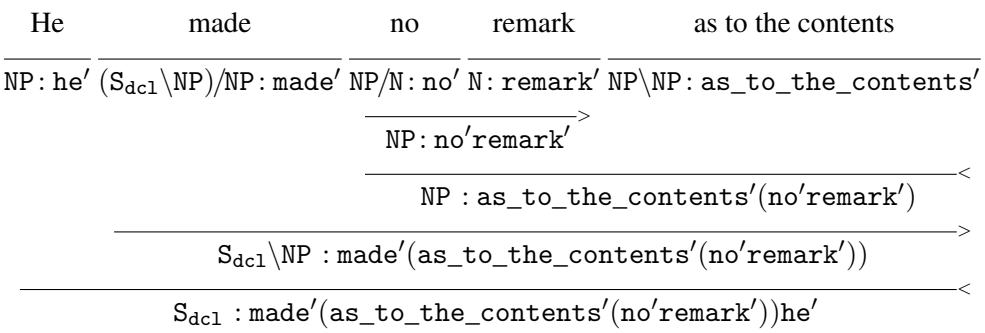


Figure 4: CCGbank derivation tree of sentence (3)

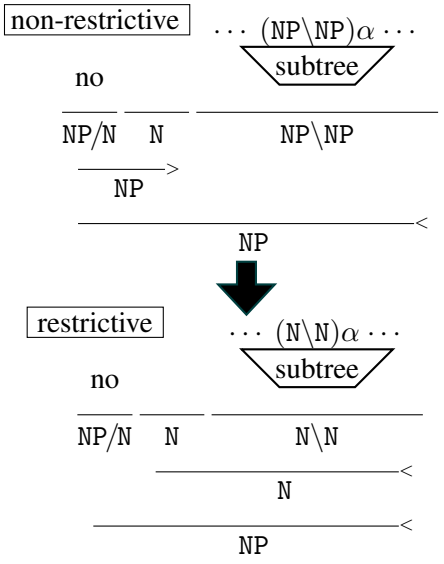


Figure 5: Conversion from non-restrictive to restrictive post-nominal modification.

CCGbank with those defined in the ConanDoyle-neg. The results demonstrated that CCGbank's scopes of the negation cue "not" align well with the ConanDoyle-neg, whereas those of "no" show notable disagreements. We attribute this results to CCGbank's invalid handling of quantifiers and restrictive versus non-restrictive post-nominal modification. However, after modifying CCGbank derivation trees, the negation scopes for "no" align more closely with the gold standards. This finding indicates that there are still challenges in building a semantically valid CCGbank. Therefore, it is essential to address these issues through appropriate corrections on CCGbank to ensure a more valid analysis.

7 Limitation

This study is anchored to NegBERT's performance, given that the gold-standard negation scope was derived from its output. A more rigorous evaluation would therefore require manual annotation of negation scope within CCGbank.

Furthermore, our analysis was limited to the negation cues "no" and "not" and to the phenomena of quantification and restrictive post-nominal modification. Future research should broaden this inquiry by examining additional negation cues and a wider array of linguistic constructions to more comprehensively evaluate the validity of CCGbank.

Acknowledgments

This work was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 25K03418) of JSPS.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Con-*

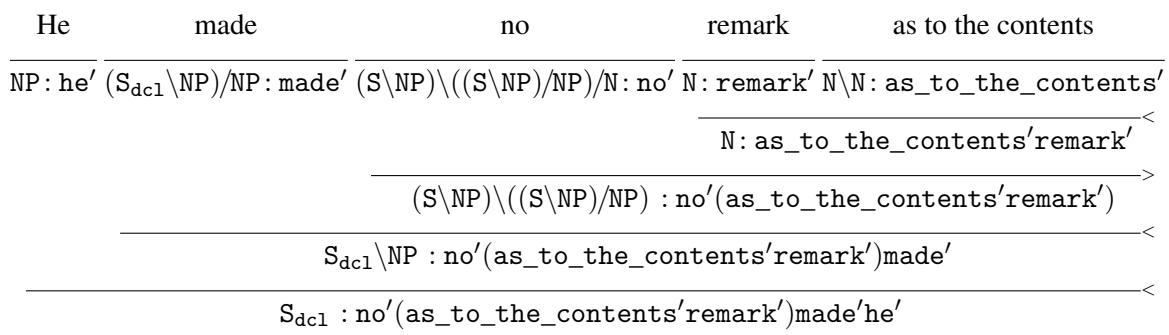


Figure 6: Modified version of CCGbank derivation tree of sentence (3)

ference on Empirical Methods in Natural Language Processing, pages 2492–2502.

Sebastian Beschke and Wolfgang Menzel. 2018. Graph algebraic Combinatory Categorial Grammar. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 54–64.

Stephen Boxwell and Michael White. 2008. Projecting Propbank roles onto the CCGbank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 103–110.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–63.

Julia. Hockenmaier and Mark Steedman. 2005. CCGbank: User's Manual.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Matthew Honnibal, James R. Curran, and Johan Bos. 2010. Rebanking CCGbank for improved NP interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 207–215.

Hai Hu, Qi Chen, and Larry Moss. 2019. Natural language inference with monotonicity. In *Proceedings of the 13th International Conference on Computational Semantics*, pages 8–15.

Hai Hu and Larry Moss. 2018. Polarity computations in flexible categorial grammar. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 124–129.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5739–5748.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3190–3195.

Mike Lewis, Kenton Lee, and Luke Zettlemoyer. 2016. LSTM CCG parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 221–231.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–720.

Nick McKenna and Mark Steedman. 2020. Learning negation scope from syntactic structure. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics*, pages 137–142.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1563–1568.

Roser Morante, Sara Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope : Guidelines v1.0. Technical report, University of Antwerp.

Mark Steedman. 2000. *The syntactic process*. MIT Press, Cambridge, MA, USA.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Ryosuke Yamaki, Tadahiro Taniguchi, and Daichi Mochihashi. 2023. Holographic CCG parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 262–276.

Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. 2017. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 277–287.

VIDAI: VIDukathAI Interpretation Through Analysis of In-context Reasoning in Tamil using LLMs

R S Mughil Srinivasan[1]
106122104@nitt.edu

Kesavan T[1]
106122068@nitt.edu

Abhijith Balan[1]
406123001@nitt.edu

Abhinav P M[2]
abhinav.pm@research.iiit.ac.in

Parameswari Krishnamurthy[2]
param.krishna@iiit.ac.in

Oswald C[1]
oswald@nitt.edu

[1]National Institute of Technology, Tiruchirappalli
[2]International Institute of Information Technology, Hyderabad

Abstract

This work investigates VIDAI, a study on Tamil vidukathai, where *vidai* means 'answer' and *vidukathai* means 'riddle' in the Tamil language, focusing on the challenges Large Language Models (LLMs) face in solving them. Tamil, a morphologically rich and culturally embedded classical Dravidian language of India, with over 2000 years of history, is spoken officially in three countries and across global diaspora communities. Although models such as LLaMA, Phi, Gemma, and Qwen excel in general NLP tasks, they struggle with Tamil riddles due to their reliance on next-token prediction and limited reasoning ability. Tamil riddles frequently use metaphors, puns, cultural references, and abstract logic, posing difficulties for models trained primarily on generic corpora. We curated a dataset of 2,283 riddles[1] and evaluated the models under various prompting strategies. The highest performance achieved was a BERTScore of 0.846 with a random 1-shot no-CoT prompt. VIDAI's findings highlight riddles as a promising benchmark for testing reasoning in LLMs.

Keywords: CoT, In-Context Learning, LLM, Question Answering, Riddle, Tamil

1 Introduction

Large Language Models (LLMs) such as GPT, LLaMA, Phi, and Gemma have advanced natural language processing, excelling in summarization, translation, question answering, and text generation (Sanchez-Bayona and Agerri, 2025; Tong et al., 2024; Giadikiaroglou et al., 2024; Jiang et al., 2023). These strengths stem from large-scale pretraining and fine-tuning on diverse datasets. However, LLMs struggle with tasks that require symbolic abstraction, associative reasoning, and cultural grounding, such as solving a riddle (Liu et al., 2024).

[1]https://anonymous.4open.science/r/TamilRiddlesDataset-48C3

LLMs are trained on next token prediction objectives (Lin et al., 2025), favoring high-frequency, contextually probable outputs. Although effective for factual tasks, this biases models against riddles, which rely on ambiguity, metaphor, and wordplay. Solving riddles requires inference, analogy, and creative abilities that were not explicitly learned during training.

These challenges are amplified in morphologically rich and culturally grounded languages such as Tamil. With more than 2000 years of literary history, Tamil has agglutinative morphology, rich inflections, eight grammatical cases (Lehmann, 1993), and distinctive phonology documented in the *Tolkāppiyam* (Tolkāppiyar, Ancient). Its oral traditions feature riddles (*Vidukathai* in Tamil language) using metaphors, puns, idioms, and poetic conventions such as *venpa*, *ullurai*, and *iraicchi* (Schiffman, 1999; Iyyanarithanar, 9th Century CE). These cultural markers are rarely present in large corpora in English, causing models to default to generic or plausible answers.

VIDAI introduces the first dedicated Tamil riddles dataset, providing a benchmark to evaluate LLM performance by inferring and understanding Tamil riddles, and assesses open-source models using structured prompting, semantic example selection, and multi-metric evaluation. VIDAI's study highlights reasoning gaps and conditions that improve model performance, offering insight into handling metaphor-rich, culturally specific tasks.

The paper is organized as follows: §1 introduces the motivation, contributions, and problem statement. Related riddle-solving and multilingual LLM reasoning work are reviewed in §2. §3 explains the creation of the Tamil Dataset. The details of the design of VIDAI and the prompting strategies are explained in §4. §5 discusses the evaluation, metrics, and observations of the experiments. §6 presents the key takeaways. The conclusion and directions for future work are presented

in §7.

1.1 Contributions

- **Dataset Creation:** Compiled a cleaned dataset of 2,250+ Tamil riddles from books and public sources, categorized into Objects, Nature, Actions, People, and Abstract Concepts.
- **Few-Shot Prompting and CoT:** Designed random and semantically similar few-shot settings (1, 2, 3, 5 shots) and generated CoT explanations using ChatGPT for selected examples and manually curated to ensure alignment with human interpretations in solving a riddle.
- **Multi-Metric Evaluation:** Used Exact Match, Levenshtein Distance, and BERTScore to assess both literal and semantic correctness.
- **Local LLM Execution:** Ran experiments on Phi-4, Gemma2: 9b, Gemma3.1: 12b, LLaMA3: 8b and Qwen2.5: 7b via Ollama[2].
- **Riddle Reasoning in Tamil:** Focused on metaphor-rich riddles in a low-resource language to test cultural and linguistic reasoning.

1.2 Problem Description

LLMs excel at next-token prediction, favoring common, contextually likely continuations, which suits tasks like summarization or dialogue. However, the riddles are based on misdirection, wordplay, and layered meanings, demanding abstraction and creative reasoning. Tamil riddles add challenges with cultural references, phonetic puns, and idioms, rare in mainstream corpora, reducing model effectiveness, and highlighting a gap in multi-step, metaphorical reasoning.

The problem statement is defined as follows: Let $\mathcal{R} = \{(r_1, a_1), (r_2, a_2), \ldots, (r_n, a_n)\}$ be a collection of Tamil riddles, where each r_i represents a riddle, a_i - its corresponding answer, and $n = |\mathcal{R}|$, for $1 \leq i \leq n$. The task is to curate $\mathcal{R}$, then evaluate a series of Large Language Models to answer the riddles by predicting an answer b_i and computing the similarity of (b_i, a_i) with the help of various metrics.

[2] https://ollama.com/

2 Related Work

Recent studies have begun to treat riddle solving as a test of LLM reasoning. (Panagiotopoulos et al., 2025) proposed a context-reconstructed augmentation method that improves performance by providing structurally similar riddles as few-shot prompts. However, their work focuses on multiple-choice formats rather than VIDAI's method of QA style. (Giadikiaroglou et al., 2024) survey puzzles, classifying them into rule-based and rule-less forms, and note that the riddles remain challenging for LLMs. Although they call for better datasets and hybrid methods, their analysis is largely language-agnostic and not tailored to low-resource languages. (Lin et al., 2021) and (Jiang et al., 2023) address commonsense and linguistic creativity in English riddles and lateral puzzles. However, their methods are language-specific and do not address metaphor in non-English contexts. Similarly, (Tan et al., 2016) tackles the Chinese character riddles using language-specific fine-tuning, limiting cross-lingual applicability. Few-shot learning studies (Brown et al., 2020; Agarwal et al., 2025) informed VIDAI's shot size design, while (Wei et al., 2022b) inspired VIDAI's CoT approach. However, these works primarily test English tasks. (Zhang and Wan, 2022) and (Xu et al., 2023) confirm riddles as multilingual challenges but focus on multiple-choice formats, unlike VIDAI's QA format. The work of (Liu et al., 2022) validates the benefit of semantically similar examples, aligning with VIDAI's sampling strategy. Reasoning-oriented prompting research (Kojima et al., 2022; Fu et al., 2023) offers useful insights, but is based on mathematical and logic puzzles. Theoretical studies (Han et al., 2024; Wei et al., 2022a) explore general reasoning but neglect the resolution of culturally rich metaphors. In general, no major study goals are open in any Indian language, such as Tamil, which requires deep cultural and symbolic reasoning.

3 Dataset Creation

3.1 Dataset Sources

We collected the Tamil riddle dataset from various public sources, including books and online repositories. The sources are from classical Tamil riddle books such as (Muthaiah, 1987) and (Manivasan, 2018). We incorporate riddles from educational websites such as (Dheivegam, 2023),

(FreshTamil.com, 2020), (FreshTamil.com, 2024), and quiz portals such as (Vinaval, 2025) and (Vidukathaigal, 2025). The VIDAI's final dataset consisted of 2,283 unique Tamil riddles, each paired with a single ground-truth answer. This represents the largest available collection of Tamil riddles from which we could extract the best from online sources.

3.2 Riddle Structure

Riddles ranged from one to five poetic lines in colloquial Tamil, with answers as single words or short phrases. While some mapped to concrete concepts, others required abstract or symbolic interpretation.

3.3 Categorization of Riddles

We classified VIDAI's dataset based on the Tamil riddle answers into five broad semantic categories.

- **Natural Elements and Weather**: Answers related to natural phenomena and cycles, often expressed through environmental metaphors. Examples: சூரியன் - (Sūriyaṉ, Sun), காற்று - (Kāṯru, Wind).
- **Human Body and Senses**: Refers to body parts or sensory actions, typically using anatomical or functional metaphors. Examples: மூச்சு - (Mūccu, Breath), நாக்கு - (Nākku, Tongue).
- **Objects and Tools**: Man-made items are described through their form, function, or purpose. Examples: நாற்காலி - (Nāṟkāli, Chair), கடிகாரம் - (Kaṭikāram, Clock).
- **Food and Plants**: Edible items or plants, often described using taste, texture, or appearance. Examples: வெங்காயம் - (Veṅkāyam, Onion), கரும்பு - (Karumpu, Sugarcane).
- **Animals and Insects**: Creatures referenced through behavior, sounds, or cultural associations. Examples: சிலந்தி - (Silanti, Spider), நாய் - (Nāy, Dog).

4 Design

Figure 1 shows the architecture of VIDAI, using five open-source LLMs: LLaMA 3.1 (8B), Phi-4, Gemma 2 (9B), Gemma 3.1 (12B), and Qwen 2.5 (7B) with zero, one, and few-shot prompting. In few-shot settings, the models received 2, 3, and 5 riddle-answer examples.

We employ two sampling strategies: Random Sampling and Semantic Similarity Sampling. In Random Sampling, examples were selected arbitrarily from the training set, with each example likely originating from a different class. In contrast, Semantic Similarity Sampling involved first selecting one of the five predefined categories, established through manual classification of the dataset based on the answers, as described in Section *Categorization of Riddles*, and then randomly drawing all examples from that category. For instance, in a three-shot prompt, Random Sampling would yield three examples from potentially different categories, whereas Semantic Similarity Sampling would select a single category at random and then draw three examples exclusively from it.

4.1 Chain-of-Thought

Initially, the examples included only riddle-answer pairs. In a second phase, we extended the prompts using **Chain-of-Thought (CoT)** reasoning, adding explanatory reasoning to each example. CoT explanations were presented in a consistent deductive format in all examples. Each riddle is first restated and explained in English, followed by the revelation of the answer. The explanation then proceeds by mapping each segment of the riddle to its underlying meaning, interpreting phrases in relation to the proposed answer, and the justification explicitly shows how the answer satisfies each part of the riddle. This ensures that the meanings embedded in figurative, metaphorical, or cultural contexts are extracted and clarified. The length of the explanations naturally varies according to the complexity of the riddle.

4.2 Direct Prompting

In this setup, the model was given only the riddle, without any additional guidance, and tasked with producing an answer. This zero-shot approach relies entirely on the internal knowledge and reasoning of the model to interpret and solve the riddle.

An example prompt given without CoT explanation :

```
Provide only the final answer in
Tamil without any translations or
explanations in English for the given
Tamil riddle below.
Question : பறக்கும் ஆனால் பறந்து போகாது,
```

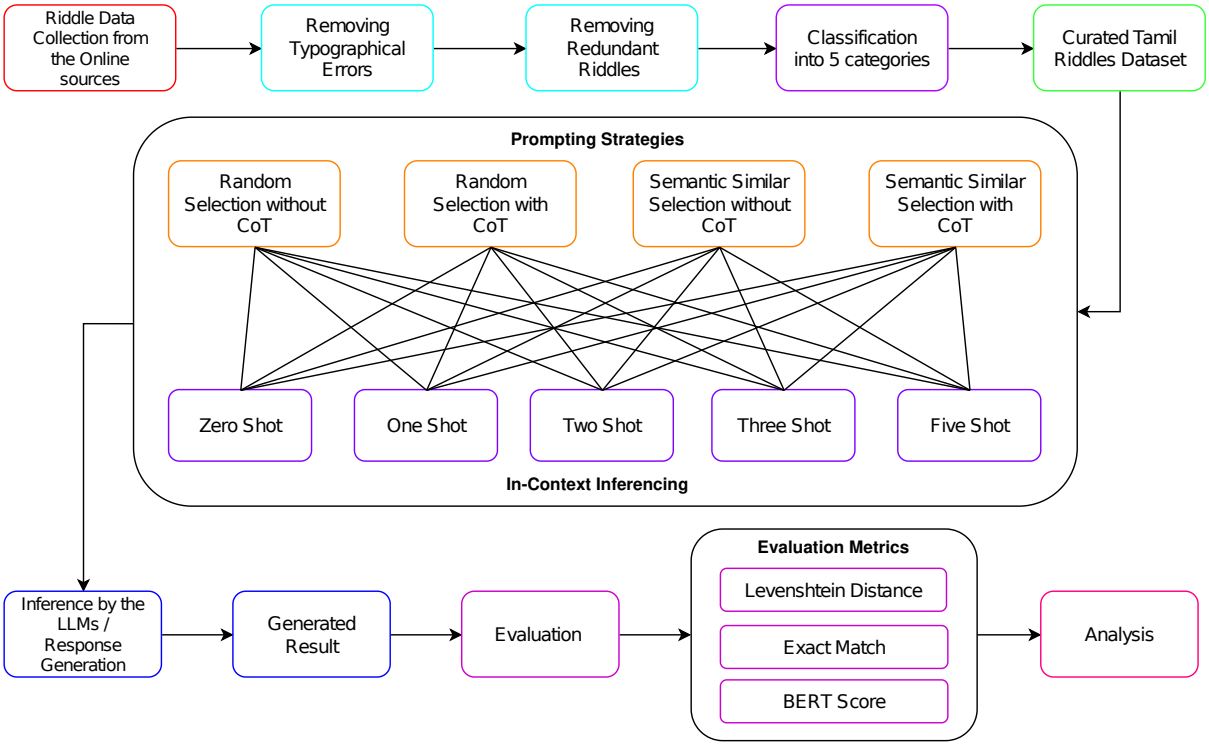


Figure 1: Architecture diagram of VIDAI: VIDukathAI Interpretation Through Analysis of In-context Reasoning in Tamil using LLMs

அது என்ன?

Answer:

4.3 Few-Shot Prompting

We adopt a few-shot strategy with Chain-of-Thought (CoT) reasoning, where each example includes the riddle, its answer, and step-by-step reasoning to interpret metaphors, recognize symbols, and eliminate unlikely options. These CoT explanations, generated using ChatGPT-4o with the riddle and its known answer, were then manually curated to ensure coherence, contextual relevance, logical foundation and alignment with human interpretations.

An example of a 1-shot prompt given with CoT explanations is:

Question: வீட்டுக்குள்ளே இருப்பாள், விந்தையாகப் பேசுவாள், வந்தவரை வா என்பாள், வாசல் தாண்டிப் போகமாட்டாள். அவள் யார்?

Answer: நாக்கு

Explanation: The riddle describes something that stays inside the house, speaks mysteriously, invites people, and never crosses the threshold. The answer is நாக்கு (tongue). வீட்டுக்குள்ளே இருப்பாள் refers to the tongue that remains inside the mouth (the house). விந்தையாகப் பேசுவாள் highlights its role in speaking, வந்தவரை வா என்பாள் shows how it welcomes speech, and வாசல் தாண்டிப் போகமாட்டாள் means it never leaves the mouth.

Provide only the final answer in Tamil without any translations or explanations in English for the given Tamil riddle below.

Question: பாதுகாப்பான பெட்டிக்குள்ளே பலரும் விரும்பும் கடிகாரம், அது என்ன?

Answer:

Enriched examples are especially useful for solving riddles, where metaphorical and cultural understanding is the key. Comparing these two prompting approaches helps evaluate how in-context learning and example-based reasoning affect LLM performance on complex, ambiguous queries. All experiments were run locally using models downloaded via Ollama, depending on available resources.

5 Evaluation, Results and Observations

To assess model performance in solving Tamil riddles, we employed four key evaluation metrics, each capturing different aspects of answer quality:

- **Exact Match:** Checks if the model answer exactly matches the ground truth. It is simple but rigid, often penalizing correct synonyms or alternate Tamil wordings. It is defined as:

  $$\text{Exact Match (EM)} = \frac{\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}{N}$$

- **Levenshtein Distance:** Measures similarity via minimal character edits, normalized as [0, 1]. It detects typos and near matches, but misses semantically correct words with different spellings. It is defined as:

  $$\text{Levenshtein Similarity} = 1 - \frac{d_{\text{lev}}(\hat{y}, y)}{\max(|\hat{y}|, |y|)}$$

- **BERTScore:** Uses transformer embeddings to compare tokens, capturing semantic meaning and paraphrases, effective for Tamil riddles with subtle contextual nuances. It is defined as:

  $$\text{BERTScore}_{F1} = \frac{1}{N} \sum_{i=1}^{N} F1(\hat{y}_i, y_i)$$

Tamil riddles, rich in metaphor and cultural nuance, require evaluation beyond strict string matching. We use BERTScore with surface metrics to capture semantic equivalence across varying wordings, analyzing the effects of context, model size, and explanation style across multiple prompting setups.

5.1 Algorithm Trace Example

This section presents a sample execution of Algorithm 1 for one configuration. The model is Phi-4, using the *Semantic similarity* prompting strategy with 3 shots. The test riddle is மேலே மேலே போகும், கீழே கீழே போகாது. என்ன? (Mēlē mēlē pōkum, kīḻē kīḻē pōgātu. Enna?) – It keeps going upward, but never goes downward. The ground truth answer is வயது (Vayatu) – Age. The evaluation begins with `GetFewShotExamples(model, riddle, shot=3, strategy="semantic")`, which selects three semantically related examples, focusing on abstract concepts from *Human Body & Senses* category. Then, `LoadTestSet()` loads the riddle and its ground truth answer. Next, `ConstructPrompt(few_shot_examples, test_riddle)` builds the prompt by combining the examples with the target riddle in QA format. Passing this on to `GenerateAnswer(model, prompt)` yields புகை (Pogaei) – Smoke. Although incorrect, this highlights the model's reliance on surface-level cues rather than abstract temporal reasoning. Finally, `Evaluate()` computes metrics such as *BERTScore*, reflecting partial semantic proximity in the metaphorical interpretation between smoke and age.

Algorithm 1 Evaluate Models on Riddle QA with Few-Shot Prompts

Input: *TestSet*, *FewShotData*, *ShotCounts*, *Models*, *PromptModes*
for all Model ∈ Models **do**
  Load model via local or remote interface
  for all PromptMode ∈ PromptModes **do**
    for all ShotCount ∈ ShotCounts **do**
      if ShotCount > 0 **then**
        Examples ← `GetFewShotExamples(` FewShotData, PromptMode, ShotCount)
      else
        Examples ← ∅
      end if
      (Questions, Answers) ← `LoadTestSet` (TestSet)
      Predictions ← [ ]
      for all Question ∈ Questions **do**
        Prompt ← `ConstructPrompt(` Examples, Question, PromptMode)
        Response ← `GenerateAnswer(` Model, Prompt)
        Append Response to Predictions
      end for
      Metrics ← `Evaluate`(Predictions, Answers)
      Evaluation Metrics: *Exact Match*, *Cosine Similarity*, *BERTScore*, *Levenshtein Similarity*
    end for
  end for
end for

5.2 Results and Discussions

We ran the five open-source LLMs on 14th Gen Intel Core i7 CPU, NVIDIA T1000 8 GB GPU with 16 GB Main Memory and 512 GB Hard Disk on Linux platform using Python code. The models LLaMA 3.1 (8B), Phi-4, Gemma 2 (9B), Gemma 3.1 (12B), and Qwen 2.5 (7B) were evaluated across four setups: semantically similar or random riddles, with or without CoT explanations, under zero, one, and few-shot (2, 3, 5) conditions. Performance was measured using exact match, Levenshtein distance, and BERTScore. The results are tabulated in Table 1 and Table 2 and visualized in Figures 3, 4, 5 and 6.

The direct comparison of State-of-the-art was not relevant because previous work differs substantially; for example, RiSCORE (Panagiotopoulos et al., 2025) treats riddles as multiple choice with ranking metrics, while RiddleSense (Lin et al.,

2021) and BRAINTEASER (Jiang et al., 2023) address English riddles using knowledge-based solvers in different formats. VIDAI's focus was on raw inference in natural QA. In this setup, each prompt contained only a Tamil riddle, and the model was asked to generate the answer directly without any additional clues or multiple choice options.

5.3 LLaMA 3.1 (8B)

LLaMA 3.1 showed steady results: Best Exact Match 0.007 (5-shot CoT, semantic), Levenshtein 0.140 (3-shot CoT, semantic), BERTScore 0.775 (1-shot no-CoT). The results highlight its strong multilingual pretraining yet limited adaptability to the Tamil riddle's metaphorical style. CoT slightly improved generalization, but the modest gains suggest a token prediction bias rather than true abstract reasoning.

5.4 Phi-4

Phi-4 peaked with 5-shot CoT semantic runs: Exact Match 0.007, Levenshtein 0.141, BERTScore 0.767. Its compact, reasoning-oriented training made it suitable for structured prompts. CoT aligned well with the modeling intermediate steps. Although absolute scores remain modest, consistent improvements show adaptability. Still, the lack of deep contextual embeddings limits its figurative understanding of Tamil riddles.

5.5 Gemma 2 (9B)

Gemma 2 varied widely. Best Exact Match 0.022 (5-shot CoT random), BERTScore 0.846 (1-shot no-CoT random), Levenshtein 0.303 (5-shot no-CoT random). These spikes suggest reliance on lexical overlap rather than reasoning, aided by token similarity. CoT often reduced performance, implying a mismatch with training. The results show that size alone does not guarantee reasoning capacity.

5.6 Gemma 3.1 (12B)

Gemma 3.1 produced consistent results: BERTScore 0.785 (5-shot no-CoT semantic), Exact Match 0.018 (zero-shot CoT), Levenshtein 0.170 (3-shot CoT random). Its larger size likely supports better semantic generalization and metaphor detection. However, CoT offered little benefit, suggesting that internal reasoning suffices. In general, Gemma 3.1 balances surface accuracy and semantic similarity, adapting to prompting conditions.

5.7 Qwen 2.5 (7B)

Qwen 2.5 scored lowest overall: Exact Match 0.004, Levenshtein 0.149, BERTScore 0.768 (all no-CoT, both prompts). The results show a weakness in multi-step reasoning and poor CoT handling, reflecting multilingual pretraining not tuned for Tamil. Its training favors fluency and factual QA over metaphorical reasoning. Low scores highlight difficulty with analogy, figurative interpretation, and inference.

5.8 Cross-Model Observations

- **Few-shot prompting** (3–5 shots) consistently outperforms **zero/one-shot**, mainly in **exact match** and **Levenshtein**.
- **Semantically similar example selection** outperforms **random sampling** in smaller models by providing more relevant **contextual alignment**.
- **CoT explanations** improve mid-sized models but sometimes reduce performance in larger ones due to **prompt-structure mismatch**.
- **Embedding-based metrics** better capture **semantic understanding** than **exact match**, reflecting the high linguistic diversity of valid **Tamil riddle answers**.

In summary, **architecture**, **training data**, and **reasoning alignment** significantly influence performance in decoding **Tamil riddles**.

6 Key Takeaways

- **Riddle solving needs more than facts:** Tamil riddles rely on metaphor, symbolism, and cultural cues that LLMs often miss.
- **Cultural grounding matters:** Most models overlook idioms and poetic clues, reducing accuracy.
- **Bigger models aren't always better:** Large LLMs are stable, but smaller ones can sometimes outperform them.
- **Current metrics fall short:** The exact match is too strict, and even BERTScore cannot fully assess understanding.

Model - Shots	Semantically Similar Riddles			Randomly Selected Riddles		
	Exact Match	Leven-shtein	BERT Score	Exact Match	Leven-shtein	BERT Score
Llama3.1:8b - 0	0.31	13.73	77.12	0.13	13.53	76.94
Llama3.1:8b - 1	0.48	13.17	77.57	0.48	13.44	77.30
Llama3.1:8b - 2	0.35	12.54	77.16	0.44	13.69	77.40
Llama3.1:8b - 3	0.35	13.24	77.13	0.39	13.31	77.30
Llama3.1:8b - 5	0.66	13.44	77.24	0.57	13.22	77.15
Phi4 - 0	0.13	13.00	74.97	0.18	12.93	74.91
Phi4 - 1	0.26	11.61	72.62	0.44	13.47	75.36
Phi4 - 2	0.13	10.38	71.84	0.22	11.30	73.06
Phi4 - 3	0.44	11.91	73.04	0.44	11.51	73.90
Phi4 - 5	0.39	12.64	74.45	0.53	12.83	74.61
Gemma2:9b - 0	1.40	16.72	77.60	0.00	20.54	79.48
Gemma2:9b - 1	1.36	16.50	77.63	0.00	20.00	84.61
Gemma2:9b - 2	1.80	16.56	77.78	0.00	25.00	82.39
Gemma2:9b - 3	1.62	16.59	77.89	0.00	21.25	81.17
Gemma2:9b - 5	1.88	16.56	77.93	0.00	30.29	82.91
Gemma3.1:12b - 0	1.88	15.80	78.19	1.80	15.78	78.17
Gemma3.1:12b - 1	1.53	15.74	78.39	1.71	15.47	78.30
Gemma3.1:12b - 2	1.53	15.99	78.18	1.31	15.66	78.38
Gemma3.1:12b - 3	1.49	15.72	78.31	1.31	15.78	78.47
Gemma3.1:12b - 5	1.75	16.12	78.53	1.05	16.61	77.77
Qwen2.5:7b - 0	0.00	14.00	75.88	0.04	14.07	75.93
Qwen2.5:7b - 1	0.00	14.18	75.73	0.31	14.59	76.80
Qwen2.5:7b - 2	0.18	13.49	76.19	0.22	14.16	75.86
Qwen2.5:7b - 3	0.26	15.00	76.04	0.22	14.29	76.34
Qwen2.5:7b - 5	0.39	14.84	76.24	0.35	14.42	75.89

Table 1: Performance of various LLMs on Tamil riddles using Exact Match, Levenshtein Distance, and BERTScore without CoT explanations (values are scaled by a factor of 100)

Model - Shots	Semantically Similar Riddles			Randomly Selected Riddles		
	Exact Match	Leven-shtein	BERT Score	Exact Match	Leven-shtein	BERT Score
Llama3.1:8b - 0	0.18	13.10	77.04	0.31	13.17	76.92
Llama3.1:8b - 1	0.39	13.71	77.21	0.39	13.55	77.22
Llama3.1:8b - 2	0.35	12.85	77.00	0.48	13.65	77.31
Llama3.1:8b - 3	0.66	14.09	77.37	0.26	13.21	77.15
Llama3.1:8b - 5	0.79	13.61	77.11	0.48	13.51	77.33
Phi4 - 0	0.09	13.38	75.13	0.13	13.13	74.66
Phi4 - 1	0.39	13.37	75.44	0.44	12.86	75.19
Phi4 - 2	0.26	12.47	74.76	0.44	13.32	75.60
Phi4 - 3	0.48	13.96	76.31	0.53	13.14	75.56
Phi4 - 5	0.70	14.10	76.72	0.35	12.89	74.96
Gemma2:9b - 0	1.31	16.46	77.68	1.49	17.00	77.55
Gemma2:9b - 1	1.45	16.16	77.50	1.45	15.98	77.61
Gemma2:9b - 2	1.18	15.68	77.70	1.58	16.24	77.80
Gemma2:9b - 3	1.66	16.51	77.90	1.80	16.79	78.00
Gemma2:9b - 5	1.71	16.21	78.11	2.23	16.99	77.98
Gemma3.1:12b - 0	1.88	16.06	78.19	1.84	15.77	78.20
Gemma3.1:12b - 1	0.96	15.35	77.82	1.40	13.90	78.18
Gemma3.1:12b - 2	1.84	16.09	78.27	1.05	15.77	77.76
Gemma3.1:12b - 3	1.23	15.96	77.76	1.58	16.97	77.69
Gemma3.1:12b - 5	1.27	15.95	78.36	0.88	16.59	77.48
Qwen2.5:7b - 0	0.04	13.97	75.95	0.09	13.82	76.01
Qwen2.5:7b - 1	0.04	13.65	75.68	0.09	14.02	76.06
Qwen2.5:7b - 2	0.13	14.02	76.25	0.18	14.45	75.64
Qwen2.5:7b - 3	0.22	14.47	76.28	0.22	14.26	75.94
Qwen2.5:7b - 5	0.39	14.65	76.47	0.18	14.39	75.28

Table 2: Performance of various LLMs on Tamil riddles using Exact Match, Levenshtein Distance, and BERTScore with CoT explanations (values are scaled by a factor of 100)

- **Riddles are a strong benchmark:** They challenge creative and cultural reasoning beyond the generation of fluent text.

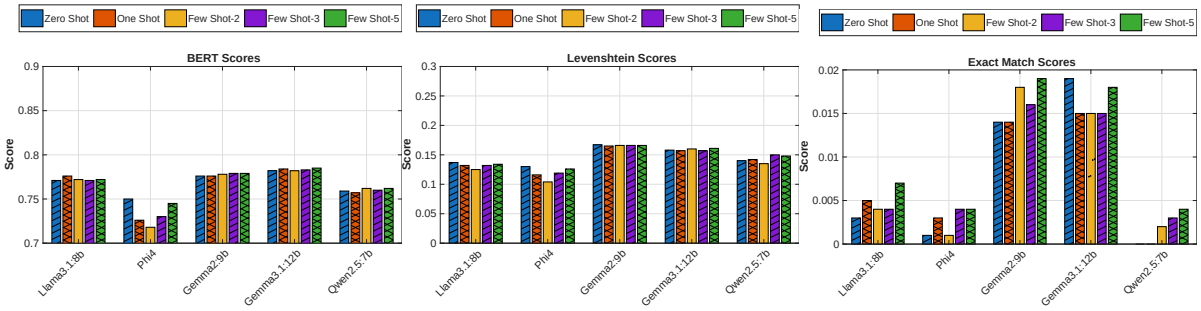
7 Conclusions and Future Work

This study evaluated modern open-source LLMs on Tamil riddles, highlighting their difficulty in handling metaphor, cultural nuance, and symbolic reasoning. Although structured prompts and CoT offered slight gains, the model relied primarily on surface patterns rather than deep understanding.

Future work shall focus on culturally rich datasets with annotated reasoning, explore hybrid neuro-symbolic methods, and incorporate cross-modal and retrieval-augmented approaches. Richer evaluation metrics and human-in-the-loop assessments are essential to push LLM toward genuine cognitive and cultural comprehension beyond fluent language generation.

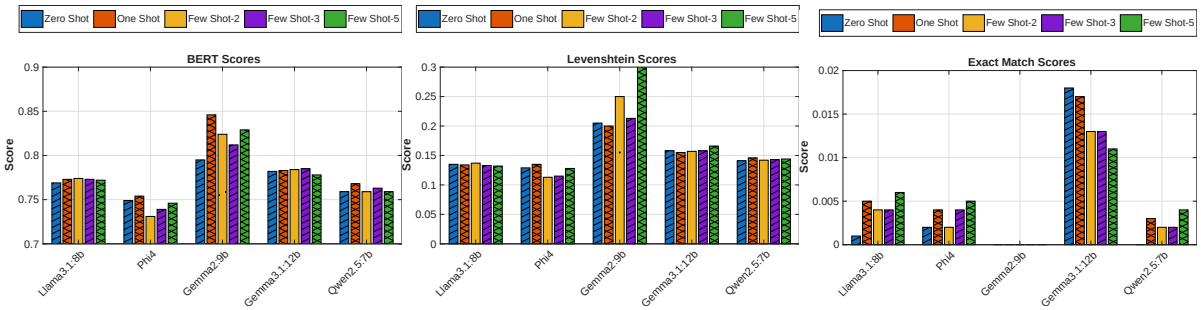
Acknowledgment

We gratefully acknowledge the authors of the online resources from which the riddles were scraped to create our dataset and the use of generative AI tools for help in paraphrasing and small edits.



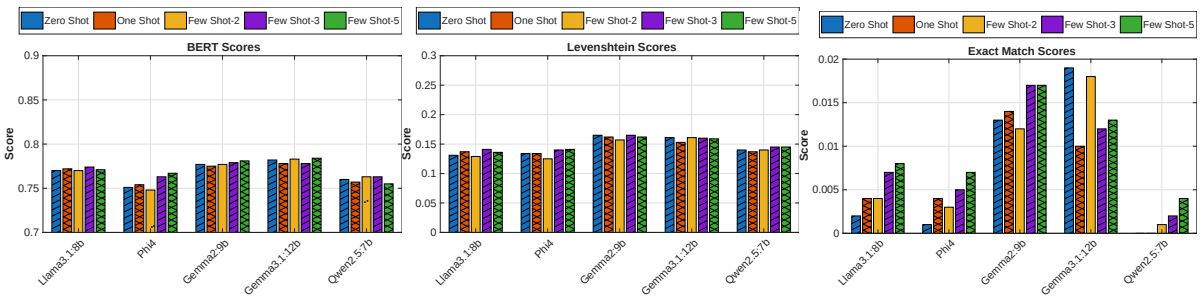
(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 3: Graphical Representation of the Metric Comparisons for Semantic Similar Selection without CoT



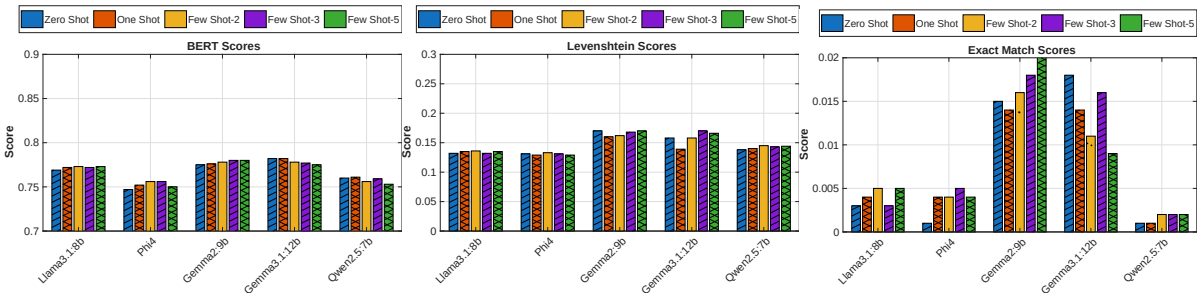
(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 4: Graphical Representation of the Metric Comparisons for Random Selection without CoT



(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 5: Graphical Representation of the Metric Comparisons for Semantic Similar Selection with CoT



(a) BERT Score (b) Levenshtein Score (c) Exact Match Score

Figure 6: Graphical Representation of the Metric Comparisons for Random Selection with CoT

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2025. Many-shot in-context learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Dheivegam. 2023. Tamiḻ viṭukataikaḷ 300.

FreshTamil.com. 2020. Tamiḻ viṭukataikaḷ | 100+ vidukathaigal in tamil | tamil riddles.

FreshTamil.com. 2024. Tamiḻ viṭukataikaḷ viṭaiyuṭaṉ 2024 | tamil vidukathaigal.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA. Association for Computational Linguistics.

Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. volume 83, page 101155.

Iyyanarithanar. 9th Century CE. *Purapporul Venbamaalai*. Classical Tamil grammatical treatise on the *puram* genre.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Thomas Lehmann. 1993. *A Grammar of Modern Tamil*. Pondicherry Institute of Linguistics and Culture, Pondicherry, India.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515, Online. Association for Computational Linguistics.

Pengxiao Lin, Zhongwang Zhang, and Zhi-Qin John Xu. 2025. Reasoning bias of next token prediction training.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

N. Manivasan. 2018. *500 Viṭukataikaḷum ataṉ aṟputa patilkaḷum*. aedahamlibrary.

Mullai Muthaiah. 1987. *1000 Viṭukataikaḷ*, 2 edition. New Century Book House Private Limited, 41-B. CITCO Industrial Estate, Chennai - 600098.

Ioannis Panagiotopoulos, George Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2025. RISCORE: Enhancing in-context riddle solving in language models through context-reconstructed example augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9431–9455, Abu Dhabi, UAE. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477, Vienna, Austria. Association for Computational Linguistics.

Harold F. Schiffman. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge University Press, Cambridge, UK.

Chuanqi Tan, Furu Wei, Li Dong, Weifeng Lv, and Ming Zhou. 2016. Solving and generating Chinese character riddles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 846–855, Austin, Texas. Association for Computational Linguistics.

Tolkāppiyar. Ancient. *Tolkāppiyam*. Earliest extant Tamil grammatical text, dating between 500 BCE–500 CE depending on tradition.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Vidukathaigal. 2025. Tamil molikal: Vidukathaigal.

Vinaval. 2025. Vidukathai vina vidaighal | vinaval.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fan Xu, Yunxiang Zhang, and Xiaojun Wan. 2023. Cc-riddle: A question answering dataset of chinese character riddles.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles.

Do Multimodal Large Language Models Have Good Taste? An Evaluation of MLLMs via the Visual Aesthetic Sensitivity Test

Yi Yao
City University of Macau
yiyao.tansy.gmail.com

Zhuang Qiu
City University of Macau
zhuangqiu.cityu.edu.mo

Abstract

The surge of large language models (LLMs) has sparked an ongoing debate on the extent to which LLMs emulate human cognition, and the emergence of multimodal large language models (MLLMs) opens a new window for benchmarking machine capabilities, including the processing of aesthetic input. Although MLLMs have the ability to process images, research focusing specifically on their ability to process abstract visual aesthetic input remains limited. In this study, we subjected five state-of-the-art MLLMs, namely, idefics2-8, llava-1.5-7b-hf, gemma-3n-E2B-it, moondream2, and llama-3.1-Nemotron-Nano-VL-8B-V1, to the Visual Aesthetic Sensitivity Test (VAST, Goetz et al., 1979) using zero-shot prompting to evaluate their ability to process abstract visual aesthetic stimuli. We found that MLLMs' processing of abstract visual aesthetic input is influenced by the interaction between prompt formulation and the model type. A noticeable gap exists between human and MLLMs responses in the VAST task, as reflected by their differing accuracy rates. Overall, while MLLMs show promising potential in aesthetic processing, their behavior differs noticeably from that of humans.

1 Introduction

In recent years, large language models (LLMs) have advanced at an unprecedented pace, achieving remarkable performance across a wide range of natural language understanding and generation tasks (Abe et al., 2024; Qin et al., 2024; Qiu et al., 2024; Wang et al., 2024). With the emergence of multimodal large language models (MLLMs), these systems have expanded into domains once regarded as uniquely human, including artistic creation and aesthetic evaluation (Hakopian, 2024; Yeo and Um, 2025; Khadangi et al., 2025). Although LLMs often generate outputs that appear human-like superficially, their underlying processing mechanisms may be fundamentally different from those of human cognition. There is an ongoing debate about whether these models truly emulate human-like thinking or simply mimic human patterns in response to prompts. Increasing amounts of research have tackled this debate head-on with an empirical approach, subjecting LLMs to many psychological experiments (Binz and Schulz, 2023; Kosinski, 2024; Cai et al., 2023; Qiu et al., 2025). For example, Binz and Schulz (2023) subjected GPT-3 to psychological experiments originally designed to study aspects of human cognition, such as decision-making, information search, and causal reasoning. They found that GPT-3 exhibited human-like or even better-than-human performance in tasks like gamble decisions and multiarmed bandit tasks, with signs of model-based reinforcement learning. In this study, we focused on another important aspect of human cognition: the ability to process the aesthetics of visual stimuli, such as appreciating the beauty of artworks, natural landscapes, and other visually pleasing patterns.

In philosophy and psychology, aesthetic judgment is a crucial subject (Beardsley, 1981; Martindale, 1988), and has been deeply explored by renowned figures like Immanuel Kant (Kant, 2024) and David Hume (Hume, 2017). In Critique of Judgment, Kant conceptualized the judgment of taste as the foundation of aesthetics, grounded in universal human sensibility. In contrast, Hume emphasized that aesthetics involves both cognitive and bodily factors. He suggested that aesthetic preferences are shaped by practice and by the ability to make nuanced comparisons. These philosophical perspectives laid the foundation for cognitive models of aesthetic judgment, aiming to explain how such judgments operate in human cognition. Building on these philosophical insights, Leder et al. (2004) introduced a cognitive model that outlines the five processes by which humans engage with aesthetics, including percep-

tual analysis, implicit memory activation, evaluative categorization, and emotional response and evaluation. This approach emphasizes the relationship between visual cues and learned cognitive responses, providing a framework for examining the processing of aesthetic stimuli. Further research shows that key features such as symmetry, complexity, and harmony are important determinants of aesthetic judgments in humans (Enquist and Arak, 1994). Eysenck (1940) proposed a new concept to explain the individual ability to make aesthetic judgment called aesthetic sensitivity. Aesthetic sensitivity refers to the ability to recognize and respond to subtle design elements that define beauty. These studies provide theoretical foundations for investigating AI systems that simulate human aesthetic perception.

Previous studies have explored the capability of MLLMs in visual recognition using large datasets of human labeled art (Huang et al., 2024; Fumanal-Idocin et al., 2023; Murray et al., 2012). For example, AesBench (Huang et al., 2024) offers a benchmark specifically designed to evaluate the ability of MLLMs to perceive and assess the aesthetic quality of images. However, all the above studies focus on concrete images, which limits their generalizability to highly abstract aesthetic content. In contrast, the Visual Aesthetic Sensitivity Test (VAST) aims to assess individual sensitivity to abstract visual forms. Over the years, VAST has been extensively used in human-centered studies to evaluate aesthetic preferences, demonstrating its reliability and validity as a measure of aesthetic sensitivity (Eysenck et al., 1984; Fróis and Eysenck, 1995; Myszkowski and Storme, 2017; Gear, 1986; Chan et al., 1980). Although extensively used in human-centered research, the potential application of VAST in artificial intelligence, especially in evaluating the aesthetic sensitivity of MLLMs, remains underexplored.

This study represents the first attempt to use well-validated psychological tests to investigate MLLMs' ability to process abstract visual input, offering new insights into whether AI systems can demonstrate abstract taste or aesthetic sensitivity. The research questions explored in this study are as follows:

- To what extent do MLLMs process abstract visual aesthetics in a way similar to humans?
- How do prompts influence MLLMs interpretation and evaluation of abstract visual aesthetics?

Model A	idefics2-8b
Developer	Hugging Face
Size	8B
Description	Instructional image-text model with OCR and visual reasoning ability.
Model B	llava-1.5-7b-hf
Developer	LLaVA Community
Size	7B
Description	Chat model based on LLaMA/Vicuna.
Model C	gemma-3n-E2B-it
Developer	Google
Size	∼2B effective
Description	Lightweight multimodal model for text, image, audio, video; 32K context.
Model D	moondream2
Developer	Vikhyat Korrapati
Size	∼1.9B
Description	Tiny, edge-optimized model for vision-language tasks like VQA and captioning.
Model E	llama-3.1-Nemotron-Nano-VL-8B-V1
Developer	NVIDIA
Size	8B
Description	Document-intelligent model with OCR, summarization, long-context support.

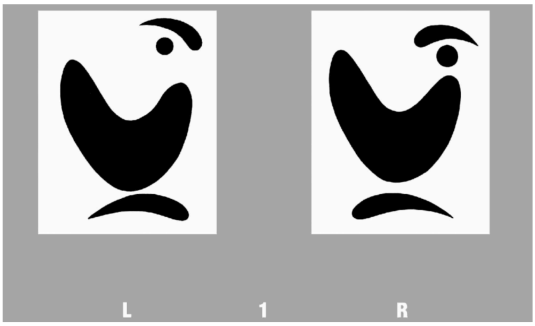
Table 1: Vertically arranged summary of the five MLLMs used in this study.

Our main contributions are as follows.

- Methodological Innovation: We introduce the first adaptation of the VAST for LLM-based evaluation, which enables the evaluation of aesthetic sensitivity in these models.
- Model Evaluation: We evaluate five MLLMs, including IDEFICS2-8, Llava-HF/Llava-1.5-7B-HF, Gemma-3N-E2B-IT, Moondream2, and Llama-3.1-Nemotron-Nano-VL-8B-V1, across 50 VAST items, and compared their responses with human normative data.
- Cognitive Insight: Building on Leder et al. (Leder et al., 2004), our approach provides new insights into how LLMs may simulate the cognitive processes behind aesthetic judgment.

2 Methodology

To address our research questions, we applied the visual aesthetic sensitivity test (VAST, Goetz et al., 1979) to obtain judgment data from five different MLLMs. The VAST is a reliable and well-validated instrument for assessing individual sensitivity to abstract visual forms. In the revised version of Goetz's study, participants viewed 50 pairs



There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.

Figure 1: An example VAST trial showing two abstract rooster-shaped images arranged side by side.

of non-representational pictures, of which one image in each pair had been intentionally altered by incorporating certain design faults. Participants were required to select the more aesthetically pleasing option, producing a quantitative score of individual aesthetic sensitivity. We selected five models for the VAST task based on their accessibility, cost efficiency, and ratings among the Hugging Face Image-Text-To-Text models. Table 1 provides a summary of the models included in our research. These models were accessed via the Hugging Face Hub and run on Colab GPUs.

We subjected each MLLM to the VAST task in a way similar to that of a human study. Each trial of the task represented a conversational session in which a PNG file containing two horizontally arranged pictures was presented, and the MLLM was prompted to select the better picture based on their visual aesthetic properties. Figure 1 illustrates an example trial of this experiment. In this trial, two abstract images that mimic the shape of a rooster were presented as visual input, while the text prompt read: 'There are two pictures in this image: the one on the left is labeled L and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice and justify it. Keep your response under 200 words.'

To explore our second research question, we designed five text prompts with slightly different wording and focus. These are listed in Table 2. Among them, Prompt 3 is adapted from the original formulation of the human study by Goetz et al. (1979, 796), which instructed participants to identify 'the most harmonious' design. Prompt 1 retains the original structure but removes the explicit reference to harmony, instead focusing on a general aesthetic comparison. Prompt 2 simplifies Prompt 1 by removing the requirement for picture descriptions. Prompt 4 replaces 'visual harmony' with 'aesthetic appeal' to shift the evaluative focus. Prompt 5 asks MLLMs to explicitly rely on their 'sensitivity to aesthetics', emphasizing subjective evaluation over descriptive analysis. We elicited the models' judgments in a zeroshot setting, where each model received only one text prompt and one PNG file in one conversational session. Each model analyzed all 50 VAST items five times, each with a different prompt. This led to 1250 pieces of model outputs from five models in total.

Prompt 1:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 2:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 3:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, and compare them in terms of their visual harmony. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 4:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, and compare them in terms of their aesthetic appeal. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 5:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, using your sensitivity to aesthetics to evaluate them. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.

Table 2: The five prompts used to instruct MLLMs during the VAST trials.

To code the model output, we adopted a meta-evaluation approach in which each raw MLLM response was judged by two separate LLMs, namely Mistral 7B Instruct and LLaMA 3 8B Instruct, to classify the choice as left (L), right (R), or unclear (N). The judging models' decisions were parsed from their outputs and recorded as L, R, or NA.

Model	Count	Accuracy
A	210	0.538
B	241	0.481
C	250	0.468
D	241	0.452
E	238	0.542

Table 3: Count and accuracy by model. A: idefics2-8b, B: llava-1.5-7b-hf, C: gemma-3n-E2B-it, D: moondream2, E: llama-3.1-Nemotron-Nano-VL-8B-V1

Prompt	Count	Accuracy
1	238	0.496
2	238	0.487
3	235	0.506
4	233	0.476
5	236	0.508

Table 4: Count and accuracy by prompt.

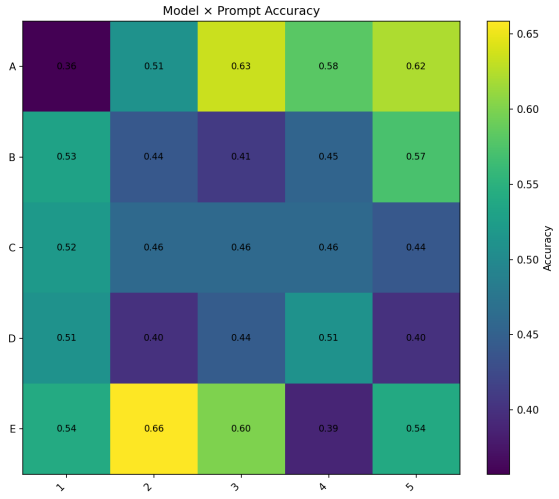


Figure 2: Heatmap of Accuracy for ModelPrompt Pairs

Empty or invalid responses were also coded as NA. The coded results, along with the original responses, were stored for further analysis. Disagreements between the two judging models were explicitly marked and resolved by two human researchers. The script and data used for the statistical analysis in the following section are publicly available via GitHub[1]

3 Results

Of the 1250 model responses to the VAST task, 70 responses were discarded due to the absence of an explicit choice between the right and left picture, leaving 1180 valid data points for analysis. Across all valid trials, the overall accuracy was 49.5%. Accuracy varied across models, ranging from 45.2% for Model D (moondream2) to 54.2% for Model E (llama-3.1-Nemotron-Nano-VL-8B-V1). Model A (idefics2-8b) and Model E achieved the highest accuracies, while Model D and Model C (gemma-3n-E2B-it) were the least accurate (see Table 3). Accuracy by prompt was relatively stable, ranging from 47.6% (Prompt 4) to 50.8% (Prompt 5)(see Table 4). The model-prompt breakdown revealed considerable variability. As shown in Figure 2, Model A performed best with Prompt 3 (63.4%) but poorly with Prompt 1 (35.7%). Conversely, Model E achieved its highest accuracy with Prompt 2 (65.9%) but its lowest with Prompt 4 (38.8%).

To assess the influence of model and prompt on accuracy, we fit a mixed-effects logistic regression model with random intercepts by image. The model included dummy-coded predictors for the model (reference: Model A) and prompt (reference: Prompt 1)[2]. Results from the GLMM showed that, relative to Model A, Model B ($\beta = 0.32$, 95% CI $= [0.06, 0.58]$), Model C ($\beta = 0.27$, 95% CI $= [0.01, 0.52]$), and Model E ($\beta = 0.38$, 95% CI $= [0.11, 0.64]$) had significantly higher accuracy. Model D did not differ significantly from Model A (95% CI $= [-0.03, 0.49]$). An omnibus Wald test showed that while the main effect of the model ($p = 0.04$) and the interaction effect ($p < 0.0001$) were significant, the main effect of prompt did not reach significance ($p > 0.05$). The comparison between model accuracy and human accuracy on the VAST task is shown in Table 5.

4 Discussion

This study investigated whether MLLMs possess an aesthetic sensitivity comparable to that of humans. We adopted a cognitive test originally designed to measure human visual aesthetic sensitivity (VAST, Goetz et al., 1979), and subjected MLLMs to the test. Our analysis revealed some noticeable response patterns across five leading MLLMs under five distinct prompt conditions. First, the result was significantly influenced by the interaction between prompt formulation and

[1] https://github.com/PON2020/vast

[2] Following a reviewer's suggestion, we constructed another GLMM, adding the interaction between the model and prompt. We confirmed the significance of the interaction effect; however, a significant interaction generally makes the main effects difficult to interpret. We included the script and results of that GLMM in the published GitHub repository for readers who are interested.

System	Mean Accuracy	Range
Human	0.6 (children), 0.7 (adults)	–
Model A	0.54	0.36–0.63
Model B	0.48	0.41–0.57
Model C	0.47	0.44–0.52
Model D	0.45	0.40–0.51
Model E	0.54	0.39–0.66

Table 5: Comparison of human and model accuracy on the VAST task. The human data was reported in Goetz et al. (1979). Range refers to the lowest and highest accuracy achieved. Model A: idefics2-8b, Model B: llava-1.5-7b-hf, Model C: gemma-3n-E2B-it, Model D: moondream2, Model E: llama-3.1-Nemotron-Nano-VL-8B-V1

the model type. This pattern indicates that current models may prioritize surface-level linguistic cues rather than engage in deeper forms of aesthetic reasoning. Second, we observed a significant variance in prompt robustness across models. Idefics2-8b and llama-3.1-Nemotron-Nano-VL-8B-V1 exhibited the highest variability across prompt conditions, suggesting a weaker generalization ability in aesthetic tasks. In contrast, Gemma-3n-E2B-it maintained relatively consistent performance, implying more stable internal representations. Third, the overall aesthetic judgment capabilities of MLLMs remained below human baselines. As shown in Table 5, human participants in the VAST task achieved an average accuracy of 0.7 for adults and 0.6 for children (Goetz et al., 1979). Although there was a clear variability among participants, human mean performance remained noticeably above chance, indicating that most human participants could perform the task with a fair degree of reliability. In contrast, our best-performing models reached a mean accuracy of 0.54, falling short of human means and never approaching the highest human scores (about 80% accuracy). This suggests that current MLLMs still lack the nuanced processing required for abstract visual evaluation.

Our findings are consistent with previous studies such as AesBench (Huang et al., 2024), showing a noticeable difference between MLLMs and human performance in aesthetic tasks. While AesBench evaluated models' performance in labeling and rating realistic photographs, our study leveraged the VAST framework, focusing on the cognitive ability of processing the aesthetics in abstract shapes. The observed differences between humans and MLLMs in the VAST could be explained by the cognitive model of aesthetic appreciation proposed by Leder et al. (2004), which describes aesthetic judgment as a process involving multiple stages. Human observers engage in this process by combining perceptual experience with cultural priors, personal background, and memory-based associations; by contrast, the processing of aesthetic images by MLLMs relies on statistical regularities of the input without accessing sensory or other embodied experiences. This distinction explains the patterns we observed in this study and offers valuable insight into ongoing debates on whether LLMs exhibit human-like cognitive mechanisms.

In conclusion, MLLMs processing of abstract visual aesthetic input is significantly influenced by the interaction between prompt formulation and the model type. A noticeable gap exists between human and MLLM responses in the processing of abstract visual input, as reflected in their different accuracy rates in the VAST task. In general, while MLLMs show promising potential in aesthetic processing, current MLLMs' behavior differs noticeably from that of humans.

4.1 Limitations and Future Work

We adopted the original VAST (Goetz et al., 1979) as a benchmark of aesthetic processing because of its standardized format and demonstrated reliability and validity in previous research. However, the normative human data was collected in 1979, raising concerns about its ability to reflect a contemporary aesthetic landscape. To address these gaps, future research should collect up-to-date human data to capture mainstream aesthetic preferences and then compare contemporary human responses with model data. It is also informative to test whether human responses are prompt-dependent, which allows researchers to claim more confidently whether the effect of the prompt in this study is unique to models or shared with humans. Last but not least, it is worth exploring the performance of frontier closed-source models (e.g., GPT-4V) on the same task to learn about upper bounds and transferability.

References

Yoshia Abe, Tatsuya Daikoku, and Yasuo Kuniyoshi. 2024. Assessing the aesthetic evaluation capabilities of gpt-4 with vision: Insights from group and individual assessments. In *38 (2024)*, pages 2Q1IS301–2Q1IS301. .

Monroe C Beardsley. 1981. *Aesthetics, problems in the philosophy of criticism*. Hackett Publishing.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Zhenguang G Cai, Xufeng Duan, David A Haslett, Shuqi Wang, and Martin J Pickering. 2023. Do large language models resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

J Chan, Hans J Eysenck, and Karl O Götz. 1980. A new visual aesthetic sensitivity test: Iii. cross-cultural comparison between hong kong children and adults, and english and japanese samples. *Perceptual and Motor Skills*, 50(3_suppl):1325–1326.

Magnus Enquist and Anthony Arak. 1994. Symmetry, beauty and evolution. *Nature*, 372(6502):169–172.

Hans J Eysenck, KO Götz, Han Yee Long, DKB Nias, and M Ross. 1984. A new visual aesthetic sensitivity testiv. cross-cultural comparisons between a chinese sample from singapore and an english sample. *Personality and Individual Differences*, 5(5):599–600.

Hans Jurgen Eysenck. 1940. The general factor in aesthetic judgements 1. *British Journal of Psychology. General Section*, 31(1):94–102.

João Pedro Fróis and Hans J Eysenck. 1995. The visual aesthetic sensitivity test applied to portuguese children and fine arts students. *Creativity Research Journal*, 8(3):277–284.

Javier Fumanal-Idocin, Javier Andreu-Perez, Oscar Cordón, Hani Hagras, and Humberto Bustince. 2023. Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *IEEE Transactions on Fuzzy Systems*, 32(4):1915–1926.

Jane Gear. 1986. Eysenck's visual aesthetic sensitivity test (vast) as an example of the need for explicitness and awareness of context in empirical aesthetics. *Poetics*, 15(4-6):555–564.

Karl O Goetz, R Lynn, A. R. Borisy, and Hans J Eysenck. 1979. A new visual aesthetic sensitivity test: I. construction and psychometric properties. *Perceptual and Motor Skills*, 49(3):795–802.

Mashinka Firunts Hakopian. 2024. Art histories from nowhere: on the coloniality of experiments in art and artificial intelligence. *AI & SOCIETY*, 39(1):29–41.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

David Hume. 2017. Of the standard of taste. In *Aesthetics*, pages 483–488. Routledge.

Immanuel Kant. 2024. *Critique of judgment*, volume 10. Minerva Heritage Press.

Afshin Khadangi, Amir Sartipi, Igor Tchappi, and Gilbert Fridgen. 2025. Cognartive: Large language models for automating art analysis and decoding aesthetic elements. *arXiv preprint arXiv:2502.04353*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508.

Colin Martindale. 1988. Aesthetics, psychobiology, and cognition. In Frank H. Farley and Ronald W. Neperud, editors, *The Foundations of Aesthetics, Art, and Art Education*, pages 7–42. Praeger, New York.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE.

Nils Myszkowski and Martin Storme. 2017. Measuring good taste with the visual aesthetic sensitivity test-revised (vast-r). *Personality and Individual Differences*, 117:91–100.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Zhuang Qiu, Xufeng Duan, and Zhenguang G Cai. 2025. Grammaticality representation in chatgpt as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1):1–15.

Zhuang Qiu, Peizhi Yan, and Zhenguang Cai. 2024. Large language models for second language english writing assessments: An exploratory comparison. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 363–370.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

Yunha Yeo and Daeho Um. 2025. Can ai recognize the style of art? analyzing aesthetics through the lens of style transfer. *arXiv preprint arXiv:2504.14272*.

L3Cube-MahaEmotions: A Marathi Emotion Recognition Dataset with Synthetic Annotations using CoTR prompting and Large Language Models

Nidhi Kowtal[1], Raviraj Joshi[2,3]
[1] Pune Institute of Computer Technology, Pune, Maharashtra India
[2] Indian Institute of Technology Madras, Chennai, Tamil Nadu India
[3] L3Cube Labs, Pune
{kowtalnidhi, ravirajoshi}@gmail.com

Abstract

Emotion recognition in low-resource languages like Marathi remains challenging due to limited annotated data. We present L3Cube-MahaEmotions, a high-quality Marathi emotion recognition dataset with 11 fine-grained emotion labels. The training data is synthetically annotated using large language models (LLMs), while the validation and test sets are manually labeled to serve as a reliable gold-standard benchmark. Building on the MahaSent dataset, we apply the Chain-of-Translation (CoTR) prompting technique, where Marathi sentences are translated into English and emotion labeled via a single prompt. GPT-4 and Llama3-405B were evaluated, with GPT-4 selected for training data annotation due to superior label quality. We evaluate model performance using standard metrics and explore label aggregation strategies (e.g., Union, Intersection). While GPT-4 predictions outperform fine-tuned BERT models, BERT-based models trained on synthetic labels fail to surpass GPT-4. This highlights both the importance of high-quality human-labeled data and the inherent complexity of emotion recognition. An important finding of this work is that generic LLMs like GPT-4 and Llama3-405B generalize better than fine-tuned BERT for complex low-resource emotion recognition tasks. The dataset and model are shared publicly at https://github.com/l3cube-pune/MarathiNLP.

1 Introduction

Recent advances in NLP have mainly benefited high-resource languages like English and Chinese, which have ample data and annotations (Thabah and Purkayastha, 2021). Low-resource languages, however, face challenges due to limited high-quality data and complex grammar, leading to poor model performance (Yang et al., 2023). Even multilingual LLMs, effective in translation, struggle with direct prompts in these languages (Luong et al., 2023; Xiang Zhang, 2023). We focus on Marathi, spoken by about 83 million people, which remains underrepresented in NLP due to scarce tools and datasets (Joshi, 2022b; Narzary et al., 2022; Joshi, 2022a). Its syntactic complexity adds to the modeling difficulty (Luong et al., 2023).

To overcome these challenges, we created an emotion classification dataset for Marathi by leveraging the capabilities of large language models like GPT-4 and Llama3-405B. A key limitation in emotion classification for Marathi is the lack of labeled emotional datasets. Manual labeling is expensive and time-consuming, which makes progress in low-resource languages slower. To address this, we combined manual validation with annotation using LLMs to produce a high-quality dataset efficiently. We manually labeled the validation and test sets to ensure a gold-standard benchmark and also annotated these sets using GPT-4 and Llama3-405B to evaluate their performance. Since GPT produced more accurate results, we used it to annotate the training set as well, accelerating progress for Marathi NLP through the strategic use of LLMs.

Interestingly, we observe that GPT-4 significantly outperforms BERT-based models trained on its own generated labels. This indicates that fine-tuning smaller models on noisy or automatically annotated data does not necessarily lead to better performance than the original LLM. The results underscore the inherent complexity of multi-label emotion recognition—an intricate task where generic

LLMs like GPT-4 are better equipped to capture subtle emotional cues than fine-tuned, smaller models. This contrasts with findings from (Jadhav et al., 2024), where BERT models trained on clean, high-quality data were shown to outperform LLMs in low-resource scenarios. In our case, the presence of residual noise in the training labels or the complexity of the task itself likely hinders the ability of BERT-based models to generalize effectively.

We use a prompting technique called Chain-of-Translation Prompting (CoTR) to improve the quality of emotion annotation for a low-resource language like Marathi (Deshpande et al., 2024). The CoTR approach, illustrated in Figure 1, has been shown to outperform standard prompting strategies, and is adopted in this study for its effectiveness. Given the scarcity of Marathi training data, LLMs may struggle to accurately predict emotion labels directly from Marathi sentences. To address this, we translate Marathi inputs into English and then generate emotion labels using the translated text. This enables LLMs to leverage their stronger English language understanding. CoTR leads to more reliable emotion classification while preserving the intent of the original Marathi content. We independently validate its effectiveness on the MahaEmotions dataset.

The main contributions of this work are as follows:

- We curate MahaEmotions[*†], a new Marathi Emotion Classification dataset annotated with eleven emotion categories, containing both model-generated and human annotated labels to ensure good annotation quality. The dataset consists of (12k, 1.5k, 1.5k) train, test, and validation samples, respectively.

- We use Chain-of-Translation (CoTR) as an effective prompting strategy to use multilingual LLMs for emotion tagging. Instead of direct categorization in Marathi, CoTR translates Marathi input into English before labeling the data, considerably enhancing the tagging accuracy. Notably, we observe an absolute 6% improvement in the GPT-4 performance using CoTR prompting.

- We benchmark the performance of multiple models on this task, including GPT-4, LLaMA3-405B, and a fine-tuned MahaBERT-V2 model. Our results show that GPT-4 outperforms LLaMA-3, which in turn outperforms fine-tuned MahaBERT-V2, both in terms of accuracy and F1-score.

- We manually annotate a high-quality test set to evaluate how the LLMs perform on the tagging task.

2 Related Work

Low-resource languages have consistently faced challenges in NLP due to the lack of sufficient linguistic resources, standardized benchmarks, and annotated corpora. As a result, they remain significantly underrepresented in mainstream NLP research (Alexandre Magueresse, 2020). The emergence of multilingual pretrained language models has helped address some of these issues through cross-lingual transfer, enabling better performance across languages with limited data.

Multilingual architectures such as mBERT, mT5, and XLM-R have shown reasonable zero-shot and few-shot performance on downstream tasks in low-resource settings (Luong et al., 2023; Kelechi Ogueji, 2021). Prompt-based learning techniques have also proven effective in adapting pretrained models to new tasks, particularly in scenarios where task-specific fine-tuning is not feasible due to data scarcity (Yang et al., 2023).

Recent advances like L3Cube-MahaNLP and MahaBERT have boosted research in syntactic parsing, classification, and sentiment analysis for Marathi by providing large monolingual datasets and transformer models (Joshi, 2022b,a; Pingle et al., 2023; Kulkarni et al., 2021; Velankar et al., 2022). However, there's still limited work on deeper tasks like emotion recognition. Marathi's complex grammar

[*] https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaEmotions

[†] https://huggingface.co/l3cube-pune/marathi-emotion-detect

and differences from English make it hard to directly apply models trained on high-resource languages. Similar trends appear in other low-resource languages like Khasi, where encoder-decoder transformers have improved tasks like translation despite limited data (Thabah and Purkayastha, 2021).

In the broader NLP community, there has been growing interest in emotion recognition, especially in the context of multilingual and multimodal systems. However, similar advancements for Marathi are still quite limited. In comparison, significant progress has been made for Hindi and Hindi-English code-mixed text, with several emotion classification models and datasets available (Kumar and andf Girish Sharma, 2023; Anshul Wadhawan, 2021; Singh et al., 2022). A good example is the EmoInHindi corpus, a low-resource benchmark that provides multi-label emotion annotations along with dialogue-level context (Singh et al., 2022). Recent surveys also highlight the importance of using customized model architectures, cross-lingual transfer, and domain adaptation techniques for improving emotion classification in low- and mid-resource languages (Shabnam Tafreshi, 2024).

Additionally, research on LLMs' reliability for non-English inputs is ongoing. Zhang et al. (Xiang Zhang, 2023) critically assess GPT-4 and other LLMs, showing performance drops for underrepresented, morphologically rich languages like Marathi. This questions whether such models can be directly used for low-resource emotion recognition without translation or augmentation. Prompt engineering adapted to linguistic traits (Patel et al., 2024) offers a practical way to overcome these limits. In multilingual contexts, translation-based prompting notably improves semantic understanding and emotion consistency.

In this study, we expand on these discoveries and provide a Chain-of-Translation (CoTR) prompting architecture for Marathi text emotion recognition that makes use of multilingual language models (Deshpande et al., 2024). Our method uses a single prompt that first translates the Marathi sentence into English and then predicts the emotion using English-based prompt templates.

3 Methodology

Our methodology involves curating a high-quality Marathi emotion dataset, applying Chain-of-Translation (CoTR) prompting for emotion tagging using both human annotators and multilingual LLMs, and training a classifier on the annotated data. As shown in Figure 2, the process includes dataset preprocessing, CoTR-based emotion labeling, model comparisons, and final evaluation using standard classification metrics.

3.1 Dataset Description

For this study, we have used the publicly available L3Cube's MahaSent-GT dataset (Joshi, 2022b), a sentiment analysis corpus in Marathi. The dataset contains textual content primarily sourced from Twitter. Each sentence is originally labeled with sentiment (Positive, Negative, Neutral), and we extend this dataset by introducing emotion labels. The dataset contains a total of 15,000 Marathi sentences. It provides a suitable foundation for emotion classification tasks due to its coverage of real-world, emotion-rich textual inputs. The distribution of emotion labels across the train, validation, and test sets, along with example sentences, is shown in Table 1.

3.2 Emotion Label Taxonomy and Annotation Scheme

We utilize a fixed set of eleven basic emotion labels: *Happiness, Sadness, Anger, Fear, Surprise, Disgust, Excitement, Pride, Respect, Sarcasm*, and *Neutral*. A careful selection process was used to ensure that this set of emotions was both simple enough to allow for consistent classification over a large number of phrases and expressive enough to reflect a wide range of sentiments.

A label is assigned to each sentence in the dataset according to the primary emotion it conveys. Although this set of emotion labels is based on popular psychological models like Ekman's basic emotions and Plutchik's emotion wheel, it is a simplified version made for practical use. Marathi is a diverse language, with many emotional states that are hard to define in a pre-defined set of

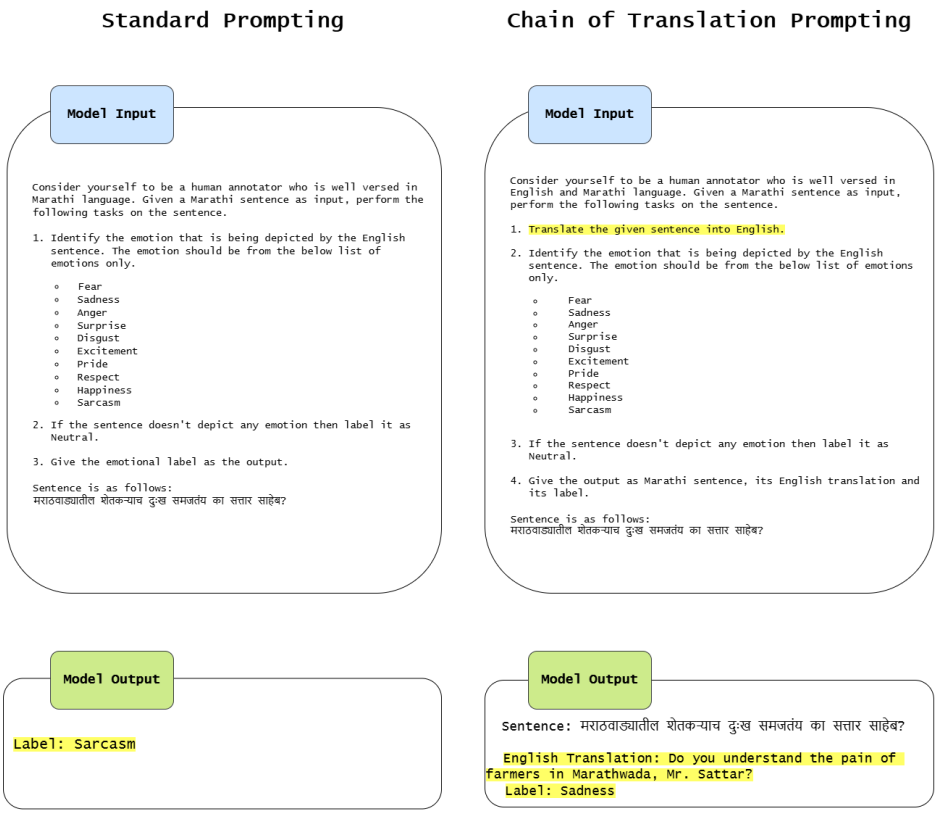


Figure 1: Prompt used in Chain of Translation Prompting (CoTR)

Figure 2: Emotion Tagging using Human and LLMs (CoTR)

categories. We chose to concentrate on a more manageable and useful set of labels. After carefully reviewing the dataset and performing manual analysis and preprocessing, we selected emotion categories that were most frequently observed in day-to-day usage and could be annotated consistently at scale.

Since there are not many extensive emotion datasets for Marathi, we used this set of eleven emotions to enable consistent and scalable annotation. Both language models and human annotators benefit from this fixed list since it helps them concentrate on distinct, non-overlapping categories. The primary emotion that each sentence in the sample conveys is labeled. The strongest or most obvious emotion is selected when a text comprises multiple emotions.

3.3 Prompt-Based Annotation Strategy

We designed a structured prompt to guide the language model during tagging. Since most large language models (LLMs) are trained primarily on English, we include translation in the same prompt. The Marathi sentence is first translated to English, and then the model predicts the emotion from a predefined set of categories: *Fear, Sadness, Anger, Surprise, Disgust, Excitement, Pride, Respect, Happiness, Sarcasm*, and *Neutral*. If a sentence contains more than one emotion, the most prominent one is assigned to it. If no emotion is clearly expressed, the sentence is labeled as *Neutral*.

3.4 Models Used

1. **GPT-4o:**
   GPT-4o is developed by OpenAI, with 1.8 trillion parameters (unofficial). It is a closed-source model and accessible through APIs provided by OpenAI. GPT-4o builds on the advancements of its previous versions, offering enhanced capabilities in natural language understanding, generation, and reasoning across a wide range of tasks.

2. **Llama 3.1 405B:**
   Llama 3.1 (Large Language Model for Multilingual Applications) is the third iteration in the Meta Llama series, designed with multiple variants, including a 405 billion parameter version and an 8 billion parameter version. These models are typically open-source. Llama3 models are optimized for multilingual tasks, incorporating vast and diverse datasets to improve performance across different languages.

3. **MahaBERT-V2:**
   MahaBERT-V2 is a transformer-based language model pre-trained specifically on a large corpus of Marathi text. It captures rich morphological and syntactic patterns of the Marathi language, making it well-suited for downstream NLP tasks in Marathi. Despite being domain-specific, its performance on emotion classification was moderate, with an accuracy of 63% and an F1 score of 0.47.

4. **MuRIL:**
   MuRIL (Multilingual Representations for Indian Languages) is a multilingual BERT model developed by Google, trained on 17 Indian languages including Marathi. It supports both transliterated and native scripts, and enables zero-shot and multilingual transfer learning. In our experiments, MuRIL achieved an accuracy of 60% and an F1 score of 0.42, slightly underperforming compared to MahaBERT-V2, likely due to its generalization across many languages rather than specialization in Marathi.

4 Results

4.1 Gold Test Set

We manually annotated test and validation sets containing 1500 sentences each. These human annotations are treated as the ground truth for evaluating model performance.

4.2 GPT-4 vs Llama3-405B

We evaluated the performance of GPT-4 and Llama-405B by prompting each model individually to classify the same set of sentences. The classification was performed after translating the Marathi inputs into English. A model's prediction was considered correct only if it matched the human-provided label.

We considered multiple evaluation scenarios:

Emotion	Train Set	Validation Set	Test Set	Example Sentence
Neutral	3903	546	499	अजूनही कारवाही झालेली नाही. काम सुरूच आहे.
Anger/Disgust	3374 (1257/2117)	398 (228/170)	405 (216/189)	कधी आणि कशी देणार सरस्वती ह्याला सद्बुद्धी? हा अशिक्षित भाजीवाला तिला मानतच नाही ना!
Happiness	1295	201	140	कोजागरी पौर्णिमेच्या सर्व नागरिकांना हार्दिक शुभेच्छा..!
Respect	1240	104	147	कर्तव्य बजावत असताना शहीद झालेल्या हुतात्म्यांना पोलीस स्मृती दिनानिमित्त शतशः नमन..!!
Pride	662	68	76	या सर्वांचे राष्ट्रवादी काँग्रेस मध्ये मनःपूर्वक स्वागत, गर्व आहे आम्हाला
Sadness	499	75	81	या' अभिनेत्रींनी कमी वयातच केली आत्महत्या...
Surprise	364	48	85	भाजपाचे सरकार असतेवेळी असे कसे झाले ?
Excitement	299	32	44	आहे हा चित्रपट पाहण्यासाठी मी खूप उत्सुक आहे
Fear	185	20	13	खरच लक्ष द्या नाहीतर खूप मोठे संकट आ वासून उभा आहे
Sarcasm	177	8	10	आणि पवार साहेबांनी काँग्रेस फोडली ती काय चण्याक्याणिती
Total Samples	11998	1500	1500	

Table 1: Number of samples per emotion label in the train, validation, and test sets, along with example sentences

Statistics	Validation Set	Test Set
GPT-4 Correct	1265	1284
Llama Correct	962	1051
Llama Correct, GPT-4 Incorrect	106	100
GPT-4 Correct, Llama Incorrect	409	333
Both Correct	856	951
At Least One Correct	1371	1384
Both Incorrect	129	116

Table 2: Model performance statistics for validation and test sets (each containing 1500 sentences)

- **Correct Prediction:** The model label matches the human-annotated gold label.
- **Disagreement Resolution:** If both models gave labels different from the human label, no credit was given to either.
- **Overlap Analysis:** We analyzed agreement and disagreement patterns, including cases where both models were correct, only one was correct, or both were incorrect.

Based on the comparative analysis of these models, we found GPT-4 to be the more consistent and accurate model. Since the performance of GPT-4 alone was comparable to the combination of GPT-4 and Llama3-405B, we chose GPT-4 for the large-scale annotation of the training dataset.

We evaluated the performance of GPT-4 and Llama3-405B on both the validation and test datasets, each consisting of 1500 Marathi sentences. Table 2 summarizes the correctness statistics across both models.

Model	Accuracy	Precision	Recall	F1 Score
MahaBERT-V2	0.63	0.65	0.62	0.64
MuRIL	0.59	0.62	0.59	0.60
GPT-4	0.83	0.85	0.82	0.83
GPT-4 (CoTR)	**0.86**	**0.88**	**0.85**	**0.86**
Llama3-405B (CoTR)	0.70	0.74	0.70	0.72

Table 3: Evaluation metrics for different models on the MahaEmotions test set (Weighted metrics). Note that Anger and Disgust are merged into a single class during both training and evaluation.

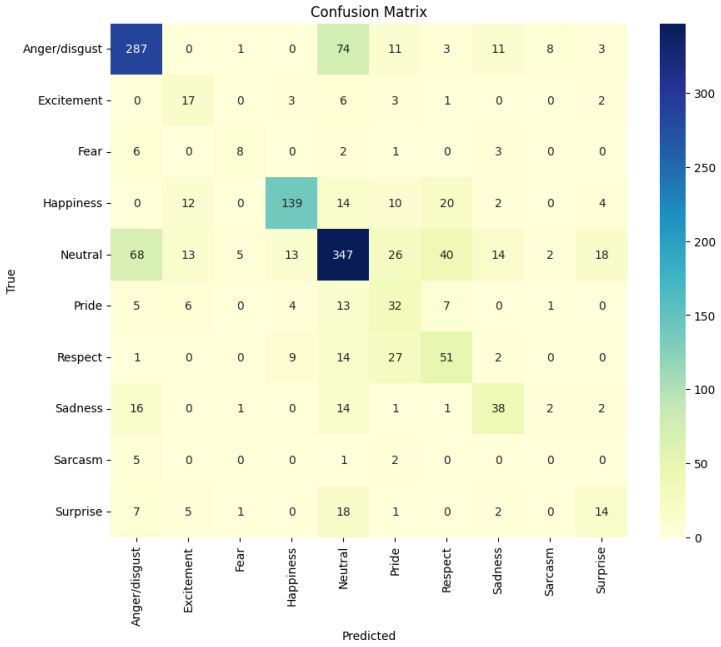


Figure 3: Confusion matrix for MahaEmotions classification task using L3Cube's MahaBERT-V2

GPT-4 consistently outperformed Llama in both validation and test set. On the test set, GPT-4 correctly classified 1284 sentences, while Llama correctly classified 1051. GPT-4 showed better accuracy, with 333 instances where GPT-4 was correct and Llama was incorrect, compared to only 100 instances where Llama was correct and GPT-4 was wrong.

Given the OR of both models' predictions is similar with GPT-4's performance (1384 for OR vs. 1284 for GPT), we decided to tag the training data exclusively using GPT-4 for the final classifier model.

After annotation, we trained a classifier on the GPT-labeled dataset. The overall performance of this classifier on the test set is shown in Table 3. It achieved an accuracy of 63%, with a precision of 0.65, recall of 0.62, and F1 score of 0.64. The detailed confusion matrix is presented in Figure 3, which shows the classification behavior across all emotion categories

The confusion matrix shows that all emotion categories are sometimes predicted as *Neutral*. This is expected, as many Marathi sentences have emotions that are expressed in a very subtle way, making them harder for the model to detect. In such cases, the model often chooses the Neutral label. This problem is common in low-resource languages, where emotions depend more on cultural and contextual clues than on clear emotional words.

There are also some clear patterns of confusion between emotions that are similar in meaning. For example, *Pride* and *Respect* are often mixed up. In Marathi, pride is often expressed with respectful language, and respectful statements can sound like pride. Similarly, *Happiness* and *Excitement* are confused with

each other because they are both positive emotions, with the main difference being the level of intensity. *Fear* and *Surprise* are also mixed up, as both can be caused by unexpected events.

In some cases, *Sarcasm* is classified as *Anger/Disgust*, which makes sense because sarcasm can carry a tone of irritation or contempt. *Sadness* is sometimes labeled as *Neutral* when expressed in a mild way, and as *Anger/Disgust* when it includes frustration. These patterns show two main challenges: the tendency of the *Neutral* class to attract unclear cases, and the difficulty of separating emotions that are similar in meaning or context. Better use of context and targeted data augmentation could help improve the model's performance in these cases.

4.3 Chain of Translation Prompting (CoTR) vs Non-CoTR Approach

As shown in Table 3, using CoTR leads to consistent improvements in accuracy, precision, recall, and F1 score. By translating Marathi inputs into English, multilingual LLMs can more effectively apply their English-language capabilities, enhancing emotion classification performance in low-resource languages like Marathi.

5 Limitations

One limitation of our work is that we used large language models (LLMs) that are mostly trained on English or multilingual data, not specifically on Marathi. Because of this, the models may not fully understand the deeper meanings or cultural context in Marathi sentences.

Another limitation is that our dataset has fewer examples of rare or complex emotions, which makes it harder for the model to learn and predict such emotions correctly. Emotions like गहिवर (Emotional overwhelm) and कातरता (Gentle sorrow) are especially difficult to label consistently.

6 Future Work and Conclusion

In this work, we focused on the task of emotion classification for Marathi, a low-resource language. We created a high-quality dataset by combining predictions from large language models (LLMs) like GPT-4 and Llama-405B with manual checks. To improve accuracy, we used a method called Chain-of-Translation (CoTR), where Marathi sentences were first translated to English before labeling. GPT-4 showed consistent and reliable results, which made it suitable for large-scale annotation.

In the future, we plan to train LLMs using more Marathi-specific emotion data. This will help the models better understand the language and its emotional tone. We also want to include more sentences that show complex and subtle emotions, such as निराशा (Disappointment), गहिवर, and कातरता.

We also aim to test our Chain-of-Translation (CoTR) method on more LLMs such as Gemma, Grok, DeepSeek, and Mistral, to see how well it works with other models.

Acknowledgments

This work was done under the mentorship of Mr. Raviraj Joshi (Mentor, L3Cube Pune). I would like to express our gratitude towards him for his continuous support and encouragement.

References

Evan Heetderks Alexandre Magueresse, Vincent Carles. 2020. Low-resource languages: A review of past work and future challenges.

Akshita Aggarwal Anshul Wadhawan. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. In *Computation and Language*.

Dr.R.R.Deshmukh Bharati Borade. 2023. Emotional speech recognition for marathi language.

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages. *arXiv preprint arXiv:2409.04512*.

Kishor Bhangale; Dipali Dhake; Rupali Kawade; Triveni Dhamale; Vaishnavi Patil; Nehul Gupta. 2023. Deep learning-based analysis of affective computing for marathi corpus. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On limitations of llm as annotator for low resource languages. *arXiv preprint arXiv:2411.17637*.

Charibeth Cheng Jan Christian Blaise Cruz. 2020. Establishing baselines for text classification in low-resource languages.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Jimmy Lin Kelechi Ogueji, Yuxin Zhu. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. *ACL Anthology*.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. 2014. Emotion recognition and its applications. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Tapesh Kumar and Mehul Mahrishi andf Girish Sharma. 2023. Emotion recognition in hindi text using multilingual bert transformer.

Minh-Thang Luong, Quoc V. Le, and Thang Luong. 2023. Multilingual neural machine translation with a special focus on low-resource languages. *Transactions of the Association for Computational Linguistics (TACL)*.

Sanjib Narzary, Maharaj Brahma, and Mwnthai Narzary. 2022. Generating monolingual dataset for low resource language bodo from old books using google keep. In *Proceedings of ACL*.

Krish Patel, Gaurav Keshari, Dhaval Powle, Saad Ansari, Tejaswini Chavan, and Anindita Khade. 2024. Hybrid nlp model for multilingual sentiment and emotion analysis in poetry. In *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, pages 1–8. IEEE.

Pravin K. Patil and Satish R. Kolhe. 2024. Sarcasm detection for marathi and the role of emoticons.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 274–281.

Mona Diab Shabnam Tafreshi, Shubham Vatsal. 2024. Emotion classification in low and moderate resource languages. *arxiv*.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. *LREC 2022*.

N. Donald Jefferson Thabah and Bipul Syam Purkayastha. 2021. Low resource neural machine translation from english to khasi: A transformer-based approach. In *Low Resource Neural Machine Translation from English to Khasi: A Transformer-Based Approach*.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. *Aggression and Cyberbullying (TRAC 2022)*, page 1.

Bradley Hauer Xiang Zhang, Senyu Li. 2023. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yuqing Yang, Jie Fu, and Pascal Poupart. 2023. Prompt learning for low-resource language understanding with pretrained models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Exploring Synonymy Representation in Large Language Models: A Comparative Analysis with Referential and Use-Based Lexical Resources

Sara Besharati
Université du Québec à Montréal (UQAM)
Montréal, Québec, Canada
besharati.sara@courrier.uqam.ca

Abstract

This study investigates how synonymy is represented in Large Language Models (LLMs) and examines the extent to which their representations align with human intuitions about context-dependent and use-based synonymy, in comparison to traditional lexical resources like WordNet. Focusing exclusively on English verbs, we utilized sentences from the Concepts in Context (CoInCo) dataset to generate substitute candidates from three sources: LLM-generated synonyms (RoBERTa), gold use-based synonyms (CoInCo), and referential synonyms (WordNet). Substitutions were evaluated by calculating cosine similarity scores within the T5 model's embedding space to assess semantic alignment.

The results show that RoBERTa (mean 0.656) and CoInCo (mean 0.625) exhibited the highest mean similarity scores to the target verbs, indicating that contextual modeling is highly effective in capturing fine-grained, context-dependent aspects of verb meaning. Conversely, WordNet (mean 0.548) provided a more conservative distribution. Analysis of the overlap among the top four candidate synonyms from each source revealed minimal convergence. The agreement between WordNet and RoBERTa was particularly low, with an average overlap of only 4.94%. This lack of convergence demonstrates that the methods generate largely distinct sets and offer complementary rather than redundant lexical knowledge.

Furthermore, while RoBERTa demonstrates enhanced contextual flexibility, its similarity scores show greater variability (standard deviation 0.133) compared to CoInCo and WordNet. This variability signals potential semantic drift, contrasting with WordNet's more stable but less contextually dynamic synonym sets.

Keywords: Verb synonymy, Large Language Models, RoBERTa, T5, WordNet, CoInCo, context-dependent synonymy, use-based synonymy, cosine similarity, lexical resources, word embeddings

1 Introduction

Synonymy refers to a paradigmatic lexical relation between two or more linguistic expressions that share the same or nearly the same meaning in certain contexts (Maienborn et al., 2011).

Among various types of synonymy, one well-known type consists of synonym pairs that share the same referent that refer to the same entity, event, or property in the world. These are what we term **referential synonyms** in this work. Lexical databases like WordNet (Miller, 1995) offer a structured means of identifying such referential synonymy by organizing words into synsets — sets of words that are interchangeable in at least one context because they share a common referent or denotation. Importantly, referential meaning is often the only kind of meaning considered in formal semantics and formal pragmatics. In these traditions, meaning is typically equated with truth-conditional content, where the central concern is whether expressions refer to the same entities or have the same extensions in possible worlds.

Although lexicographers group such words with similar meaning as synonyms, they may not always function interchangeably in any context. For example, the clear synonymy relation between the words "fiddle" and "violin" depends on the formality or informality of the context in which they appear; therefore, the degree of interchangeability between them depends on the context. In a more detailed example, if we search for a synonym for the word "kill", we may find "murder" which is not an absolute synonymy for this word, but it can be a near synonymy which is different in terms of intentionality (Cruse, 1986).

Synonymy judgment is constrained by context (Murphy, 2003). For example, the words "prize" and "award", which are referential synonyms of each other, can be perfect synonyms in a context like (a). They might still be considered

synonyms in a neutral context like (b). However, their meanings might not be similar enough to be considered synonyms in context (c).

(a) Joe won the prize (award) for the best drawing.
(b) what is a synonym for prize? award
(c) The planintiff received a hefty award (≠ prize) in the lawsuit.

The importance of context in synonymy is expressed in several approaches to semantics. Harris's distributional hypothesis (Harris, 1954) posits that words that occur in similar contexts tend to have similar meanings. This principle underlies the foundation of distributional semantics, where word meanings are derived from patterns of co-occurrence in large corpora. Moreover, words possess a flexible and generative nature, highlighting the context-dependent meaning of words (Pustejovsky, 1998).

Context is also essential in how computational models like transformer-based Large Language Models (LLMs) (Vaswani, 2017; Lin et al., 2022) represent natural language expressions. These models use token and positional embeddings, and attention mechanisms to capture relationships between words in a sentence (Vaswani, 2017). Thus, LLMs provide a powerful framework for analyzing synonymy by capturing context-sensitive relationships between words.

In more details, we compare the cosine similarities between the T5 large language model (Raffel et al., 2020)'s representation of a target verb in a sentence with that of the following groups: (1) the recommended substitutions by human annotators for the target words in specific sentences (**gold use-based synonyms**), (2) the suggested synonyms by WordNet for the target words (**referential synonyms**), (3) the synonyms generated by LLMs for the target words (**LLMs synonyms**), and (4) a random group including unrelated substitutions with no lexical connection to the original target words. Given that **referential** synonyms do not take context into account, while **use-based** synonyms and LLM-generated alternatives do, we want to investigate how close referential synonyms are to the others in meaning, and how much their overlaps differ. In this study, we investigate what cosine similarity scores within LLMs' embedding space reveal about the degree of alignment among these different sources of synonymy. To ensure a robust and unbiased evaluation of these different synonym types, we calculate similarity scores within the embedding space of a T5 model, chosen for its strong performance in capturing sentence-level meaning and to avoid the potential bias of using our synonym-generating model (RoBERTa) for evaluation.

2 Literature Review

Advancements in contextualized language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have driven significant progress in understanding the syntactic and semantic relationships in language. Probing studies such as (Rogers et al., 2021) have highlighted BERT's capacity to encode rich linguistic knowledge, including entity types, semantic roles, and proto-roles, which supports its application in semantic analysis. These studies establish that BERT's architecture allows it to differentiate word meanings in various contexts effectively. Similarly, (Staliūnaitė and Iacobacci, 2020) showed that RoBERTa can handle lexical relations such as antonymy, highlighting the appropriateness of its contextualized embeddings for semantic analysis. This aligns with (Vulić et al., 2020), who revealed RoBERTa's strength in predicting lexical relations, particularly homonymy and synonymy and (Nair et al., 2020), which highlighted that BERT's alignment with human judgments on word sense relatedness is strong for distinctly separated senses (homonymous).

Lexical substitution (LS) (McCarthy and Navigli, 2007) is the task of replacing a target word in context with a semantically similar alternative while preserving the meaning of the original sentence. Although LS was initially proposed as a method for evaluating word sense disambiguation systems (McCarthy, 2002), subsequent research has focused on identifying the most suitable substitutions to enhance performance in various natural language processing tasks, including paraphrase generation (Fu et al., 2019), text simplification (Štajner et al., 2022), machine translation (Agrawal and Carpuat, 2019), and style transfer (Helbig et al., 2020), among others. In this research, we build on the idea of LS to compare and analyze the similarity levels of words selected with contextual information during substitution with those chosen using thesaurus-based resources.

In another related study, a new benchmark called SWORDS (Lee et al., 2021) was proposed, asking Candidates to first generate substitution using a context-free resource such as a thesaurus, and

then evaluated by humans for their contextual fit. This method is grounded in the intuition that it is cognitively easier for humans to evaluate a given list than to retrieve suitable words from scratch like CoInCo which is limited to human recall. Based on their results, Compared to CoInCo, SWORDS provides 4.1 times more substitutes per target word while maintaining contextual appropriateness by obtaining 1.5 times more appropriate than those found in previous datasets for the same number of candidates. Moreover, the limitations of applying BERT naively to lexical substitution is addressed in (Zhou et al., 2019) by explaining that masking the target word and sampling predictions often yields semantically inappropriate substitutes. Therfore, they proporsed a novel embedding dropout technique which helps BERT to generate substitute candidates that are sensitive to context but not overly biased toward the original word, and tried to increase the diversity and relevance of proposed substitutes. Moreover, to ensure substitute quality, the proposed approach incorporates a validation step that measures the semantic consistency between the original and substituted sentences using BERT's contextualized sentence representations.

As mentioned, one of the essential criteria of lexical substitution is preserving the meaning of the original sentence. While using pre-trained language models (PLMs) ensures contextual relevance, they might produce words that are semantically distant from the original target. In another recent work (Vladika et al., 2025), by concatenating the masked sentence with the original sentence, the authors introduced CONCAT, an augmented lexical substitution approach designed to enhance the contextual information available to the model during masked token prediction, thus improving the generation of contextually appropriate and grammatically coherent substitutes.

Unlike most lexical substitution (LS) tasks, our objective is not to identify the best single substitute for a target word. Instead, we aim to examine how different substitution strategies influence the contextual representation of the target word. By using various lexical resources to generate candidate substitutes—such as human-annotated datasets, thesaurus-based lists, or embeddings-derived candidates, we investigate how the choice of substitution method affects the semantic similarity between the original and modified sentences. This allows us to evaluate not only the appropriateness of substitutes, but also the sensitivity of contextual word meaning to different substitution sources.

In totall, the goal of this work is to examine to what extent LLM-generated synonyms align with both referential and use-based substitutes in context. Specifically, this study evaluates the degree of similarity between LLM-generated synonyms and those provided by CoInCo, a resource reflecting use-based synonymy derived from human judgments, and WordNet, which encodes referential synonymy through structured lexical relations. By focusing on verb synonymy and utilizing cosine similarity to compare embeddings, this work adds to the field by directly assessing whether LLMs encode synonymy in ways that align more closely with context-dependent human judgments or traditional lexical resources. In addition, looking at the overlap between different synonym sets not only reveals how much lexical information is shared across them, but also identifies the unique contributions of each source, offering insights into whether LLMs capture synonymy in ways that converge with context-dependent human judgments or remain closer to static, dictionary-based relations. Taken together, these complementary analyses provide a multidimensional evaluation framework that deepens our understanding of LLMs' representational capabilities.

3 Method

As the LLMs from the BERT family have shown their effectiveness in natural language understanding (Zhang et al., 2020; Staliūnaitė and Iacobacci, 2020), we selected the RoBERTa [1] model (Liu et al., 2019) for generating the LLM-based synonyms—Compared to other LLMs in the BERT family, RoBERTa showed the ability to learn more about lexical features (Staliūnaitė and Iacobacci, 2020).

Additionally, we used the Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2020) to represent both the original sentences and their paraphrases in order to avoid potential bias that could arise from using RoBERTa for both synonym generation and embedding-based evaluation. In other words, using the same model for generating substitutes and for measuring their semantic similarity could risk inflating alignment scores due to model-specific representational artifacts. Therefore, by employing T5 we aimed to reduce this

[1] https://huggingface.co/docs/transformers/en/model_doc/roberta

confound and ensure that the evaluation reflects model-independent semantic similarity.

Moreover, as the effect of context on lexical meaning was a key factor in our investigation, we selected the CoInCo (Concepts in Context) corpus (Kremer et al., 2014) for our experiments, which provides access to 35,000 target words and at least six synonyms for each word recommended by human annotators based on context.

3.1 Data Processing

This study focused exclusively on English verbs, as verbs play a central role in encoding actions and relational meanings in a sentence. We used 4,081 sentences from CoInCo with a target verb and its synonyms, but we only retained sentences where the target verbs had a synonym available in WordNet. Furthermore, we imposed a vocabulary constraint by requiring that the target verbs and all its synonym candidates existed within the T5 model's vocabulary, ensuring accurate word representations in the model's embedding space. After applying these filters, 4,039 target sentences remained, forming the final dataset.

The alternative sentences for each target verb were created by replacing the target verb in the sentence with synonyms from four different sources, resulting in four groups. The first group utilized CoInCo's gold use-based synonyms substitutions, which reflected use-based (contextual) synonymy. For each target word, CoInCo provides several synonyms and a label representing the frequency of each synonym. We selected the substitute with the highest frequency based on the label "freq" in CoInCo. The second group relied on synonyms extracted from WordNet, we used NLTK[2] library in Python to collect all the WordNet synonyms. To determine the intended sense of each verb in context, we employed the classic Lesk algorithm (Lesk, 1986) for word sense disambiguation (WSD). The Lesk algorithm is a knowledge-based WSD method that selects the most appropriate WordNet synset for a word by measuring overlap between the context of the target word and the dictionary definitions (glosses) of its possible senses. Specifically, the algorithm compares the words surrounding the target word in the sentence with the definitions and example usages of its candidate synsets in WordNet, selecting the sense with the greatest lexical overlap.

We used the implementation available in the NLTK library, which applies the simplified Lesk variant introduced by (Banerjee and Pedersen, 2002), considering context windows and WordNet relations such as glosses and examples.
The third group consisted of synonyms generated by RoBERTa, which provided a contextual representation of the target word. To extract the synonyms suggested by RoBERTa, we used a masked language modeling approach. Specifically, for each target word in a sentence, we replaced it with a mask token (<mask>), allowing the RoBERTa model provided by the Transformers [3] library in Python to predict the most likely word to fill in the blank. Among the model's predictions, we selected the most probable word that was different from the original target word, assuming it to be the best synonym candidate in the given context.

Lastly, a random group was developed, which included unrelated substitutions with no lexical connection to the original target words. In more detail, first, we used NLTK to extract all English verbs in the *omw-1.4* (Open Multilingual WordNet) corpus. Then, for each target verb, we selected a verb in the corpus that was neither the target verb nor among its WordNet synonyms. An example of a target sentence and its substitutions in each group is shown in Table 1, where the verb "end" is replaced by different alternatives.

Target	A mission to **end** a war
CoInCo	A mission to **stop** a war
RoBERTa	A mission to **stop** a war
WordNet	A mission to **finish** a war
Random	A mission to **take_a_dare** a war

Table 1: Comparison of Different Groups

3.2 Model

The main analysis in this study was based on the evaluation of the cosine similarity between the T5 model's representations of the target words and their substitutions from four sources: CoInCo (gold use-based synonyms), WordNet (referential synonyms), RoBERTa (LLM-generated synonyms), and Random (unrelated substitutions).

We used the T5 model as a contextual encoder to obtain sentence-level representations. T5 treats every NLP task as a text-to-text problem, making

[2]https://www.nltk.org/

[3]https://huggingface.co/docs/transformers/en/index

it highly adaptable for contextual semantic tasks. In our setup, we encode sentences containing a target verb using a pre-trained T5 model (e.g., T5-base), generating contextual embeddings that reflect the influence of surrounding lexical material on the target word. These representations serve as a semantic basis for comparing substituted and original sentences, allowing us to assess how well different models preserve or shift meaning when proposing substitutes —Our choice of the T5 model for evaluation was a deliberate methodological decision. First, to ensure an unbiased assessment, we used a different model for evaluation than for synonym generation (RoBERTa). This separation prevents the risk of inflated similarity scores that can arise from model-specific artifacts. Second, T5's encoder-decoder architecture is distinct from masked language models like RoBERTa and is particularly adept at capturing the broad contextual dependencies and subtle semantic shifts essential for our analysis, as it excels at tasks involving sentence-level meaning.

Tokenization: To represent the sentences in each group and extract the embedding of the target words, we first tokenized the input sentence using T5's tokenizer. Next, we lemmatized the target word and the T5's tokens using spaCy library in Python (Honnibal and Montani, 2017) to obtain their base forms for accurate comparison. Then we compared the target words with the T5 tokens (base forms) to identify the index of the token that matched the target word.

Representation: We passed the tokenized sentences through the T5 model to generate contextual embeddings for all tokens. Next, using the identified index, we extracted the high-dimensional embedding corresponding to the target word from the model's last hidden state.

3.3 Evaluation

Similarity Measurement: After collecting the embeddings of the four substitutes for the target words within the sentence contexts, for each sentence, we calculated the cosine similarity between each substitute and the target word, using the SciPy library in Python[4], to analyze the semantic alignment between the groups—Semantic alignment refers to the cosine similarity score between the contextual embedding of a substitute word and the target word within the T5 model's embedding space. A higher score indicates that the substitute is semantically closer to the original word in that specific context. For example, in *A mission to end a war*, the substitute "stop" would demonstrate a high semantic alignment, while the random substitute "tak_a_dare" would show a very low alignment.

Intersection Statistics To assess the degree of lexical similarity and agreement between the substitute word sets generated by CoInCo, RoBERTa, and WordNet, we analyzed the overlap of multiple top candidate synonyms rather than focusing solely on the best single substitute from each method. For each target word in a sentence, we extracted the top 4 substitute candidates from each source —CoInCo, RoBERTa, and WordNet — forming three sets of potential synonyms. We then computed pairwise intersections between these sets to determine the proportion of shared substitute words between each pair of methods. Specifically, for every sentence, we calculated the pairwise intersection between CoInCo and RoBERTa, CoInCo and WordNet, and RoBERTa and WordNet among their respective top candidates, expressed as a percentage. These overlap percentages were computed individually for each sentence to preserve context-specific lexical variation, and then averaged across the entire dataset to obtain an overall measure of alignment and divergence between methods. This approach captures not only agreement on the single most likely synonym, but also the broader semantic neighborhood suggested by each method, providing richer insight into how these lexical resources and models complement or diverge from each other in proposing substitute words. A high average overlap indicates that different methods tend to propose similar sets of plausible synonyms in context, whereas a low overlap reflects distinctive lexical suggestions unique to each method.

Statistical Significance The variables used in the statistical analysis were the cosine similarity values, calculated using T5 embeddings to represent both the original sentence and its paraphrases. Furthermore, the statistical methods applied to the data included Welch's ANOVA and the Games-Howell post-hoc test.

To assess the assumptions for parametric testing, the Levene test and Bartlett test were first performed using the SciPy library in Python. Both tests confirmed significant differences in group variances (Levene's test: Statistic = 277.167, p =

[4]https://scipy.org/

0.000; Bartlett test: Statistic = 894.552, p = 0.000), indicating a violation of the homogeneity of variances assumption. Additionally, normality checks showed that the overall distribution (pooled from all groups) was approximately normal, with skewness = –0.084 and excess kurtosis = 0.556, both of which fall within commonly accepted thresholds for normality.

Given these results, Welch's ANOVA was selected as it is robust to unequal variances and performs reliably even under moderate deviations from normality, particularly when the sample size is sufficiently large. This test revealed statistically significant differences across groups.

To determine which group means differed significantly, the Games-Howell test was used for pairwise comparisons. This post-hoc test is appropriate when variances are unequal and does not require equal group sizes, making it suitable for our analysis.

4 Result

The average overlap in the top four substitutes per target verb varied considerably across the three resources. The comparison between CoInCo and RoBERTa yielded the lowest agreement, with an average overlap of only **5.03%**, indicating that these two methods rarely propose the same candidate substitutions in context. A higher, yet still limited, degree of overlap was observed between CoInCo and WordNet at **12.60%**, suggesting slightly greater alignment between gold use-based synonyms substitutes and referential synonyms. The overlap between WordNet and RoBERTa was also low, at **4.94%**, demonstrating that the distributional predictions of the language model differ substantially from the lexical entries in WordNet. Overall, these results indicate that the three methods generate largely distinct sets of candidate substitutes, with minimal convergence beyond chance levels.

While cosine similarity metrics indicated some semantic alignment between methods, the overlap analysis reveals minimal agreement across the broader candidate synonym sets. The particularly low overlap values suggest that even when methods identify similar top-ranked substitutes, they tend to propose largely distinct alternatives overall. This pattern reflects complementary rather than redundant lexical knowledge among CoInCo, WordNet, and RoBERTa's substitute proposals.

In addition, statistical analyses demonstrated significant differences in the degree to which the four synonym groups aligned semantically with their target words. Welch's ANOVA revealed a highly significant effect of synonym source on mean cosine similarity scores ($F(3, \cdot) = 7510.70$, $p < .001$), with a large effect size ($\eta_p^2 = 0.489$), indicating that nearly half of the variance in similarity can be attributed to differences between groups.

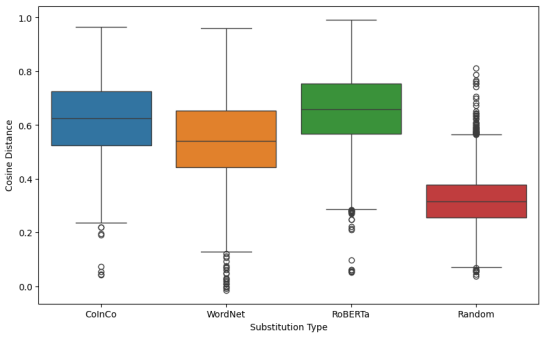


Figure 1: Cosine Similarities by Synonym Types

Post-hoc pairwise comparisons using the Games-Howell test (Table 3, Appendix A.2) confirmed that all group differences were statistically significant ($p < .001$). The largest effects were observed between CoInCo vs. Random and RoBERTa vs. Random, while moderate to strong effects were found in comparisons involving WordNet (e.g., RoBERTa vs. WordNet, CoInCo vs. WordNet). These results suggest that synonym sets derived from contextualized (LLM-based) and use-based resources exhibit stronger semantic alignment with target words than those generated from random selection or static lexical databases.

Descriptive statistics further elucidate these differences (Table 4, Figure 1). RoBERTa and CoInCo exhibit the highest mean similarity scores (0.656 and 0.625, respectively), indicating that their substitutes generally preserve semantic similarity with the target word better than those from WordNet (mean = 0.548) or Random (mean = 0.321). Median values follow the same pattern, with RoBERTa's median (0.659) notably higher than the other methods, reflecting a consistent tendency to produce contextually relevant substitutes.

However, RoBERTa's similarity scores show greater variability (standard deviation = 0.133) compared to CoInCo and WordNet, with a wider range spanning from a minimum of 0.051 to a maximum near 0.99. This suggests that although RoBERTa frequently generates semantically close synonyms, it can also produce less similar or con-

textually inappropriate substitutes, highlighting the challenges LLMs face in fully capturing nuanced or rare word senses.

In contrast, WordNet displays a more conservative similarity distribution, with lower mean and median scores and narrower variability, indicative of a lexically constrained but stable synonym set that is less sensitive to context. The broader distributions for CoInCo and RoBERTa likely reflect their enhanced contextual adaptability, with RoBERTa especially able to generate a wider semantic spectrum of substitutes.

Table 2 presents a detailed comparison of the top substitute candidates generated by CoInCo, RoBERTa, and WordNet for the target verb "begin" in a specific context. This example illustrates the key findings regarding overlap and semantic strategy divergence.

Target Sentence	The company will begin mailing materials to shareholders.
CoInCo	start, commence, initiate
RoBERTa	start, initiate, shipping, sending, posting
WordNet	start, commence, set_out, start_out, get_down

Table 2: Examples of Different Group Substitutes

Demonstrating Complementary Lexical Knowledge (Low Overlap) This example illustrates the finding that the candidate sets are largely distinct and complementary, rather than redundant, as reflected by the overall low average overlap (e.g., WordNet vs. RoBERTa at 4.94%).

- CoInCo focuses on traditional initiation verbs: start, commence, initiate.
- WordNet includes similar terms (start, commence), but also phrasal verbs that may be less contextually appropriate, such as set_out or get_down.
- RoBERTa provides a mix of precise synonyms (initiate) and contextually derived words that describe the subsequent action (e.g., shipping, sending, posting).

The inclusion of verbs related to the direct object (e.g., shipping materials) by RoBERTa, alongside abstract synonyms (initiate), demonstrates a distinct, context-sensitive approach that differs substantially from the sets generated by CoInCo and WordNet, supporting the conclusion of minimal convergence across the broader candidate sets.

Demonstrating RoBERTa's Contextual Success and Flexibility This example highlights RoBERTa's ability to capture the fine-grained, context-dependent aspects of verb meaning and its greater contextual adaptability.

For the target verb "begin" in the context of mailing materials:

- RoBERTa proposes several substitutes that are highly specific to the context of distributing physical items: shipping, sending, and posting. These terms are not general synonyms for "begin" but are contextually functional substitutes that preserve the intended meaning of the action being initiated. This use of highly specific verbs validates the finding that RoBERTa excels in generating contextually relevant substitutes, contributing to its highest overall mean similarity score (0.656).

Demonstrating RoBERTa's Variability vs. Referential Stability This comparison addresses the finding of a trade-off between semantic precision and contextual flexibility.

- RoBERTa's Variability: RoBERTa shows high contextual flexibility by including shipping and posting. While contextually useful, these are semantically distant from the target verb's core "initiation" meaning, contributing to the model's greater variability (Std Dev = 0.133). The model risks semantic drift by moving too far into the related action space.
- WordNet's Stability: WordNet exclusively provides terms focused on the referential meaning of initiation (start, commence, set_out). This results in a conservative similarity distribution, demonstrating a stable but less contextually dynamic set.

This divergence illustrates that LLMs capture a broader semantic spectrum, sometimes sacrificing general semantic proximity for high contextual relevance (e.g., shipping vs. start), whereas static resources maintain referential consistency regardless of the specific surrounding context (the mailing materials clause). Overall, these findings highlight a trade-off between semantic precision and contextual flexibility: CoInCo and RoBERTa

generate more semantically aligned and context-sensitive synonyms on average, while RoBERTa's broader distribution signals occasional semantic drift. WordNet provides more stable but less contextually dynamic synonym sets, and Random substitutions unsurprisingly yield the lowest semantic alignment.

5 Conclusion

The closer semantic proximity between LLM-generated substitutes and use-based ones showed that LLMs produce word substitutes that are closer to use-based approaches like CoInCo. Furthermore, the closer proximity of LLM-generated and use-based substitutes to target words underscores the effectiveness of contextual modeling in capturing fine-grained, context-dependent aspects of verb meaning.

Nevertheless, the limited overlap across the full sets of candidate synonyms produced by different methods reveals that each approach explores distinct lexical neighborhoods, offering complementary perspectives on lexical semantics rather than converging on identical substitutions. This divergence highlights the complex and gradient nature of synonymy, where semantic equivalence varies depending on pragmatic and usage contexts.

Moreover, while LLMs exhibit greater contextual flexibility and adaptability, this strength is accompanied by increased variability in synonym quality, occasionally leading to semantic drift. In contrast, traditional lexical resources provide more stable but less context-sensitive synonym sets. These contrasting properties suggest that no single resource suffices to comprehensively model the full spectrum of verb synonymy.

Together, these findings reinforce the effectiveness of context-aware approaches, especially those powered by LLMs in capturing subtle nuances of verb similarity, while also underscoring that methods may agree semantically without converging on the same lexical choices. This divergence highlights the potential benefits of combining multiple approaches to broaden coverage and enhance robustness in synonym substitution tasks.

6 Limitations

While this study provides valuable insights into the representation of verb synonymy in LLMs and their comparison with lexical resources, several limitations must be acknowledged. First, the analysis was limited to verbs extracted from the CoInCo dataset, which may constrain the generalizability of findings to other parts of speech or less commonly used verbs. Second, this study relied on cosine similarity as the sole metric of alignment. This metric may not capture all semantic subtleties, particularly in cases involving polysemy or context ambiguity.

Additionally, while WordNet and CoInCo served as strong lexical benchmarks, incorporating other lexical databases or expert-annotated datasets could further strengthen the external validity of the findings. Future work could also include (i) asking human annotators to rate the synonyms generated by WordNet and RoBERTa as an additional metric for comparison, and (ii) extending the analysis beyond English to examine cross-linguistic patterns. Moreover, although this study mitigated potential bias by using a different model for evaluation, it did not systematically explore how variations in LLM architectures (e.g., GPT, XLNet, BERT variants) might influence synonym generation and semantic alignment. Addressing these gaps by covering a broader range of linguistic categories, employing diverse evaluation methods, and performing cross-architecture comparisons would provide deeper insights into model-specific behaviors in capturing context-aware synonymy.

Statement: *During the preparation of this work the authors used ChatGPT-4o to help with text editing, after which, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.*

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. *arXiv preprint arXiv:1911.00835*.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136–145. Springer.

D.A Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. *Advances in Neural Information Processing Systems*, 32.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

David Helbig, Enrica Troiano, and Roman Klinger. 2020. Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. *arXiv preprint arXiv:2005.07617*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A benchmark for lexical substitution with improved data coverage and quality. *arXiv preprint arXiv:2106.04102*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI open*, 3:111–132.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Claudia Maienborn, Klaus von Heusinger, and Paul Portner. 2011. *Semantics: An international handbook of natural language meaning*, volume 1. Walter de Gruyter.

Diana McCarthy. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 089–115.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

MLynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. *arXiv preprint arXiv:2010.13057*.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in artificial intelligence*, 5:991242.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. *arXiv preprint arXiv:2009.08257*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Juraj Vladika, Stephen Meisenbacher, and Florian Matthes. 2025. Lexical substitution is not synonym substitution: On the importance of producing contextually relevant word substitutes. *arXiv preprint arXiv:2502.04173*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9628–9635.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3368–3373.

7 Appendix

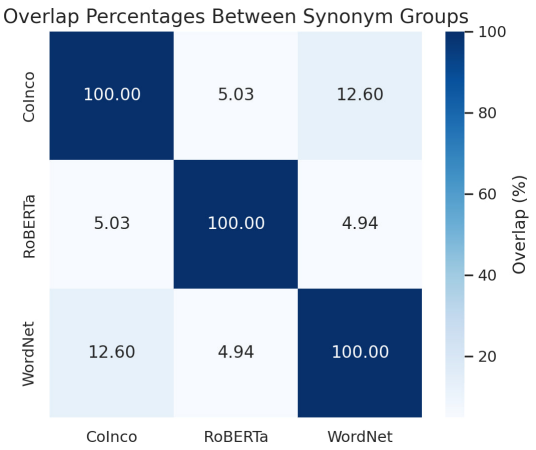


Figure 2: A.1: Overlap Percentages Between Synonym Groups

Group Comparison	**Mean Difference**	**p-value**	**Hedge's g**
CoInCo vs Random	0.3044	< 0.001	2.5167
CoInCo vs RoBERTa	-0.0307	< 0.001	-0.2245
CoInCo vs WordNet	0.0769	< 0.001	0.5155
Random vs RoBERTa	-0.3351	< 0.001	-2.8629
Random vs WordNet	-0.2275	< 0.001	-1.7304
RoBERTa vs WordNet	0.1076	< 0.001	0.7367

Table 3: Appendix A.2: Games-Howell Post-Hoc Test Results

Statistic	**CoInCo**	**RoBERTa**	**WordNet**	**Random**
Count	4039	4039	4039	4039
Mean	0.6251	0.6557	0.5481	0.3206
Std	0.1400	0.1332	0.1578	0.0983
Min	0.0424	0.0512	-0.0141	0.0384
25%	0.5248	0.5664	0.4418	0.2552
Median	0.6255	0.6589	0.5418	0.3160
75%	0.7263	0.7535	0.6535	0.3792
Max	0.9648	0.9922	0.9601	0.8109

Table 4: Appendix A.3: Descriptive Statistics of the Cosine Similarities

Non-English Code Generation with Cross-Lingual Chain of Thought

Yuha Nishigata
Japan Women's University
Tokyo, Japan
m2116061ny@ug.jwu.ac.jp

Waka Ito
Japan Women's University
Tokyo, Japan
m2016013iw@ug.jwu.ac.jp

Kimio Kuramitsu
Japan Women's University
Tokyo, Japan
kuramitsuk@fc.jwu.ac.jp

Abstract

Large language models (LLMs) that support multiple languages are becoming increasingly important. A key challenge in building such models is the performance gap between English and non-English languages, largely caused by the imbalance of pre-training data across languages.

We introduce Cross-Lingual Chain-of-Thought (CL-CoT), a fine-tuning approach designed to improve code generation in low-resource languages through cross-lingual transfer. CL-CoT works by adding English chain-of-thought prompts to instruction texts in low-resource languages, such as Japanese, encouraging the model to reason in English while generating code.

We evaluated CL-CoT on code generation tasks in three low-resource languages. Our method outperformed conventional approaches in most cases, indicating that CL-CoT effectively promotes cross-lingual transfer.

1 Introduction

Large language models (LLMs) have become foundational components of modern information systems. For widespread adoption and practical implementation, users must be able to utilize LLMs smoothly in their native languages. However, developing LLMs that operate effectively in non-English languages presents significant challenges.

The fundamental cause is the imbalance of language resources in pre-training datasets (Li et al., 2025; Zhang et al., 2024).LLM pre-training relies heavily on web text data, which is predominantly English(Penedo et al., 2025). This tendency is particularly pronounced in specialized domains such as engineering and medicine (Zhao et al., 2024).

Cross-lingual transfer has attracted attention as one approach to address this challenge. Cross-lingual transfer refers to transferring knowledge acquired through learning in a specific language to another language (Wu and Dredze, 2019). In particular, by transferring knowledge from high-resource languages such as English to low-resource languages such as Japanese, performance improvements can be expected for instructions in low-resource languages (Xu et al., 2025). However, many aspects of the mechanisms underlying cross-lingual transfer remain unclear.

We propose Cross-Lingual Chain-of-Thought (CL-CoT), an instruction tuning method based on the internal translation hypothesis. This hypothesis suggests that LLMs internally perform reasoning in English when processing non-English instructions(Schut et al., 2025). CL-CoT aims to promote cross-lingual transfer by explicitly encouraging English-based reasoning.

In this study, we evaluated the effectiveness of CL-CoT in promoting cross-lingual transfer using code generation tasks. Code generation has been reported to significantly improve programming efficiency and reduce the burden on developers (Paradis et al., 2024). If we consider programming languages as a type of language, code generation can be viewed as a translation task from natural language to code. However, what distinguishes code generation from translation tasks is that established methods exist for quantitatively evaluating the correctness of generated code(Chen et al., 2021).

In code generation tasks, it is known that significant performance gaps exist between instructions in English and instructions in low-resource languages (Li et al., 2024; Sato et al., 2024). Additionally, since the output is quantitatively evaluable code, this is a task that enables the analysis of cross-lingual transfer. In this paper, we analyze cross-lingual transfer in code generation not only between English and Japanese, which we have previously investigated, but also with other Asian languages (Vietnamese and Korean) to verify the general applicability of CL-CoT.

The remainder of this paper is organized as fol-

lows. Section 2 defines the problem regarding language resource quantity and cross-lingual performance gaps, and explains why we focused on cross-lingual transfer and internal translation. Section 3 proposes the CL-CoT method. Section 4 reports the evaluation of code generation performance using the CL-CoT method. Section 5 summarizes related work, and finally, Section 6 concludes this research.

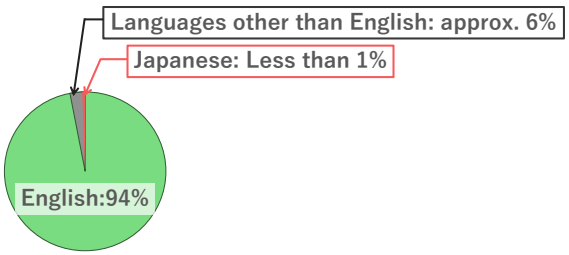


Figure 1: Language distribution of code comments in The Stack dataset as a percentage-based pie chart.

2 Background

In this section, we describe code generation tasks, language resource imbalances, and cross-lingual transfer methods that inform our approach.

2.1 Code Generation

Code generation converts natural language instructions into executable code (Jiang et al., 2024). Similar to machine translation, it involves converting natural language descriptions into another representation system (programming languages). However, generating code requires understanding the intent behind natural language descriptions and producing algorithmic solutions. Therefore, code generation tasks serve as a testbed for fundamental NLP capabilities such as compositional understanding and abstract reasoning.

Code generation tasks have robust semantic evaluation metrics that directly measure whether a model truly understands and solves given problems.

Traditional evaluation metrics like BLEU scores often have low correlation with human judgment, making objective assessment challenging. Code generation tasks address this limitation by enabling robust semantic evaluation through execution-based metrics such as Pass@k (Chen et al., 2021), providing clear, quantitative performance evaluation independent of subjective interpretation.

2.2 Language Resources

Code generation faces significant language resource imbalances. The latest information sources about code are disseminated through documentation, tutorials, and academic papers written in English. Additionally, discussions about code implementation methods and error resolution are primarily conducted in English on developer-oriented platforms such as GitHub and Stack Overflow. According to Kocetkov et al. (2022), in The Stack, one of the code datasets used for pre-training, 94% of code comments are written in English as shown in Figure 1, with the proportion of low-resource languages being extremely limited. Given this background, English is the only high-resource language, while many languages including Chinese are considered as low-resource languages.

The imbalance in language resource quantity also affects actual code generation performance, and clear performance differences have been confirmed between instructions in English and non-English languages. In a study by Wang et al. (2024), a comparison of code generation performance for instructions in English and Chinese showed that performance decreased by more than 13% for Chinese instructions. The study identified the lack of language resources for non-English languages as the primary factor behind this performance degradation.

2.3 Cross-Lingual Transfer

Cross-lingual transfer enables knowledge learned in one language to be utilized in another language (Wu and Dredze, 2019). It has attracted attention as a means to resolve language resource imbalances and cross-lingual performance gaps. In code generation, utilizing knowledge acquired in English enables efficient performance improvements even for instructions in low-resource languages.

The underlying principles and mechanisms of cross-lingual transfer remain unclear. The internal translation hypothesis suggests that when LLMs receive instructions in languages other than English, they internally translate the instructions into English and then generate responses using knowledge learned in English (Wendler et al., 2024; Schut et al., 2025). We focus on an instruction tuning method based on the internal translation hypothesis to promote cross-lingual transfer.

3 Proposed Method

In this section, we describe existing instruction tuning methods and our proposed method, CL-CoT. Figure 2 shows an overview of the proposed method using Japanese as an example.

3.1 Instruction Tuning

Instruction tuning is supervised fine-tuning using paired data consisting of instruction texts and corresponding outputs, enabling LLMs to follow given instructions (Wei et al., 2021).

A key characteristic is that effective improvements can be achieved with far less training data compared to pre-training (Zhou et al., 2023).

Below is an example of instruction tuning data in Japanese:

Instruction

2つの数の乗算を計算する関数をかきなさい.

Output

```
def calculate_multiple(a, b):
    return a * b
```

Multilingual LLMs are created by composing instruction texts in target languages (Wang et al., 2024) and applying tuning. However, adding instruction texts in multiple languages can cause catastrophic forgetting (Fujii et al., 2024), leading to performance degradation in certain languages.

3.2 Para-lingual SFT

Sato et al. proposed Para-lingual supervised fine-tuning with parallel instruction texts in high-resource and low-resource languages to promote cross-lingual transfer (Sato et al., 2025). We refer to this approach as Para SFT. For convenience, we refer to the conventional monolingual SFT as Mono SFT.

Below is an example of Para SFT when Japanese is the low-resource language:

Instruction

2つの数の乗算を計算する関数をかきなさい.
Write a function that calculates the multiplication of two numbers.

Output

```
def calculate_multiple(a, b):
    return a * b
```

Para SFT is based on the internal translation hypothesis. By providing instructions in the order of low-resource language followed by English, Para SFT implicitly encourages translation from the low-resource language to English, and further expects code generation utilizing English knowledge.

3.3 Cross-Lingual Chain-of-Thought

Para SFT implicitly encourages switching from low-resource languages to English through parallel translation. In contrast, we propose Cross-Lingual Chain-of-Thought (CL-CoT), which explicitly promotes cross-lingual transfer by adding Chain-of-Thought(CoT) instruction text that induces reasoning.

The CoT instruction text is a cross-lingual transfer version of the CoT prompt ("Let's think step by step"): "Let's think the instruction in English."

Below is an example of CL-CoT for a low-resource language. The part enclosed by <reason> </reason> tags is the CoT instruction text.

Instruction

2つの数の乗算を計算する関数をかきなさい.
<reason>
Let's think the instruction in English.
</reason>

Output

```
def calculate_multiple(a, b):
    return a * b
```

4 Experiments

This section reports on the effectiveness of the CL-CoT proposed in the previous section. The purpose of this experiment is to verify whether CL-CoT is effective in promoting cross-lingual transfer and improving code generation performance in low-resource languages.

4.1 Experimental Settings

4.1.1 Target Languages

We selected Japanese, Vietnamese, and Korean as target low-resource languages. According to the analysis of The Stack pre-training corpus by Kocetkov et al. (2022), natural-language descriptions

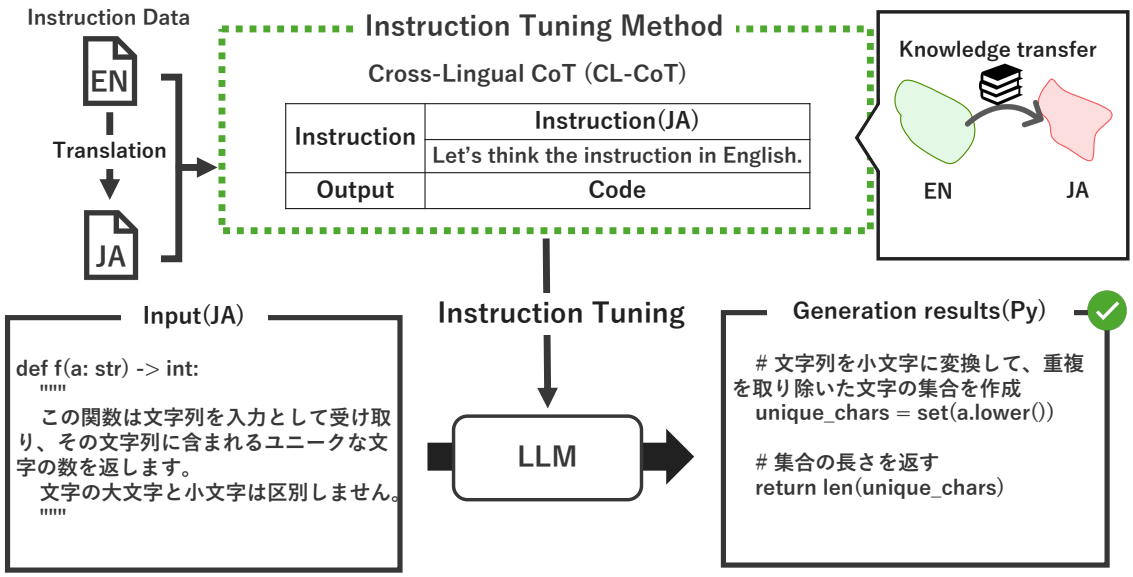


Figure 2: Overview of Cross-Lingual Chain-of-Thought (CL-CoT) instruction tuning method with Japanese input example.

account for less than 1 % in Japanese and Korean, while Vietnamese is not included at all. These languages are therefore well suited to evaluating the effectiveness of instruction tuning in low-resource settings.

4.1.2 Instruction Tuning Methods for Comparison

We investigated the instruction tuning methods described in Section3.

1. **Mono SFT**: Standard instruction tuning method that uses only instruction texts in the low-resource language.

2. **Para SFT**: A method where parallel instruction texts are arranged in the order of low-resource language followed by English.

3. **CL-CoT**: A method where a CoT prompt that encourages reasoning in English is added after the instruction text in the low-resource language.

4.1.3 Instruction Tuning Dataset

We created our dataset using the educational_instruct dataset[1] from the OpenCoder project:

1. From the 118,278 samples in the educational_instruct dataset, extract the top 1,000 English instruction–code pairs, sorted in descending order of instruction length.

2. For the above English instruction texts, we machine-translated them into low-resource languages using the DeepL API[2]. The code was used as-is without translation.

High-quality machine translation is crucial for multilingual LLM development.

Following Zhou et al. (2023), we adopt a sample size of 1,000 pairs, which has been shown effective for instruction tuning experiments.

We compared several translation services by translating a subset of instruction data and checking for errors and omissions. Based on translation fidelity and proven performance in previous multilingual LLM projects, we selected DeepL.

4.1.4 Target LLMs for Evaluation

We evaluated three LLMs with different pre-training data compositions:

- **meta-llama/Meta-Llama-3-8B**: A model developed by Meta, pre-trained on over 15 trillion tokens of data primarily in English.

- **Qwen/Qwen2.5-7B**: A model developed by Alibaba that supports multilingual generation. This model is pre-trained on 18 trillion tokens

[1] OpenCoder-LLM/opc-sft-stage2

[2] https://www.deepl.com

of multilingual data, and the officially supported languages include English, Japanese, Vietnamese, and Korean.

- **infly/OpenCoder-8B-Base**: A code-specialized model jointly developed by INF Technology and M-A-P. This model is pre-trained on 2.5 trillion tokens with a composition ratio of 9:1 between code and code-related web data.

We performed instruction tuning on these models using the entire English dataset from educational_instruct to create baseline models. This baseline served to improve English code generation performance under the same conditions and subsequently examine cross-lingual transfer capabilities.

4.1.5 Evaluation Method

We evaluated code generation performance for low-resource languages using CL-HumanEval (Sato et al., 2024), which is based on HumanEval (Chen et al., 2021). The benchmark includes English, Japanese, Vietnamese, and Korean, with 164 problems provided for each language. CL-HumanEval differs from the original HumanEval in two main ways. First, CL-HumanEval replaces English-derived function and variable names with language-independent identifiers. Second, this benchmark removes hints such as code execution examples. We used Pass@1 as the evaluation metric. Similar to HumanEval, this metric indicates whether the model can generate correct code when generating a single code sample.

4.2 Experimental results

Table 1 reports Pass@1 scores for the three instruction tuning methods across three LLMs. The highest score for each model–language pair is highlighted in bold. CL-CoT achieved the highest scores in most combinations, outperforming the baseline in seven out of nine combinations of three models and three low-resource languages. Our results show that the proposed method, CL-CoT, improves code generation performance for low-resource languages.

In contrast, Para SFT underperformed the baseline in four of the nine model–language combinations. Furthermore, CL-CoT matched or outperformed Mono SFT in six of the nine model–language combinations.

These results suggest that instruction tuning methods that explicitly encourages reasoning in English are effective in promoting cross-lingual transfer.

4.2.1 Performance Changes by Instruction Tuning Data Size

Based on the previous section's results, we investigated how the size of instruction tuning data affects performance. Figure 3 shows the evaluation results for Japanese with data sizes of 1,000, 5,000, and 10,000 samples. The results indicate that increasing the data size from 1,000 to 10,000 samples yields no significant performance improvement. Based on these findings, we conclude that approximately 1,000 instruction tuning samples are sufficient to achieve the observed performance gains.

4.2.2 Proportion of Comments in Target Languages

CL-CoT is an instruction tuning method that encourages reasoning in English. Therefore, even when instructions are given in low-resource languages, there is a possibility that comments in the generated code may be written in English. To address this concern, we conducted an additional analysis of the proportion of comments in the generated code that were written in the same language as that used in the instruction. Table 2 shows the proportion of problems where comments in the generated code were written in the instruction language out of all 164 problems in CL-HumanEval. Bold values represent the highest values for each model.

From the results in Table 2, CL-CoT showed a higher proportion of comment generation in the instruction language than the baseline in seven out of nine combinations of three models and three low-resource languages. This result demonstrates that while CL-CoT promotes reasoning in English, it can appropriately maintain the instruction language in the output. This suggests that promoting cross-lingual transfer and maintaining output language consistency can be compatible, which is a result that supports the effectiveness of CL-CoT.

5 Related Work

The work most closely related to ours is xCoT, proposed by Chai et al. (2024) for mathematical-reasoning tasks. xCoT uses cross-lingual CoT prompting during inference by adding the instruction "Let's think the question in Language and then think step by step in English" to the input, which encourages LLMs to handle questions in low-resource languages while reasoning and producing answers

Table 1: Evaluation results on CL-HumanEval using the Pass@1 metric. The table compares three LLMs with three instruction tuning methods (Mono SFT, Para SFT, and CL-CoT). The highest score for each model–language pair is highlighted in bold.

model	instruction tuning method	EN	JA	VI	KR
Llama3-8B	**Baseline**	50.6	39.6	42.1	43.3
	Mono SFT	-	**45.7**	43.9	**45.7**
	Para SFT	-	38.4	43.9	36
	CL-CoT	-	43.9	**45.1**	45.1
Qwen2.5-7B	**Baseline**	61	54.3	51.2	48.2
	Mono SFT	-	54.3	**54.9**	**52.4**
	Para SFT	-	51.8	51.2	49.4
	CL-CoT	-	**54.9**	**54.9**	**52.4**
OpenCoder-8B	**Baseline**	59.8	41.5	**40.9**	43.9
	Mono SFT	-	**45.1**	39.6	42.7
	Para SFT	-	43.9	37.2	**44.5**
	CL-CoT	-	43.3	39.6	43.9

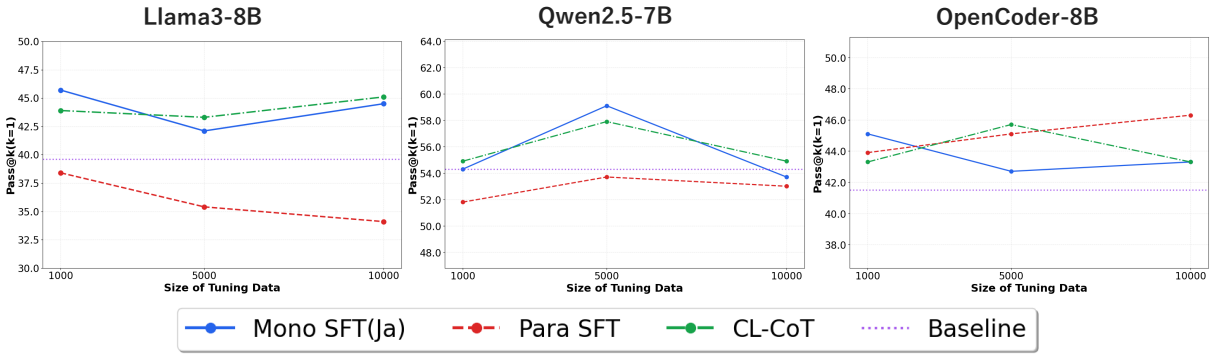


Figure 3: Changes in Pass@1 performance when the size of instruction tuning data is increased.

in English. In contrast, our method is based on instruction tuning that explicitly appends such CoT prompts during tuning while preserving the instruction language in the output for cross-lingual code generation.

In recent years, various methods have been proposed to promote cross-lingual transfer. In the following, we organize related work from three perspectives: approaches other than instruction tuning, instruction tuning-based approaches, and prompting methods.

5.1 Methods for promoting cross-lingual transfer other than instruction tuning

We first describe methods other than instruction tuning for promoting cross-lingual transfer. Representative methods include continual pre-training and model merging.

Continual pre-training refers to further training a pre-trained LLM on data specific to a target language or task, enabling the model to adapt to that domain (Fujii et al., 2024). Although this approach improves comprehension in the target language, it requires substantially more data than instruction tuning to achieve comparable gains.

Model merging is a method that integrates parameters from multiple models with different capabilities (Yang et al., 2024). This method can improve performance without requiring additional training data or large-scale computation, and its effectiveness in Japanese code generation has also been reported (Nakano et al., 2025). However, there is a constraint that integration is difficult between models with different architectures.

Our instruction tuning method has the advantage of being applicable to existing models with a small amount of data compared to these methods.

5.2 Instruction tuning methods promoting cross-lingual transfer

We describe instruction tuning methods that promote cross-lingual transfer.

Table 2: Percentage of benchmark problems in which comments in the generated code were written in the instruction language (low-resource language). The highest value for each model–language pair is highlighted in bold.

Model	Instruction tuning method	JA	VI	KR
Llama3-8B	**Baseline**	**3.7**	1.8	**5.5**
	Mono SFT	0	0	0.6
	Para SFT	0.6	1.2	0
	CL-CoT	1.8	**4.3**	0
Qwen2.5-7B	**Baseline**	26.8	39.6	27.4
	Mono SFT	23.1	33.5	47
	Para SFT	40.2	46.3	40.9
	CL-CoT	**47.5**	**54.3**	**64.6**
OpenCoder-8B	**Baseline**	74.3	93.3	88.4
	Mono SFT	**92.7**	**97**	94.5
	Para SFT	89.6	95.7	95.7
	CL-CoT	92	96.3	**95.7**

Chen et al. (2023) compared monolingual and multilingual instruction tuning. They created multilingual data using the Alpaca dataset and its machine-translated versions, demonstrating that multilingual tuning achieves performance equal to or better than individual language-specific tuning under the same computational constraints.

Shaham et al. (2024) investigated the effects of small amounts of multilingual data. They showed that adding just 40 multilingual examples to an English tuning set significantly improves multilingual instruction-following capabilities.

Research focusing on language mixing within instructions includes Yoo et al. (2024), who proposed dividing instruction text into multiple languages at the sentence or word level. Ranaldi et al. (2024) proposed CrossAlpaca, which improves cross-lingual semantic consistency through demonstrations combining cross-lingual instruction following and translation following.

Compared to these approaches, our method is distinguished by incorporating prompts that encourage English reasoning within low-resource language instructions.

5.3 Promoting cross-lingual transfer using CoT

Many methods for promoting cross-lingual transfer using CoT have also been proposed.

Shi et al. (2022) demonstrated that providing LLMs with prompts for step-by-step reasoning in English achieves high performance regardless of the language of the problem statement.

Qin et al. (2023) proposed Cross-Lingual Prompting (CLP). This method employs a two-stage approach that first aligns cross-lingual understanding and then performs task-specific reasoning, thereby improving multilingual reasoning performance.

Our method differs from these prompting approaches by incorporating CoT elements that encourage English reasoning during instruction tuning.

6 Conclusion

This study aims to narrow code generation performance gaps for low-resource languages by leveraging cross-lingual transfer. Imbalanced language resources in LLM pre-training create performance disparities across languages. Based on the internal translation hypothesis, we propose CL-CoT, which adds English reasoning prompts to low-resource language instructions.

Evaluation results confirmed that the proposed method can promote cross-lingual transfer using minimal data and improve code generation performance in low-resource languages. In particular, evaluation in three languages—Japanese, Vietnamese, and Korean—demonstrated the versatility of CL-CoT. Additionally, the elements in CL-CoT that encourage reasoning in English were found to function effectively for code generation from low-resource languages.

These findings provide valuable insights for developing LLMs aimed at enhancing code generation performance in non-English speaking regions and for designing instruction tuning methods that promote cross-lingual transfer.

Future work will focus on improving per-

formance in multilingual code generation tasks through collecting extensive data in other languages and refining the CL-CoT method. Furthermore, we plan to extend the insights gained from our proposed method to other tasks and validate the generalizability of our approach.

References

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2023. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Daiki Iida, Seiya Ohi, Kakeru Hattori, Shota Hirai, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Building a japanese-centric large language model via continued pre-training. In *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing (NLP)*. The Association for Natural Language Processing (NLP). This work is licensed under CC BY 4.0.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, and 1 others. 2022. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533*.

Mingda Li, Abhijit Mishra, and Utkarsh Mujumdar. 2024. Bridging the language gap: Enhancing multilingual prompt-based code generation in llms via zero-shot cross-lingual transfer. *arXiv preprint arXiv:2408.09701*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Noriko Nakano, Yuha Nishigata, Waka Ito, Nao Soma, Miyu Sato, and Kimio Kuramitsu. 2025. Enhancing cross-lingual transfer in code generation capabilities through model merging. In *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*.

Elise Paradis, Kate Grey, Quinn Madison, Daye Nam, Andrew Macvean, Vahid Meimand, Nan Zhang, Ben Ferrari-Church, and Satish Chandra. 2024. How much does ai impact development speed? an enterprise-based randomized controlled trial. *arXiv preprint arXiv:2410.12944*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

Miyu Sato, Yuha Nishigata, Yuka Akinobu, Toshiyuki Kurabayashi, and Kimio Kuramitsu. 2025. Does instruction tuning with parallel structure promote cross-lingual transfer? In *Proceedings of the Thirty-First Annual Meeting of the Association for Natural Language Processing*. The Association for Natural Language Processing.

Miyu Sato, Yui Obara, Nao Souma, and Kimio Kuramitsu. 2024. Cl-humaneval: A benchmark for evaluating cross-lingual transfer though code generation. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 656–664.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Chaozheng Wang, Zongjie Li, Cuiyun Gao, Wenxuan Wang, Ting Peng, Hailiang Huang, Yuetang Deng, Shuai Wang, and Michael R Lyu. 2024. Exploring multi-lingual bias of large code models in code generation. *arXiv preprint arXiv:2404.19368*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum learning for multilingual transfer in llms. *arXiv preprint arXiv:2411.02460*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Lingualift: An effective two-stage instruction tuning framework for low-resource language reasoning. *arXiv preprint arXiv:2412.12499*.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

The Relationship Between Dialogue Acts and Idea Generation in Human–Human Collaborative Story Writing

Natsumi Ezure and Michimasa Inaba
The University of Electro-Communications
Chofu, Tokyo, Japan
e2430013@edu.cc.uec.ac.jp, m-inaba@uec.ac.jp

Abstract

With the rapid advancement of generative models, human–AI co-creation systems have been widely studied. However, over-reliance on AI-generated ideas may reduce original human ideas. To inform the design of AI systems that better elicit human creativity, we investigate human–human co-creative dialogue, focusing on dialogue acts (DAs) that promote idea generation. In this study, we collected data on collaborative story creation between two human workers assigned to asymmetric roles: a leader, who writes the story, and a supporter, who assists via chat. The dataset comprises 485 dialogues annotated with dialogue acts, presurvey and postsurvey evaluations. Logistic regression analysis results revealed that the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. Furthermore, correlation analysis showed that a higher frequency of accepting proposals and positive opinions from supporters was positively correlated with a higher number of idea proposals from the leader.

1 Introduction

The development of AI technologies, including Large Language Models (LLMs) trained on vast text data, has expanded possibilities for co-creation that combines human ideas and experience with AI. Human–AI co-creation systems have been widely studied; in these systems, users interact with AI to collaboratively work on creative tasks, often by issuing commands for partial modifications or the regeneration of AI-generated content (Oh et al., 2018; Davis et al., 2016; Huang et al., 2020; Louie et al., 2020; Kumaran et al., 2023). Furthermore, systems that interact with people through natural language have been investigated (Schmitt and Buschek, 2021; Yuan et al., 2022). The human–AI co-creation process has been shown to consist of three stages: Ideation, Illumination, and Implementation (Wan et al., 2024). LLMs are frequently employed in the initial Ideation phase. However, a notable issue in ideation tasks is that users tend to accept AI suggestions (Qin et al., 2025). Koivisto and Grassini (2024) found that the best human ideas still matched or exceeded ideas of chatbots. These findings indicate that eliciting ideas from humans is important for co-creation. In contrast, during human–human collaborative dialogue, a partner does not always provide an abundance of ideas. In fact, a situation where a partner offers fewer ideas might motivate an individual to generate more of their own, fostering self-driven ideation. From this perspective, identifying the characteristics of dialogue that promote idea generation in human–human interaction could offer new insights for designing co-creative AI that more effectively elicits human creativity. This study aims to identify the features of utterances that facilitate idea generation by analyzing dialogue data from a human–human co-creative process. Specifically, we address the following research questions:

- **RQ1:** In a co-creative dialogue, what is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?
- **RQ2:** In a co-creative dialogue, what is the relationship between one's own dialogue acts and the partner's dialogue acts?

In this study, we collected data on collaborative story creation between two human workers exchanging ideas. In our experimental setting, two workers were assigned asymmetric roles: a leader, who was responsible for writing the story on an interface, and a supporter, who could only assist through text chat. Logistic regression analysis results revealed that the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. Furthermore, correlation analysis revealed that the number of idea

proposals from leaders increased when supporters offered more frequent acceptances and positive opinions.

2 Related Work

2.1 Human–AI Collaborative Creation

Several studies have been conducted on human–AI co-creation systems using generative models in various domains, including drawing (Oh et al., 2018; Davis et al., 2016) and music (Huang et al., 2020; Louie et al., 2020).

Specifically, various studies have been conducted on collaborative text writing. BunCho (Osone et al., 2021) is a plot co-creation system for Japanese novelists that generates plots based on user-inputted themes and keywords. Dramatron (Mirowski et al., 2023) is a collaborative scriptwriting system built using Chinchilla (Hoffmann et al., 2022). Dramatron first generates a story summary and then progressively generates the title, characters, plot overview, location descriptions, and dialogue for each scene. Users can instruct regeneration or manual correction at any time. CritiCS (Bae and Kim, 2024) is a framework where, after a user inputs an initial draft, multiple LLM critics and one human leader incrementally refine drafts of the plan and story over multiple rounds.

CharacterChat (Schmitt and Buschek, 2021) and Wordcraft (Yuan et al., 2022) are human–AI collaborative writing systems that allow for dialogue through natural language. CharacterChat (Schmitt and Buschek, 2021) is designed for users to create fictional characters while chatting with the system. Wordcraft (Yuan et al., 2022) is a text editor where users can write short stories with a large language model (LLM). Wordcraft users can generate and modify narratives while interacting with LLM in natural language. Similar to CharacterChat and Wordcraft, our work focuses on building a story through dialogue. However, our study distinguishes itself by analyzing human-human collaborative interactions for AI system design.

2.2 Idea Generation

Wan et al. (2024) found that a three-stage iterative Human–AI co-creativity process emerges in collaborative prewriting, consisting of Ideation, Illumination, and Implementation. However, previous research for ideation tasks suggests that humans may overly adopt LLM-generated ideas. For example, Qin et al. (2025) investigated the impact of the timing of LLM assistance on ideation tasks and found that using LLMs from the beginning reduced the number of original ideas. This suggests that AI-generated ideas can ultimately constrain human creativity. In an exercise measuring creativity in divergent thinking, Koivisto and Grassini (2024) found that the best human ideas still matched or exceeded ideas of chatbots, highlighting the importance of eliciting human creativity.

Building on this background, our study focuses specifically on the quantity of ideas to analyze the relationship between collaborative dialogue and creativity.

2.3 Dialogue Acts

Dialogue acts (DAs) are semantic tags assigned to utterances in dialogue (e.g., "suggest" or "answer"). DAs are widely used in dialogue analysis. Although the analysis of DAs in co-creative dialogue data remains limited, for example, Katuka et al. (2022) analyzed the relationship between DAs and participant satisfaction in dialogues. This study, in contrast, aims to identify DAs that may facilitate idea generation by analyzing the quantitative relationship between DAs and the number of ideas generated.

LLMs are used for various tasks in text analysis, such as summarization and classification, and have also been applied to the annotation of DAs. However, it has been reported that the performance of LLMs is inferior to that of manual annotation (Qamar et al., 2025). Therefore, in this study, we first used an LLM (GPT-4-turbo) to initially assign DAs, which were then manually corrected. This two-step approach was employed to maintain high accuracy while reducing the annotation workload compared to creating annotations from scratch.

3 Collaborative Story-Writing Dialogue

Our goal is to gain new insights into eliciting ideas from human–human co-creation, with the aim of contributing to the future design of human-AI co-creative systems. To this end, we assigned two human participants to different roles: a leader and a supporter. This asymmetric setup is designed to approximate a future human–AI co-creation environment, where humans lead the dialogue and generate ideas, and the system supports them. However, it is important to note that our focus is on analyzing genuine human–human dialogue; this is not a Wizard of Oz experiment where a human

simulates an AI.

3.1 Data Collection

Workers were recruited and matched using the crowdsourcing website Lancers [1]. They were asked to create an interesting story related to a given theme in Japanese. Each session involves two workers, divided into the roles of leader and supporter. The leader leads the dialogue and tackles the given creative task, specifically by thinking of interesting stories and writing them on the interface while discussing with the supporter. The supporter can only discuss with the leader via chat. Because this asymmetric setup is designed to approximate a future human–AI co-creation environment where a human leads and the system supports, we primarily analyze the supporter's dialogue strategy to elicit leader ideas in this study.

Workers are required to create at least one story of five to ten sentences on a given theme in 30 min. We built a story co-writing interface using Google Spreadsheets (see Appendix A). We used Google Spreadsheet because it enables the recording of worker operations and editing history through Google Apps Script. The theme is automatically generated by randomly assigning two characters and a genre to the template "(Character 1) and (Character 2) in (Genre)." For instance, the interface generates "merchants and demon lords in adventure story" or "futurists and old man in comedy." We used 52 types of characters and eight genres. The interface includes an input field where the leader worker enters the created story. Workers must create at least one story related to the given theme. They can create two or three stories simultaneously or create a second one after finishing the first one.

Data collection was performed using Zoom[2]. First, the leader worker shared the interface with the supporter using Zoom's screen sharing options. Second, the leader worker generated a theme on the interface and decided on the story's theme to be created. The leader worker was allowed to regenerate themes if they found a theme difficult to write stories about.

The co-creation dialogue began when the leader and the supporter agreed on a theme. The session was set to 30 minutes. Workers could only discuss via Zoom's text chat, as voice calls were prohibited.

[1]https://www.lancers.jp/

[2]https://www.zoom.com/

Postsurvey Item
You generated many ideas
You generated good ideas
You made pertinent points
Your partner generated many ideas
Your partner generated good ideas
Your partner made pertinent points
Your partner was easy to talk to
Your partner stimulated your idea generation
The conversation with your partner was lively
Overall, your partner was a good partner

Table 1: Items of the postsurvey

The leader worker could post to the chat and fill in the interface at any time. Workers could participate in the experiment as many times as they wanted. However, the same pair could not participate in the same roles.

Through a presurvey, we collected demographic data such as age and gender. In the postsurvey, participants used a seven-point Likert scale (Table 1) to answer questions regarding their self-evaluation and their evaluation of their partner. For example, "You generated many ideas" and "Your partner generated many ideas."

To answer RQ1, "In a co-creative dialogue, what is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?", we specifically analyze the postsurvey responses to the items concerning their own idea generation ("You generated many ideas") and their evaluation of the partner ("Overall, your partner was a good partner").

3.2 Data Collection Results

A total of 120 workers participated in our data collection. Initially, 500 dialogues were collected. Then, incomplete data were excluded. As a result, answers to a presurvey by 120 participants were obtained, as well as dialogue histories for 485 dialogues and answers to a postsurvey. A total of 118 workers participated as leaders, and 120 as supporters. Workers could take part in the experiment multiple times; on average, each worker participated 8.08 times (SD = 2.85), with a minimum of 2 and a maximum of 10 participations. Detailed demographic information for the workers is provided in Appendix C.

Figure 1 shows an example of the collected dialogues. The total number of utterances was 20,159

Theme

"Knight" and "princess" in human drama or youth story
(「騎士」と「姫」が出てくる「ヒューマンドラマ・青春物」)

Dialogue

Supporter: What kind of story are you thinking?
(どんな感じで作りますか？)
[setQuestion]

Leader: How about we make the main character a knight who's childhood friends with the princess?
(導入部分は主人公を騎士にして、姫と幼馴染的な展開にしますか？)
[suggest]

Supporter: The childhood friend idea sounds good.
(幼馴染いいですね。)
[positiveOpinion]

Leader: So maybe the princess starts falling for him without even realizing it after he protects her?
(姫は守ってもらったのをきっかけに知らずに知らずのうちに恋に落ちる的な？)
[suggest]

Supporter: I see.
(なるほど。)
[accept]
...

Created story

Once upon a time, there lived a princess and a knight who served in her castle.
(あるところにお姫様とそのお城に仕える騎士がいました。)

They were childhood friends and played together like brothers and sisters when they were young. (二人の関係は幼馴染で小さい頃は兄妹のように仲よく遊んで暮らしていました。)

As they grew older, they became more aware of their respective stations as a knight and a princess, and a distance began to grow between them.
(二人は成長しお互い騎士とお姫様という関係を実感し始め、いつしか距離を取り始めるようになりました。)
...

Figure 1: Example of dialogue during co-creation and a created story (originally written in Japanese, translated by authors). Dialogue acts (DAs) annotated are indicated in brackets.

and the total number of words was 287,077. Then, the average number of utterances per dialogue was 41.6 (SD = 16.6), and the average number of words was 591.9 (SD = 258.6).

3.3 Annotation

To analyze dialogues, we annotated the collected dialogues with dialogue acts (DAs). In this study, we designed new DAs specifically suited for co-creation dialogue.

We assigned provisional DAs to each sentence and repeated the modification of DAs. These tags were inspired by the tags in Hazumi (the multimodal dialogue corpus) (Komatani and Okada, 2021) and the ISO standard 24617–2 tags (Bunt et al., 2012). We decided to use 17 DAs as the final version, as shown in Appendix D.

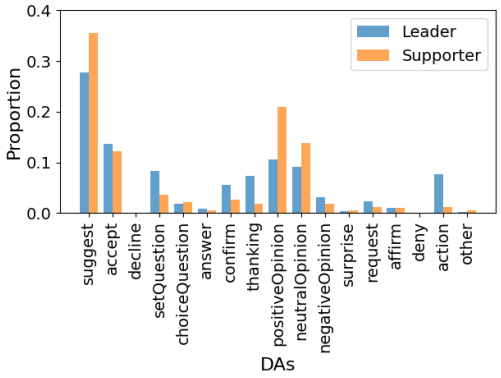


Figure 2: Distribution of DAs for supporters and leaders. The blue bars indicate the proportion of each tag relative to the total number of DAs for leaders, and the orange bars indicate the proportion for supporters.

All dialogues were annotated with GPT-4-turbo [3], followed by manual correction by human annotators. The GPT-4 prompt comprises three parts: the annotation manual, which includes tag definitions; a dialogue created by the authors; and the annotation results of the dialogue (see Appendix E). The Cohen's kappa value between the GPT-4-turbo annotations, which were manually corrected, and human annotations described in the previous paragraph was 0.78. For comparison, when two human workers annotated the five dialogues, the Cohen's kappa between them was 0.79. Therefore, this indicates that using GPT-4-turbo for initial annotation followed by manual correction achieves an accuracy level comparable to that obtained when employing manual annotation by human workers.

Figure 1 shows an example of annotation results. Tags are indicated within square brackets. Figure 2 demonstrates the distribution of the DAs of supporters and leaders. A comparison of leaders and supporters reveals that the proportion of leaders' *setQuestion* (question without options) is higher than that of supporters' *setQuestion*. This indicates that leaders seek more opinions from supporters. Additionally, the proportion of supporters' *positiveOpinion* (positive opinions or impressions) is higher than that of leaders' *positiveOpinion*. This indicates that the leader leads the dialogue, and the supporter responds to the leader's utterance, which is what we intended. Furthermore, the proportion of supporters' suggestions is higher than that of leaders' suggestions. This is because the leader directly inputs ideas into the interface.

[3] https://openai.com/

4 Analysis

Although this study analyzes human–human dialogue, the leader's role is intended to represent the user when co-creating with AI. We aim to analyze dialogues designed to elicit more ideas from leaders.

4.1 Postsurvey Analysis

To investigate the relationship between the leader's quantity of ideas and other factors, we analyzed the results of the postsurvey using logistic regression. In this analysis, as shown in Table 1, the leader's response to "You generated many ideas" was used as the dependent variable, with the other items serving as independent variables. However, to avoid issues of multicollinearity, "Overall, your partner was a good partner" was excluded from the model. Furthermore, the values of the dependent variable were binarized: values greater than or equal to the median were converted to 1, and values below the median were converted to 0. Table 2 shows the results of the logistic regression analysis. Focusing on the statistically significant results, the quality of the leader's own ideas ("You generated good ideas") and the pertinence of their points ("You made pertinent points") had a positive association with the dependent variable. This indicates that when leaders generate a high quantity of ideas, they also tend to generate high-quality ideas and make pertinent points. In contrast, the quantity of the partner's ideas ("Your partner generated many ideas") had a negative association with the dependent variable. This suggests that when the leader's quantity of ideas is high, the supporter's quantity of ideas tends to be low. Possible reasons for this include a leader subjectively feeling their own contribution is smaller when their partner's is larger, or feeling a diminished need to generate ideas themselves when their partner is highly generative.

4.2 Relationship between Evaluation Metrics and Dialogue Acts

The objective of this study is to identify the characteristics of dialogue that facilitate a leader's idea generation. To this end, we analyze the relationship between dialogue acts and the three evaluation metrics detailed below.

Quantity of ideas (subjective) The first metric subjectively evaluates the quantity of ideas generated by the leader. The leader's response score to "You generated many ideas" in the postsurvey was used as the score.

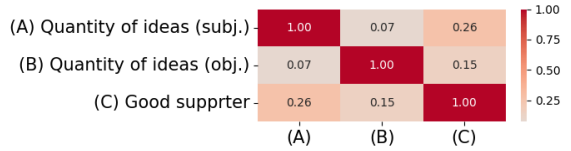


Figure 3: Correlation among three evaluation criteria (Quantity of ideas (subjective) / Quantity of ideas (objective) / Good supporter)

Quantity of ideas (objective) The second metric objectively evaluates the quantity of ideas generated by the leader. The number of *suggest* tags in the leader's utterances is used as the score. *Suggest* tag is assigned to utterances proposing ideas or action. This criterion is employed for an objective evaluation because the first criterion is subjective and there might be a discrepancy between the actual quantity of ideas and subjective evaluation.

Good supporter The third metric evaluates the supporter's positive influence on the leader. When collaborating with a dialogue system, assessing whether it was a good collaborative partner will be anticipated; therefore, we adopted this criterion in this study. The leader's response score to "Overall, your partner was a good partner" in the postsurvey was used as the score for "Good supporter." This criterion correlated with partner-related questions, which indicates the overall evaluation of the supporter, making it a suitable evaluation criterion (see Appendix B).

We conducted a Spearman rank correlation analysis among the evaluation metrics. As shown in Figure 3, the resulting correlations were generally weak. Furthermore, no significant correlation was found between the two metrics for idea quantity, "Quantity of ideas (subjective)" and "Quantity of ideas (objective)" ($\rho = 0.07, p > 0.05$). This result suggests the importance of measuring the quantity of ideas from both subjective and objective perspectives.

Figure 4 shows the results of the correlation analysis between the evaluation metrics and the number of the leader's and supporter's dialogue acts. For this analysis, we used different correlation methods based on the data type of the evaluation metric. For the relationship with "Quantity of ideas (objective)," we used the Pearson correlation coefficients. For

	Coefficient (β)	Std. Error (SE)	z-value	p-value
Intercept	−15.97	1.61	−9.89	***
You generated good ideas	2.49	0.26	9.45	***
You made pertinent points	0.43	0.17	2.54	*
Your partner generated many ideas	−0.66	0.20	−3.23	**
Your partner generated good ideas	−0.08	0.27	−0.28	
Your partner made pertinent points	0.34	0.22	1.53	
Your partner was easy to talk to	0.39	0.24	1.62	
Your partner stimulated your idea generation	0.03	0.23	0.15	
The conversation with your partner was lively	0.09	0.23	0.42	

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Table 2: The results of the logistic regression analysis with the leader's response to "You generated many ideas" as the dependent variable.

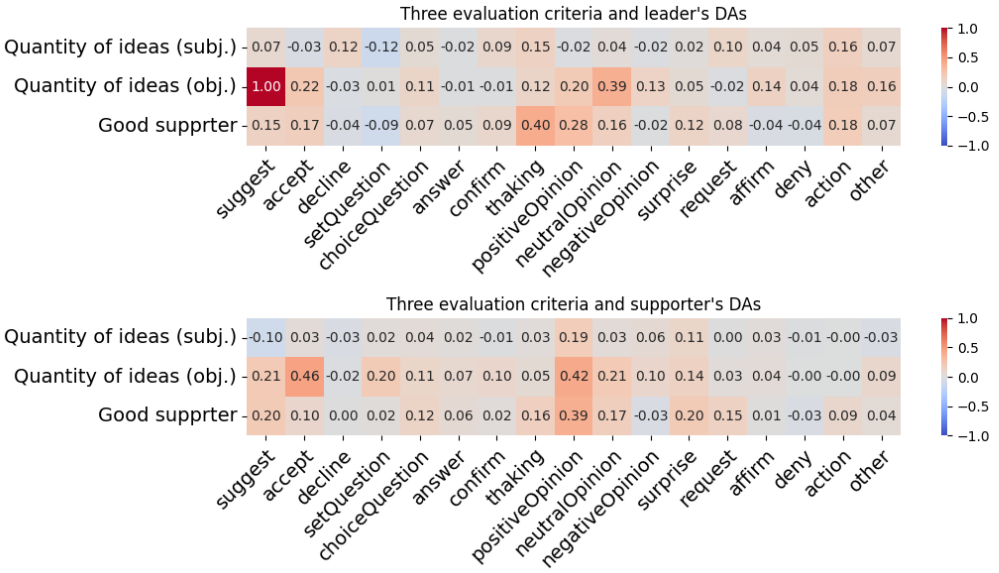


Figure 4: Correlation between three evaluation criteria (Quantity of ideas (subjective) / Quantity of ideas (objective) / Good supporter) and DAs

the relationships with "Quantity of ideas (subjective)" and "Good supporter," which are based on ordinal Likert scales, we used the Spearman rank correlation coefficients.

As shown in the top panel of Figure 4, there were weak to moderate correlations between the "Quantity of ideas" metrics and the leader's dialogue acts. Regarding the "Good supporter" metric, an interesting point is the positive correlation observed with *thanking* ($\rho = 0.40, p < 0.01$). *Thanking* is a DA that expresses gratitude, such as "Thank you." We discuss this result in detail in Section 4.3.

Then, as shown in the bottom panel of Figure 4, "Quantity of ideas (objective)" had weak to moderate correlations with the supporter's DAs. Notably, it showed a moderate positive correlation with the supporter's *accept* ($\rho = 0.46, p < 0.01$). This suggests that an increase in acceptance in the supporter's utterances could potentially lead to an increase in proposals from the leader. A moderate correlation was also found between "Quantity of ideas (objective)" and the supporter's *positiveOpinion*. This indicates that when the leader makes many proposals, the supporter tends to respond with positive feedback, such as "that's good," which can be interpreted as a reaction to the leader's high number of proposals. Furthermore, there was a moderate positive correlation between the "Good supporter" metric and the supporter's *positiveOpinion* ($\rho = 0.39, p < 0.01$). This result suggests that receiving positive opinions from a partner may lead to a more favorable overall impression of that partner. Given that the supporter's *positiveOpinion* is positively correlated with both subjective and objective measures of idea quantity, as well as with the partner evaluation, it can be considered a particularly useful dialogue act when designing strategies to elicit ideas.

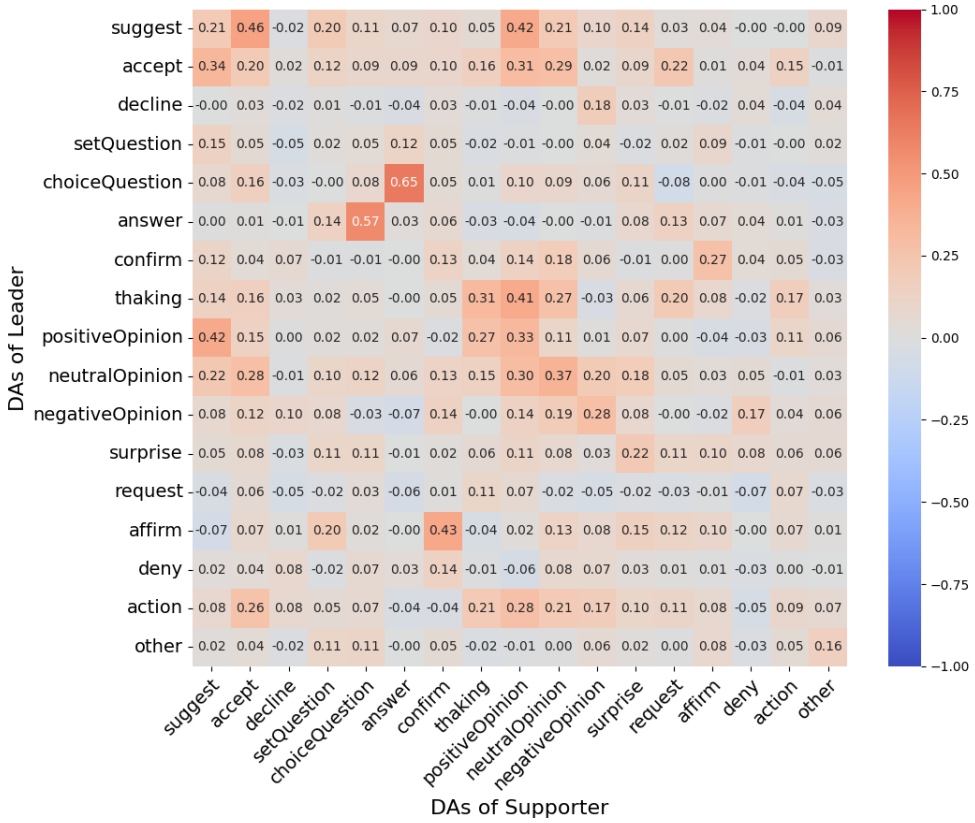


Figure 5: Pearson correlations between leaders' DAs and supporters' DAs

4.3 Analysis of Leader's and Supporter's Dialogue Acts

To analyze the relationship between the dialogue acts of the leader and the supporter, we performed a correlation analysis using the Pearson correlation coefficient.

As shown in Figure 5, some pairs of dialogue acts exhibited correlations that are expected by their definitions. For instance, a strong positive correlation was found between *choiceQuestion* (a question presenting two or more options, e.g., "Is the character female or male?") and *answer* (a response to a *choiceQuestion*, e.g., "Female."). Similarly, a moderate positive correlation was observed between *confirm* (an utterance seeking confirmation) and *affirm* (an affirmative response). These results are consistent with the definitions of the dialogue acts. Furthermore, this analysis revealed a positive correlation between the supporter's *positiveOpinion* and the leader's *thanking*. This suggests a conversational pattern where a positive utterance from the supporter is often followed by an expression of gratitude from the leader. Considering this in conjunction with the finding from Figure 4 that *thanking* correlates positively with the "Good supporter" metric, a plausible causal chain emerges. It is likely that the supporter's use of *positiveOpinion* contributes to the leader's perception of them as a "Good supporter," while also eliciting *thanking* from the leader. Therefore, the observed correlation between *thanking* and "Good supporter" may be an indirect consequence.

5 Discussion

5.1 RQ1: What is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?

As shown in Figure 3, no correlation was found between the two metrics for idea quantity, "Quantity of ideas (subjective)" and "Quantity of ideas (objective)." This discrepancy does not suggest contradictory results but highlights the potential disparity between subjective and objective evaluations of the quantity of ideas. For example, even if the quantity of the leader's ideas was quantitatively high, the evaluation might be relative to the quantity of the supporter's ideas.

In this study, we analyzed human–human dialogue to identify characteristics that elicit ideas. However, a limitation is that the relationships observed between dialogue acts, the quantity of ideas, and partner evaluations were generally weak to moderate, with no strong correlations found. This suggests that a complex creative process like idea generation is likely influenced by a variety of factors beyond the dialogue acts.

5.2 RQ2: What is the relationship between one's own dialogue acts and the partner's dialogue acts?

Our goal is to design dialogues that elicit more ideas from leaders. Therefore, we discuss the supporter's dialogue acts that were positively correlated with the objective idea quantity (Figure 5). *Accept* refers to utterances that accept a proposal. The tendency for the supporter's *accept* to be high when the leader's *suggest* is high can be attributed to the supporter actively accepting the leader's proposals. However, the direction of causality in these correlations could be reversed. For instance, it is possible that a supporter's frequent use of *accept* fosters an environment where the leader feels more comfortable making suggestions. Similarly, it is unclear whether a positive atmosphere created by the supporter fosters more ideas from the leader, or if the supporter's positive reactions are simply a consequence of the leader being highly generative.

5.3 Human–AI Co-creation Design

This study aimed to identify the features of utterances that facilitate idea generation by analyzing dialogue data from a human–human co-creative process. In this section, we discuss dialogue strategies for eliciting human ideas in human–AI co-creation based on our findings. Our results revealed that the quantity of the supporter's ideas had a negative association with the leader's quantity of ideas, suggesting that receiving fewer ideas from a partner may, in fact, stimulate one's own ideation. However, it is crucial to note that this finding is limited to the subjective measure of idea quantity; no negative correlation was found for the objective measure (i.e., between the number of *suggest* acts from leaders and supporters). This suggests the negative relationship may only apply when idea quantity is perceived subjectively. We also found that acceptance and positive opinions from supporters were positively correlated with the leader's idea proposals, indicating that these can be effective supportive strategies. These findings offer important design implications for future co-creative AI systems. For instance, instead of always presenting numerous ideas, an AI could adopt a strategy of accepting the user's proposals and providing positive feedback to better draw out their creativity. Furthermore, we found a positive correlation between the positive opinions of the supporter and the leader's overall evaluation of that partner. This suggests that positive feedback may not only increase the quantity of ideas but also enhance the user's impression of their collaborative partner. However, since our results are based on human-human interaction, whether the same outcome would occur with an AI partner requires future validation. Future work includes examining the causality of the correlations identified in this study and building and evaluating a dialogue system that implements these findings.

6 Conclusion

In this study, we investigated utterance patterns that promote idea generation in human–human co-creative dialogue, aiming to inform the design of AI systems that can better elicit human creativity. We conducted a collaborative story-writing experiment with pairs assigned leader and supporter roles and analyzed the collected dialogue data. Our results revealed the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. We also found that acceptance and positive opinions from supporters were positively correlated with the leader's idea proposals, indicating that these can be effective supportive strategies.

Limitations

This study has several limitations. First, our study's participants were limited to Japanese native speakers. This demographic specificity may restrict the generalizability of our findings.

Second, this study focused on human–human dialogues rather than human–AI interactions. While our ultimate goal is to inform the design of human–AI co-creative systems, we opted for human–human data at this exploratory stage. Building on the findings of this paper, our future work will apply these insights to conduct controlled experiments in human–AI co-creative settings.

Third, as shown in Figure 4, the primary relationships observed were generally weak to moderate correlations, with no strong correlations found.

This may suggest that a complex process like idea generation is influenced by many factors beyond dialogue acts. Future research should consider incorporating factors such as the leader's personality and their proficiency in story writing into the analysis.

Fourth, while we observed a positive correlation between a supporter's DAs and the "Quantity of ideas (objective)," the causal direction of this relationship remains undetermined. It is unclear whether a high frequency of a particular DA from the supporter actively stimulated the leader's idea generation, or if the supporter's DAs (such as acceptance) simply increased in response to the leader proposing a large number of ideas. As part of our future work, establishing this causality will require further experiments designed specifically to distinguish between these two possibilities.

References

Minwook Bae and Hyounghun Kim. 2024. Collective Critics for Creative Story Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 196–207.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. In *International Society for Music Information Retrieval (ISMIR)*.

Gloria Ashiya Katuka, Alexander R. Webber, Joseph B. Wiggins, Kristy Elizabeth Boyer, Brian Magerko, Tom McKlin, and Jason Freeman. 2022. The Relationship between Co-Creative Dialogue and High School Learners' Satisfaction with their Collaborator in Computational Music Remixing. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Mika Koivisto and Simone Grassini. 2024. Author Correction: Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 14.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SceneCraft: Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1):86–96.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Alex Faickney Osborn. 1953. *Applied imagination.* New York: Charles Scribner's Sons.

Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs Understand Dialogues? A Case Study on Dialogue Acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

Peinuan Qin, Chi-Lan Yang, Jingshu Li, Jing Wen, and Yi-Chieh Lee. 2025. Timing Matters: How Using LLMs at Different Timings Influences Writers' Perceptions and Ideation Outcomes in AI-Assisted Ideation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Oliver Schmitt and Daniel Buschek. 2021. Character-Chat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition*, C&C '21, New York, NY, USA. Association for Computing Machinery.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

A Interface for Co-creation

Figure 6 shows the story co-writing interface. The story co-writing interface was built on Google Spreadsheets. Only the leader could write the story on the interface. When a worker clicks on the theme generation button on the interface, it automatically generates and shows the theme of the story being created. The button for automatic theme generation is placed on another sheet to prevent accidental touches. The generated theme is displayed in the yellow area of the interface. Participants create their story based on this theme, entering it sentence by sentence into the designated cells. The input field consists of ten fields per story and a checkbox indicating that the story creation is complete.

B Relationship between the three evaluation criteria and the leader's responses in the postsurvey

Table 3 presents the Spearman rank correlation coefficients between the three evaluation criteria and leader's postsurvey responses. There was a strong correlation between the "Quantity of ideas (subjective)" and generating good ideas. One of Osborn's principles of brainstorming is "focus on quantity, more ideas will increase the likelihood of good ideas" (Osborn, 1953). The obtained result supports the aforementioned Osborn's brainstorming principle. We confirmed that "Quantity of ideas (subjective)" is suitable for evaluating the quality and quantity of the leader's ideas.

Additionally, there were strong correlations between "Good supporter" and each response to Questions 4–9 in the survey. This implied the overall evaluation of the supporter, making the criterion a suitable evaluation criterion.

C Participants

In this section, the demographic information of the 120 workers who participated in this experiment is provided. Of the participants, 60.8% were female and 39.2% were male. Additionally, 7.5% were in their teens or younger, 35.8% were in their 20s, 26.7% were in their 30s, 20.8% were in their 40s, 7.5% were in their 50s, and 1.7% were in their 60s.

D Definition of DAs

Table 4 shows the definition of 17 DAs which were used for annotations.

E Prompt for annotation

Initial dialogue act annotation was performed using GPT-4-turbo. The prompt for annotation consisted of three parts (Figure 7): the annotation manual (task description, tag list, and notes of annotation); a dialogue example created by the authors (input example 1); and the corresponding annotation results for that dialogue (output example 1). The dialogue that we wanted to be annotated was appended to this prompt and then input into the model.

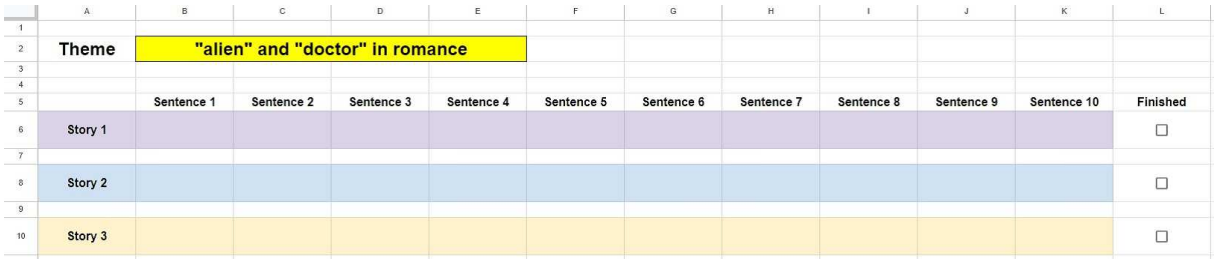


Figure 6: The story co-writing interface was built on Google Spreadsheets. The original interface is written in Japanese. The button for automatic theme generation is placed on another sheet to prevent accidental touches.The yellow area indicates the story's theme. The first story is entered sentence by sentence into the cells of the purple area, the second story into the light blue area, and the third story into the light yellow area.

No.	Postsurvey Item	Subjective Eval.		Objective Eval.
		Good supporter	**Quantity of ideas (subjective)**	**Quantity of ideas (objective)**
Self Evaluation				
1	You generated many ideas	0.26	-	0.07
2	You generated good ideas	0.27	0.78	0.02
3	You made pertinent points	0.10	0.53	−0.04
Partner Evaluation				
4	Your partner generated many ideas	0.56	0.23	0.07
5	Your partner generated good ideas	0.74	0.27	0.12
6	Your partner made pertinent points	0.68	0.31	0.10
7	Your partner was easy to talk to	0.85	0.35	0.13
8	Your partner stimulated your idea generation	0.80	0.32	0.12
9	The conversation with your partner was lively	0.81	0.37	0.11
10	Overall, your partner was a good partner	-	0.26	0.15

Table 3: Relationship between the three evaluation criteria and other responses in the postsurvey

Table 4: The definition of DAs. The underlined part represents the utterance to which the tag is annotated.

Tag Name	Tag Description	Utterance Example
suggest	Proposal of ideas or directions. It is a statement or question sentence.	Let us introduce pirates.
accept	Acceptance of proposals or agreement with the partner's opinion.	It's getting better, isn't it? → Yes.
decline	Rejection of proposals or disagreement with the partner's opinion.	It's getting better, isn't it? → No.
setQuestion	Question without options.	Where is a good setting?
choiceQuestion	Question presenting two or more options.	Are idols female? Male?
answer	Response to a choiceQuestion.	Are idols female? Male? → Male.
confirm	Confirmation of ideas or facts.	Is this okay?
thanking	Expressing gratitude.	Thank you.
positiveOpinion	Positive opinions or impressions.	That's good!
neutralOpinion	Opinions or impressions that are neither positive nor negative.	Punctuation is missing.
negativeOpinion	Negative opinions or impressions.	It doesn't add up.
surprise	Expressing surprise.	Oh!
request	Request to the partner.	Please read.
affirm	Affirmation regarding facts, with no intention of agreeing.	Does it mean like this? → Yes.
deny	Denial regarding facts, with no intention of disagreeing.	Does it mean like this? → No.
action	Declaration of future actions.	I will write it.
other	Anything that does not fit into the above tags.	It's been about 15 min.

Task Description

- This is a task to annotate text dialogue data between two people.
- The content of the dialogue is that two people who are given a theme create a story that matches the theme while brainstorming ideas.
- Please annotate considering the entire context, not just a single utterance.

Tag List

[Definition of DAs]

Notes on Annotation

…

- Output Format: Utterance<Reason for assigning the tag>[Assigned tag]
- Please assign only one tag per utterance. If there are two or more candidate tags, please select the one you think is best.

Input Example 1

…

A: Should the meeting of the princess and the pirate be on a ship?

A: Or in a castle?

B: Let's have it on a ship.

…

Output Example 1

…

A: Should the meeting of the princess and the pirate be on a ship? <Reason: When the same speaker presents choices in two separate utterances, apply the “choiceQuestion" tag to both.>[choiceQuestion]

A: Or in a castle? <Reason: When the same speaker presents choices in two separate utterances, apply the “choiceQuestion" tag to both.>[choiceQuestion]

B: Let's have it on a ship<Reason: Answering a question with choices.>[answer]

…

#Input Example 2

[Dialogue history]

#Ourput Example 2

Figure 7: Prompt for DA annotation. The original prompt is written in Japanese.

Non Idiomatic Conventionalised Expressions: A New Pain in the Neck?

Ganesh Katrapati and Manish Shrivastava
International Institute of Information Technology Hyderabad
ganesh.katrapati@research.iiit.ac.in m.shrivastava@iiit.ac.in

Abstract

Multiword Expressions (MWEs) are linguistic units that are fixed, conventionalised, and function as single semantic units. While idioms have received considerable attention in NLP, another class of conventional expressions like "as far as I know", have remained underexplored. Unlike idioms such as *kick the bucket*, which are characterised by complete non-compositionality, these expressions can often be interpreted compositionally. However, this does not capture their meaning fully. We introduce the term *Non-Idiomatic Conventionalised Expressions (NICEs)* to describe this category of expressions that are largely compositional but retain crucial non-compositional elements. NICEs play important roles in discourse and pragmatics, making their systematic study essential.

We automatically discover and manually refine candidate NICEs, validating their distinct position on the compositionality spectrum through lexical substitution tests. Results show that they differ from regular constructions and that machine translation often renders them inaccurately.

1 Introduction

Multi Word Expressions (MWE) are linguistic forms spanning word boundaries that may have idiosyncratic interpretations (Jackendoff, 1996; Nunberg et al., 1994; Sag et al., 2002). *MWEs* are also generally fixed with little or no variation in their form (Moon, 2002, Calzolari et al., 2002) making them entrenched in language and *conventionalised* in their usage. *Idioms*, such as *"kick the bucket" or "spill the beans"*, are a sub-class of *MWEs* for which the meaning of the whole cannot be predicted from their component words alone i.e. they are semantically *non-compositional*.

Compositionality and Conventionalisation are non-binary and MWEs can be found across both spectra. At one end of the compositionality spectrum, we have completely idiomatic expressions (*Idioms*) while near the other end are fully conventionalised *MWEs* where all the words contribute to the meaning of the expression (Baldwin et al., 2003, Reddy et al., 2011). For instance, *Conjunct Verbs* like *"take a bath"* or Named Entities like "International Institute of Information Technology" which is perfectly compositional but acts as a single unit nonetheless.

Expressions such as *"as far as I know"*, *"as a matter of course"*, *"at one time or another"* are fixed and occur in predictably specific contexts. They are frequently used by proficient speakers and are considered to be important for successful participation in a linguistic community (Coulmas, 1979, Wray, 2000, Yorio, 1989, Edmonds, 2014). They also show a high degree of conventionalisation - the words in such expressions are not easily substitutable. In contrast to *Idioms*, they usually have a compositional interpretation but, while being seemingly semantically transparent, the usage of these expressions in language cannot be fully predicted from their words alone.

Consider the following sentence :

(a) Her father raised his hand once, hardly sparing ***a second glance*** .

Here, the expression *"a second glance"* is not considered an idiom (Haagsma et al., 2020), and indeed, one could interpret it compositionally but that would result in an incomplete understanding.

We categorise such expressions as *Non Idiomatic Conventionalised Expressions (NICE)* and situate them in the spectrum of compositionality. We draw inspiration from the Cognitive and Construction Grammar traditions (Langacker, 2008, Tyler, 2005, Wulff, 2013) which define all constructions as conventionalised grammatical patterns.

In the following sections, we briefly look at related work (2). We then describe our approach to automatically discover and extract candidate *NICEs* (3) and the filtering process (4). To validate our claim, We present two measures based on substitution tests (5) which show that conventionalised expressions exhibit a distinct behaviour when compared to regular compositional constructions. Finally, we demonstrate that machine translation systems frequently fail to capture NICEs accurately, leading to partial or misleading renderings.

2 Related Work

Most recent research on *MWEs* primarily centers on *Idioms*, particularly on determining whether a given *Potentially Idiomatic Expression* is used literally or idiomatically in context (Haagsma et al., 2019, 2020; Sporleder and Li, 2009a; Fazly et al., 2009a). These approaches typically depend on pre-defined lexical resources or idiom dictionaries (Moon, 1998; Cowie, 1993), thereby avoiding any automatic identification of new or unseen *MWEs*. Nonetheless, a subset of studies has addressed the problem of *MWE discovery*, which is of particular relevance to this paper.

There are two main approaches for MWE discovery. One approach relies on the *fixedness* property of MWEs and estimates the degree of association and co-dependency between its component words using Pointwise Mutual Information(PMI) (Church and Hanks, 1989) and its modified formulations (Fazly and Stevenson, 2006; Bouma, 2009) along with other statistical measures (Pecina and Schlesinger, 2006).

On the other hand, Baldwin et al. (2003); Katz and Giesbrecht (2006); Kiela and Clark (2013); Salehi et al. (2015) base their work on the *non-compositionality* of *MWEs*. They use distributional semantic models to compute the difference between *MWE* representation and the representations of its component words as a measure of the degree of its non-compositionality. Our work blends these two approaches for discovering candidate expressions.

While several types of Multi Word Expressions such as Verb-Noun constructions, Verb-Particle constructions, Noun Compounds and Idioms have been explored, non idiomatic conventionalised expressions (NICE) are relegated to the generic class of *collocations* (Sag et al., 2002) which can include *any* combination of words with a high enough degree of association.

Phrasal collocations also called *Phraseological Units*, *Lexical Bundles* are loosely defined as multi word collocations at the phrase level. They too have been mined from raw corpora using association measures (Colson, 2017, Hyland, 2008). However, these collocations are not differentiated by their compositionality and the categorisation of NICEs simply as collocations or lexical bundles does not capture the fact that these expressions are not completely compositional in nature.

3 Approach

We use the *British National Corpus (BNC)* (BNC, 2007) to extract candidate expressions. We find it ideal because of its size, variety and availability. It is also widely used in the area of idiom discovery and processing (Fazly et al., 2009b, Sporleder et al., 2010, Sporleder and Li, 2009b, Salton et al., 2017, Haagsma et al., 2019).

To extract candidate expressions, we follow a PMI-based approach (Church and Hanks, 1989; Fazly and Stevenson, 2006), scanning the corpus sequentially and merging word pairs (w_a, w_b) if their PMI exceeds a threshold θ [1]. Each merge creates a new candidate $(w_a\_w_b)$, and the algorithm proceeds greedily.

Since this over-generates candidates, we prune them using a *Minimum Description Length (MDL)* criterion, which balances data likelihood and model size. *MDL* has been widely used in unsupervised morphological segmentation (Creutz and Lagus, 2002), clustering (Li and Abe, 1992), and text pattern mining (Wu et al., 2010). Our pruning iteratively evaluates candidates and retains only those offering a cost benefit.

3.1 Ranking

We consider *NICEs* as constructions that are largely compositional but retain some degree of non-compositionality, which gives them distinct distributional patterns. To capture this, we segment the *BNC* so that each candidate expression forms a single token[2]. We then train a *word2vec* model (Mikolov et al., 2013) on the segmented corpus and compute a *compositionality score*: the cosine distance between a candidate's embedding and the normalised sum of its component embed-

[1] We set $\theta = 10$

[2] Words are joined with underscores

dings. Candidates with higher scores (lower compositionality) are ranked and selected for further analysis.

It is worth noting that when we tried this approach using Contextualised Embeddings (Devlin et al., 2019), the resultant score did not reflect the compositionality of the candidates in any way. Our results echo those of Pickard (2020); Cordeiro et al. (2016); Nandakumar et al. (2018) in this matter and this merits further observation.

4 Filtering

From the ranked candidate list, we removed Named Entities, Nominal Compounds, Conjunct/Compound Verbs, and Idioms (Haagsma et al., 2020), as they are not within our scope. Three computational linguists proficient in English then filtered the list manually. An expression was marked as *NICE* if substitutions made it less natural (e.g., *"a moment or two"* vs. *"a moment or three"*) or altered its meaning entirely (e.g., replacing *little* with *less* in *"with little difficulty"*). As we could not perform an exhaustive filtering of all expressions, the top *300* ranked expressions were selected; out of these, *90* expressions were tagged as *NICEs* by all the validators [3] . *We submit that this is not at all a comprehensive list of NICEs but perhaps sufficient to illustrate the topic of our study.*

5 Measures

Here, we describe two measures to test the results of our approach against our definition of *NICEs*. Both the measures are based on substitution tests designed to compare *NICEs* with regular language usage.

For both the measures, we randomly selected a list of (a) Adjective-Noun (AN) pairs, (b) Verb-Determiner-Noun (VN) combinations and (c) NGrams (Random) of length 3, and compare them with the *NICEs*.

5.1 Naturalness

As noted in (4), substitutions within a conventionalised expression reduce its *naturalness*. In a corpus, such valid substitutions are therefore rare. To measure this, we take the nearest neighbours of each word in an expression, generate leave-one-out substitutions, and compute their occurrence

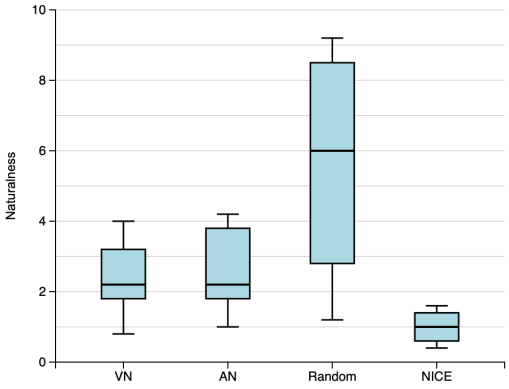


Figure 1: A box plot for naturalness metric across different types of expressions

counts along with the number of unique substitutions found.

$$Sub_N(E) = C(E) * CR(E) * MR(E) \quad (1)$$

$$CR(E) = \frac{\sum_{i \in S_E} C(s_i)}{C(E)} \quad (2)$$

$$MR(E) = \frac{\sum_{i \in S_E} min(1, C(s_i))}{S_E} \quad (3)$$

Where $C(E)$ is the frequency of expression E in the corpus, S_E is the set of possible substitutions, and $CR(E)$ is the ratio of all substitution counts to that of E (eq. 2). We also measure substitution variability $MR(E)$ (eq. 3) as the proportion of substitutions found in the corpus out of all possible ones, and finally normalise by the word length of E.

5.2 Semantic Shift

NICEs typically occur in contexts not shared by their substituted variants. To capture this, we use a transformer-based language model[4] to obtain sentence embeddings for both the *NICE* instances and their substitutions, and compute cosine distance as a measure of semantic similarity.

$$Sub_S(E) = \frac{1}{S_E} \sum_{i \in S_E} distance(\overrightarrow{V_E}, \overrightarrow{V_{S_i}}) \quad (4)$$

Where $\overrightarrow{V_E}$ and $\overrightarrow{V_{S_i}}$ are averaged sentence embeddings of the *NICE* (E) and the substituted expression S_i respectively. The measures show consistently lower means across all quartiles for conventionalised expressions compared to other types

[3] You can download all the data and the MT evaluations at https://github.com/neshkatrapati/nice-data

[4] We use https://www.sbert.net/ (Reimers and Gurevych, 2019)

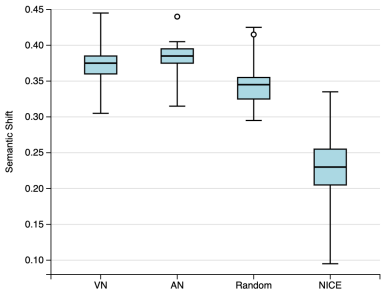


Figure 2: A box plot for semantic shift metric across different types of expressions

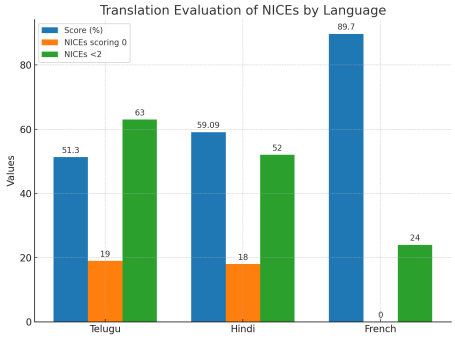


Figure 3: Translation Evaluation of NICEs by Language

(Figures 1, 2), supporting our claim that *NICEs* cannot be substituted without losing naturalness or altering meaning.

6 Impact on Machine Translation : A Preliminary Study

To investigate the effect *NICEs* may have on NLP applications, we chose to explore Machine Translation and performed a *preliminary study.*

We chose three languages, each with varying links to English. *French* for familiar syntax and shared cultural aspects, *Hindi*, an Indo-Aryan language which shares some structural features (same language family), and finally *Telugu*, a Dravidian language with marked diverge in morpho-syntax when compared to English.

We designed a translation evaluation task to assess how well NICEs are preserved across languages. Annotators were given English source sentences (taken from *BNC*) with a highlighted NICE, along with corresponding translations from an MT system [5]. Their task was to judge only the translation of the NICE , independent of the rest of the sentence. Each item was rated on a 0, 1, 2 scale: 0 for incorrect or literal translations, 1 for partially correct or unnatural ones, and 2 for correct and natural renderings. Each language was annotated by one annotator who was proficient in the respective language and also a linguist. The evaluation set consists of *220* sentences averaging *2 to 3* examples per *NICE*.

6.1 Results

French translations performed best with 89.7%, which is expected given shared linguistic and cultural structures with English. In contrast, results dropped sharply for Indic languages: Hindi, despite being in the same language family, scored only 59%, and Telugu performed worst at 51.3%. This highlights the challenge of preserving NICEs across languages with less structural and cultural overlap. This cannot be attributed solely to data scarcity. English-Indic language pairs have considerable parallel corpora (Kunchukuttan et al., 2018; Ramesh et al., 2022), more so, in the case of Hindi. It should also be noted that *NICEs* are not exactly rare phenomena.

Moreover, as you can see in Fig.3, a number of expressions are unsatisfactorily translated and still some others entirely missed or mis-translated. Even French, which has a high overall score, had 24 out of 90 expressions with less than perfect score.

7 Conclusion

Our findings highlight that *NICEs* cannot be simply subsumed under broader categories such as *MWEs*, *Lexical Bundles*, or *Phrasal Units*. Their partially non-compositional meaning requires explicit recognition and treatment as a distinct linguistic phenomenon. Ignoring this dimension risks overlooking essential aspects of their semantics, which can undermine the performance of downstream NLP tasks as evidenced by our MT results.

Limitations

Our study has several limitations. The set of *NICEs* examined is relatively small and not exhaustive, and validation relied on only a few annotators, with limited language coverage. Nonetheless, the consistency of results across different languages and annotators suggests that the observed patterns are robust and worth pursuing in future, larger-scale studies.

[5]We used Google Translate which is one of the most widely used MT system

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. pages 89–96. Association for Computational Linguistics.

BNC. 2007. British national corpus, XML edition. Oxford Text Archive.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. 16(1):76–83.

Jean Pierre Colson. 2017. The idiomsearch experiment: extracting phraseology from a probabilistic network of constructions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10596 LNAI:16–28.

Silvio Cordeiro, Carlos Ramisch, Marco A. Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997. ACL.

Florian Coulmas. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics*, 3(3-4):239–266.

A. P. Cowie. 1993. *Oxford Advanced Learners Dictionary of Current English*, 5th edition. Oxford University Press, Oxford.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. pages 21–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Amanda Edmonds. 2014. Conventional expressions. *Studies in Second Language Acquisition*, 36(1):69–99.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009a. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009b. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 337–344.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (May):279–287.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a Wide Net: Robust Extraction of Potentially Idiomatic Expressions. pages 1–34.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Ray Jackendoff. 1996. The architecture of the language faculty.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432, Seattle, Washington, USA. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english–hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ronald Langacker. 2008. *Cognitive Grammar: A Basic Introduction*, volume 12. Oxford University Press.

Hang Li and Naoki Abe. 1992. Clustering Words with MDL Priciple. *Journal of Natural Language Processing*, 4(2):71–88.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint*.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press, Oxford.

Rosamund Moon. 2002. *Fixed Expressions and Idioms in English: A Corpus-Based Approach (review)*, volume 78.

Anoop Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. How well can we predict multiword expression compositionality using embedding-based methods? In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 73–81. ALTA.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 651–658, Sydney, Australia. Association for Computational Linguistics.

Thomas Pickard. 2020. Comparing word2vec and glove for automatic measurement of mwe compositionality. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE+ELex 2020)*, pages 95–100, Barcelona, Spain (Online).

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. *IJCNLP 2011 - Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *CICLing*.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 977–983.

Giancarlo Salton, Robert Ross, and John Kelleher. 2017. Idiom type identification with smoothed lexical features and a maximum margin classifier. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 642–651, Varna, Bulgaria. INCOMA Ltd.

Caroline Sporleder and Linlin Li. 2009a. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009b. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Caroline Sporleder, Linlin Li, Philip John Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 639–646.

Andrea Tyler. 2005. Cognitive grammar. *Studies in Second Language Acquisition*, 27:650 – 651.

A Wray. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4):463–489.

Ke Wu, Jiangsheng Yu, Hanpin Wang, and Fei Cheng. 2010. Unsupervised text pattern learning using minimum description length. *2010 4th International Universal Communication Symposium, IUCS 2010 - Proceedings*, pages 161–166.

Stefanie Wulff. 2013. Words and idioms.

Carlos A. Yorio. 1989. *Idiomaticity as an indicator of second language proficiency*, page 5572. Cambridge University Press.

Forecasting Time Series with LLMs via Patch-Based Prompting and Decomposition

Mayank Bumb[1], **Anshul Vemulapalli**[1], **Sri Harsha Jella**[1], **Anish Gupta**[1], **An La**[1], **Ryan Rossi**[2], **Franck Dernoncourt**[2], **Hongjie Chen**[3], **Nesreen Ahmed**[4], **Yu Wang**[5]
[1]University of Massachusetts Amherst, [2]Adobe, [3]Dolby Labs, [4]Intel, [5]University of Oregon

Abstract

Recent advances in Large Language Models (LLMs) have demonstrated new possibilities for accurate and efficient time series analysis, but prior work often required heavy fine-tuning and/or ignored inter-series correlations. In this work, we explore simple and flexible prompt-based strategies that enable LLMs to perform time series forecasting without extensive retraining or the use of a complex external architecture. Through the exploration of specialized prompting methods that leverage time series decomposition, patch-based tokenization, and similarity-based neighbor augmentation, we find that it is possible to enhance LLM forecasting quality while maintaining simplicity and requiring minimal preprocessing of data. To this end, we propose our own method, PatchInstruct, which enables LLMs to make precise and effective predictions.

1 Introduction

Time-series forecasting (TSF) has a broad range of applications in agriculture, business, epidemiology, finance, etc. Many of these applications require robust predictions of time series, and accurately modeling the dependencies between variables remains to be a challenge (Shao et al., 2020). Traditional forecasting models such as ARIMA, LSTMs, and even Transformer/Graph-based architectures have displayed a strong performance on these tasks (Zhou et al., 2024).

More recently, Large Language Models (LLMs) have shown a promising future in modeling time series, with accurate predictions that rival state of the art (SOTA) methods, due to their strengths in pattern recognition, sequence modeling, and generalization across tasks. However, current LLM-based methods often rely on complex architectures or require heavy fine-tuning, limiting their scalability to real-world applications.

One prominent approach, S^2IP-LLM (Pan et al., 2024), embeds time series into a semantic space to enhance forecasting performance. While effective, it introduces two key limitations. First, it incurs a high computational cost during inference due to its reliance on complex decomposition and patching pipelines. Second, it does not explicitly model dependencies across related time series, which can be critical in domains such as traffic and energy forecasting where inter-series relationships play a significant role.

We aim to develop a method (see Figure 1) that maintains the predictive strength of LLM-based models while addressing the above limitations of inference speed and generalization. Therefore we guide our experimentation around the idea of whether we can create general-purpose prompts that guide LLMs to forecast time series both accurately and efficiently, without requiring model fine-tuning or architectural changes.

To this end, we introduce PatchInstruct, a prompt-based framework that tokenizes time series data into meaningful patches that encapsulate temporally relevant patterns and guides the LLM via structured natural language instructions to output precise predictions. Unlike prior work, PatchInstruct requires no model retraining or architecture modification and also significantly reduces inference time (in comparison to the baseline and complex architectures) alongside token usage while preserving or improving accuracy.

We compare PatchInstruct with several other prompting strategies—including Zero-shot, Neighbors, and PatchInstruct + Neighbors—and evaluate them on diverse, real-world datasets (Weather and Traffic), primarily using GPT-4 and GPT-4o as the LLM backbones.

Across the datasets and small forecasting horizons we study ($H \leq 12$), PatchInstruct is typically the most accurate among our baselines in MSE/-MAE and, in our setup, reduces inference overhead by 10x–100x compared to S^2IP-LLM while maintaining comparable accuracy. Neighbor augmen-

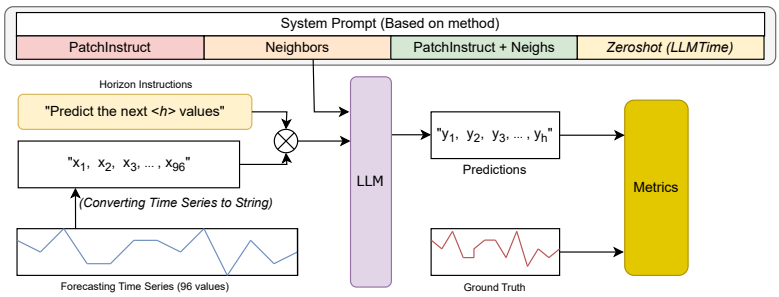


Figure 1: LLM-based Time-Series Forecasting Pipeline

tation is dataset-dependent: it helps on Weather, where retrieved series are highly informative, but can underperform on datasets with weaker cross-series alignment.

Taken together, these results indicate that careful prompt design can substitute for portions of model-specific architecture in our tested setting, enabling scalable and domain-adaptable time-series forecasting with LLMs. Our evaluations standardize on GPT for output-format reliability (other backbones frequently violated the required H-length outputs in pilots), and we fix the neighbor count to ($k = 5$) to fit a 50k input-token budget; broader cross-backbone and budget studies are left for future work.

2 Related Work

2.1 Time Series Foundation Models

Foundation Models (FMs)—large pre-trained models that learn general-purpose representations—have propelled state-of-the-art results in NLP and CV, and the same paradigm is increasingly adapted to time series (Shi et al., 2024). Liang *et al.* provide a useful taxonomy for Time Series Foundation Models (TSFMs) along four axes: data category (standard, spatial, or trajectory/event), model architecture (Transformer-, non-Transformer-, or diffusion-based), pre-training strategy (self-supervised vs. supervised), and application domain (Liang et al., 2024).

While diverse architectures exist, Transformer backbones remain prevalent due to their ability to capture long-range dependencies in sequential data (Miller et al., 2024). Diffusion-style foundations have also emerged, e.g., TimeDiT, which marries diffusion objectives with Transformer blocks for time series analysis (Cao et al., 2024b). Among open and commercial TSFMs, models such as Lag-Llama demonstrate that large-scale pretraining across heterogeneous collections improves adaptability and zero/few-shot forecasting quality (Rasul et al., 2023). In practice, the choice of pre-training signal (contrastive, masked reconstruction, forecasting-style objectives) and coverage of data domains strongly influences downstream generalization.

2.2 LLMs for Time Series Forecasting

Large Language Models (LLMs) have recently been explored for time series by casting numerical forecasting as a language problem via tokenization, prompting, and in-context learning. Foundational work on Transformer-based forecasting for raw continuous inputs includes TST (Zerveas et al., 2020), which applies a Transformer encoder to multivariate sequences. Subsequent advances such as PatchTST (Nie et al., 2022) introduce *patching*—partitioning series into localized segments (patch tokens)—and channel-independence, which together improve efficiency and accuracy. Thus, PatchTST *builds on* earlier Transformer formulations like TST; our discussion reflects this chronology.

A parallel line of research converts real-valued series into discrete tokens to better leverage the next-token prediction strengths of LLMs. Chronos (Ansari et al., 2024) quantizes and scales observations into a fixed vocabulary to enable zero-shot and transfer settings. Digit-level tokenization treats each numeric digit as a token, aligning forecasting with language modeling mechanics (Gruver et al., 2024). Prompt-based formulations go further: PromptCast frames forecasting as sentence-to-sentence generation with task-specific prompts (Xue and Salim, 2023), and GPT4TS demonstrates that a single LLM can address forecasting, anomaly detection, and classification using textual prompts alone (Zhou et al., 2023).

Despite these advances, robustness on hetero-

geneous, irregular, and partially observed series remains challenging. LLM4TS proposes a two-stage pipeline that first aligns pretrained LLMs to standardized time series structures and then fine-tunes for forecasting (Chang et al., 2024). Decomposition-aware prompting (e.g., TEMPO) explicitly models trend/seasonal/residual components to improve interpretability and performance (Cao et al., 2024a). Hybrid systems integrate spatial and relational biases—TPLLM fuses CNNs/GCNs with LLMs for traffic prediction (Ren et al., 2024), while GenTKG combines retrieval-augmented generation with parameter-efficient tuning for temporal knowledge graphs (Liao et al., 2024).

In contrast to methods that rely on heavy fine-tuning or complex multi-component stacks, our work (PATCHINSTRUCT) adopts a training-free prompting framework: we decompose inputs into compact *patch* segments and instruct LLMs directly. This minimizes token usage and implementation overhead while preserving competitive accuracy—particularly on short-horizon forecasts—across diverse domains.

3 Methodology

We propose an approach that leverages Large Language Models (LLMs) for time series forecasting through specialized prompt engineering techniques that eliminates the need for model fine-tuning or architectural modifications.

Our approach begins with a zero-shot baseline, inspired by TimeLLM (Gruver et al., 2024), where the model is prompted with raw historical time series values and tasked with predicting future values. While this baseline offers simplicity and generality, it lacks the inductive bias necessary to capture local temporal dynamics, leading to suboptimal performance in complex forecasting settings. Our second baseline S^2IP-LLM introduces significant inference-time overhead due to its reliance on complex decomposition pipelines and fine-tuning, limiting its scalability in real-world deployments.

To address these limitations, we introduce PatchInstruct (see Figure 2), a prompting strategy that encode temporal structure through patch-based representations, and provide pretrained LLMs more context on the dataset. The core idea is to decompose a time series into fixed-length overlapping patches and provide them to the LLM in a structured format, along with instructions to predict future values.

Additionally, we experimented the model's forecasting ability by supplementing the target time series with a small set of similar time series referred to as Neighbors (Neighs). Specifically, we select the five most similar time series from the dataset's past seen data, referred to as neighbors. The motivation behind this approach is to provide the LLM with additional contextual signals and recurring patterns that may not be fully observable in the target series alone.

We finally also tested a combination of these the Patch-Instruct and Neighbors strategy, enriching the prompt with structurally decomposed information from the target series (via patching), while augmenting it with relevant patterns from similar series (via nearest neighbors).

In the following subsections, we detail the construction of each approach, describe the datasets and evaluation metrics used, and present a comparative analysis of their forecasting performance.

We evaluate our method on two time series datasets: Weather and Traffic. These datasets comprised of continuous measurements sampled at regular intervals. A 96-timestep input window is used to forecast future horizons of 1,2,3,4,5,6 and 12 steps.

3.1 Overview of Framework

Our framework is designed to adapt large language models (LLMs) for time series forecasting without any fine-tuning, using carefully structured prompts that condition the model with temporal data and forecasting instructions. The pipeline is modular and supports multiple prompting strategies, including PatchInstruct, Neighbors, and PatchInstruct + Neighbors, and Zeroshot by modifying the structure of the system prompt and the input representation.

At inference time, a raw time series is converted into a sequence of string-formatted numerical values. Depending on the method, additional transformations are applied—for example, decomposing the sequence into overlapping fixed-length patches (in PatchInstruct), or retrieving similar time series (in Neighs). These inputs are concatenated with forecasting instructions (e.g., "Predict the next h values") and passed to the LLM. The output is parsed into a numerical forecast and compared with the ground truth using standard forecasting metrics.

In addition to forecasting, PatchInstruct also

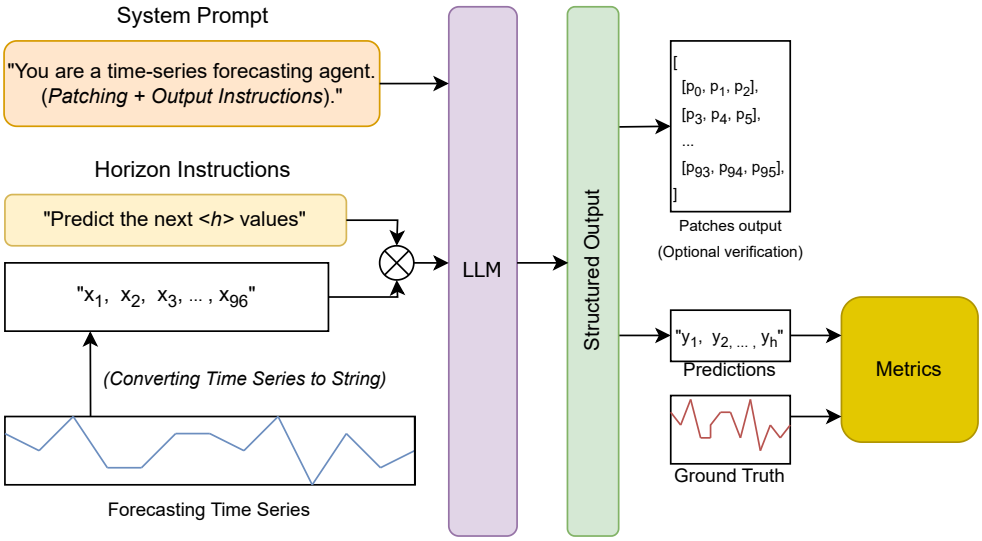


Figure 2: PatchInstruct Forecasting Pipeline

prompts the model to output reconstructed patches from the input, enabling an optional interpretability step. These predicted patches can be compared with the actual ones to assess whether the LLM is learning meaningful temporal structure and capturing local dynamics, thus providing deeper insight into the model's understanding of the task.

3.2 Prompt Design

We now delve deeper into the specific construction of each prompting strategy. This section provides detailed formulations illustrating how time series data, patching instructions, and neighboring trends are encoded within the input to the LLM. The prompts were designed through rigorous empirical testing to ensure clarity and effectiveness. Each prompt consists of a system prompt, which defines the forecasting method and describes how we construct the series, and a user prompt, which contains the actual time series data provided to the LLM for prediction.

PatchInstruct is built upon the zeroshot prompt inspired by LLMTime, this method decomposes the time series into patches and the LLM uses them to form predictions. Below, we outline the structure of the best Patch-Instruct prompt. Additional experiments exploring alternative patching strategies are presented in the Appendix A. The following prompt is a user prompt used for most methods to specify the horizon, and give the context window to the LLM.

Horizon Prompt (Input Time Series)

Continue the following sequence without producing any additional text. Sequence: $<x_1, x_2, x_3, ..., x_{96}>$. Predict the next 3 values.

PatchInstruct System Prompt

You are a forecasting assistant that sees time series data. The sequence represents the total regional humidity measured every 10 minutes. Task: (1) Split the series into overlapping patches with window size 3 and stride 1. (2) Generate the patches in natural order, then reverse the list so the most recent patch appears first. (3) Use these patch tokens to forecast the next 3 values.
Output format:

```
Patches:
[[latest_patch], ..., [oldest_patch]]
Prediction:
[y1, y2, y3]
```

No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

For the System prompt, we included both the patching instructions and dataset-specific information, such as the name of the series. Among the variants of prompts tested, we found reverse patching to perform the best. Further details and comparisions of patching strategies are provided in (Section A; see appendix for Table 5).

Neighs is built upon our zero-shot prompt. This method adds closest neighboring series in terms of

euclidean distance over all the past windows of data and construct a composite prompt by giving all the 5 neighboring prompts as additional context. We outline the structure of the Neighs prompt used for the weather dataset in the "Neighs System Prompt" box. We specified the number of neighbors that, and gave the model additional instructions.

PatchInstruct+Neighs integrates the strengths of both PatchInstruct and Neighs approaches. We combined the two methods using the system prompt in the "PatchInstruct+Neighs system prompt" box.

Neighs System Prompt

You are a forecasting assistant that sees time series data. The sequence represents the total regional humidity measured every 10 minutes. You will also be given 5 neighbor time-series similar to the one to forecast. Use it to understand the trends.
Output format: `[y1, y2, y3]`
No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

PatchInstruct + Neighs System Prompt

You are a forecasting assistant that sees time series data. The sequence represents the total regional humidity measured every 10 minutes. You will also be given 5 neighbor time-series similar to the one to forecast. Use it to understand the trends.
Task: (1) Split the series into overlapping patches with window size 3 and stride 1. (2) Generate the patches in natural order, then reverse the list so the most recent patch appears first. (3) Use these patch tokens to forecast the next 3 values.
Output format:

```
Patches:
[[latest_patch], ..., [oldest_patch]]
Prediction:
[y1, y2, y3]
```

No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

In summary, these prompt-based variants allow us to systematically assess the impact of explicit instructions for time series decomposition, patching, and neighbor augmentation within LLM-based forecasting frameworks. The results, presented in Table 4, provide a comparative analysis of these prompting strategies.

3.3 Evaluation

We evaluate forecasting performance using Mean Squared Error (MSE), Mean Absolute Error (MAE). Runtime Efficiency and Input/Output token usage.

4 Experiments

In this section, we design experiments to investigate our proposed patch instruct framework.

4.1 Backbone selection

PatchInstruct is training-free and works with any instruction-following LLM in principle. In practice, our pilot runs compared GPT, Gemini, and Llama. Both Gemini and Llama exhibited inconsistent decoding for multi-step forecasting—specifically, failing to return exactly H predictions at a given horizon despite explicit instructions and formatting templates. These output-format errors impeded fair comparison and large-scale evaluation. We therefore standardize on GPT for all reported results to ensure consistent generation of horizon-length prediction vectors.

4.2 Datasets

We evaluate our approach on two real-world datasets: Weather and Traffic. Weather captures fast-changing environmental conditions, while Traffic reflects urban flow patterns with spatial and temporal dependencies

The Weather dataset is collected from a meteorological station at the Max Planck Institute for Biogeochemistry (Jena, Germany). It contains 14 meteorological features, including temperature, humidity, and atmospheric pressure measurements. The high-frequency recordings capture intricate weather dynamics critical for testing short-term forecasting precision.

The Traffic dataset consists of sensor network data from Los Angeles, collected between March and June 2012. It records traffic flow rates and congestion patterns across urban arteries. The spatial-temporal correlations in this dataset test the model's ability to capture complex topological dependencies in transportation systems.

We summarize key statistics in Table 1. This selection provides systematic coverage of (1) different sampling frequencies, (2) variable sequence lengths, and (3) heterogeneous feature

interactions—three critical axes for stress-testing tokenization strategies in temporal learning tasks. The datasets' public availability ensures reproducibility, while their domain diversity demonstrates our method's generalizability beyond narrow application contexts.

Dataset	Features	Frequency	Time Span	Samples	Value Range
WEATHER	14	10 minutes	3 years	157,680	0.5–18.13
TRAFFIC	181	Hourly	4 months	34,172	2.5–70

Table 1: Summary of datasets used in our experiments.

4.3 Main Results

For our experiments, we adopt S^{2}IP-LLM as the primary baseline, a method that aligns time series embeddings with the semantic space of a pre-trained LLM through a tokenization framework. While effective, S^{2}IP-LLM suffers from significant computational overhead, requiring extensive training and inference time due to its fine-tuning of LLM components. All methods are evaluated in a consistent zero-shot setting without model retraining to isolate the impact of prompting strategies.

Using various prompting strategies, we instructed a pre-trained LLM to consider time-series patches and utilize them for forecasting without any additional fine-tuning or retraining. Our experiments demonstrate that such patch-based prompting methods can significantly improve forecasting performance across multiple datasets and over shorter horizons. In contrast to models like S^{2}IP-LLM, which rely on explicit decomposition, semantic alignment, and parameter tuning, our approach leverages instruction-tuned LLMs. Among all the strategies evaluated the PatchInstruct technique consistently delivered the best results. This suggests that prompting pre-trained LLMs with thoughtfully structured temporal context can match or even surpass models trained from scratch, offering a lightweight yet effective alternative for time-series forecasting.

Table 3 presents a performance comparison between S^{2}IPLLM (the baseline) and our best-performing Patch Instruct method across multiple time series datasets and forecast horizons. The results clearly indicate that Patch Instruct consistently outperforms the baseline in terms of both MSE and MAE. Finally, Table 3 compares the two methods in terms of input/output token counts and computation time. The analysis reveals that Patch Instruct not only improves forecasting accuracy but also significantly reduces computational overhead. Overall, this comparison highlights the efficiency and effectiveness of the Patch Instruct method.

Dataset	Horizon	S^{2}IP-LLM		Zeroshot		PatchInstruct	
		MSE	MAE	MSE	MAE	MSE	MAE
WEATHER	1	0.0095	0.056	0.0028	0.043	**0.0014**	**0.029**
	2	0.017	0.077	0.0085	0.072	**0.0076**	**0.067**
	3	0.0238	0.0875	**0.0106**	**0.068**	0.0110	0.085
	4	0.0326	0.1051	**0.0115**	**0.085**	0.0236	0.113
	5	0.0371	0.1120	0.0277	0.1	**0.0159**	**0.094**
	6	0.0439	0.1228	0.0204	0.11	**0.0101**	**0.083**
	12	0.0904	0.1823	0.1098	0.221	**0.0436**	**0.137**
TRAFFIC	1	21.0814	**2.4067**	43.49	3.18	**20.05**	2.76
	2	24.0935	2.4919	23.47	2.53	**9.38**	**1.89**
	3	29.9573	2.7849	22.38	2.54	**6.47**	**1.78**
	4	29.8382	2.6147	27.50	2.87	**11.15**	**1.89**
	5	36.0289	2.7971	34.27	3.20	**8.46**	**1.88**
	6	42.3193	3.0444	29.66	3.07	**25.59**	**2.75**
	12	**68.7149**	**3.8473**	296.20	7.72	235.75	5.89

Table 2: Results comparing our approach to baselines.

Dataset	Horizon	S^{2}IP-LLM			Zeroshot			PatchInstruct		
		Time (s)	IT	OT	Time (s)	IT	OT	Time (s)	IT	OT
WEATHER	1	535.42	7	1	1.36	7370	80	1.24	8500	80
	2	518.45	7	2	1.06	7370	120	1.20	8500	120
	3	533.73	7	3	1.20	7370	160	1.01	8500	160
	4	518.37	7	4	1.01	7370	200	1.16	8500	196
	5	537.85	7	5	1.14	7370	240	1.29	8500	216
	6	522.43	7	6	1.35	7370	280	2.05	8500	280
	12	558.67	7	12	1.44	7370	520	1.51	8500	499
TRAFFIC	1	50.94	7	1	2.59	7360	80	1.31	7950	86
	2	52.88	7	2	2.04	7360	116	1.05	7950	134
	3	55.34	7	3	1.17	7360	160	1.11	7950	185
	4	51.00	7	4	2.13	7360	200	1.17	7950	223
	5	49.96	7	5	1.14	7360	240	1.12	7950	239
	6	50.11	7	6	1.38	7360	268	1.23	7950	336
	12	52.66	7	12	1.78	7360	520	1.36	7950	558

Table 3: Token and Time Comparison for Forecasting.

4.4 Cost vs. Performance Analysis

Our instruction-based forecasts deliberately spend more input tokens than the baseline S^{2}IP-LLM. Across the prompt variants, a single prediction consumes about $\approx$ 800 - 1000 input tokens. By contrast, S^{2}IP-LLM needs only the horizon-length of output tokens once its patch encoder has been trained. The extra prompt length therefore represents about a 100 times increase in in front-loaded cost.

Because our method relies on an already-trained LLM and does no task-specific fine-tuning, the end-to-end latency of producing a forecast collapses from minutes to just seconds. For example, on the Weather dataset at horizon $= 1$, S^{2}IP-LLM requires 535 s, whereas Reverse Patch returns the prediction in 0.86 (see Table 3). Thus, even after accounting

for the larger prompt, our approach is two to three orders of magnitude faster in real-time settings.

The additional 800–900 input tokens result in a substantial improvement in short-range forecasting accuracy. On Weather (H=1), mean squared error (MSE) drops from 1.15×10^{-2} to 2.6×10^{-4} (a 97.7% reduction), and mean absolute error (MAE) decreases from 6.52×10^{-2} to 1.4×10^{-2}. Similar improvements are observed on Traffic, where the MSE at the same horizon is reduced by 85%, indicating that the gains generalize across domains.

Given (i) the low marginal price of LLM tokens relative to GPU training hours, and (ii) the consistent short-horizon error reductions that are operationally most valuable, the accuracy and latency benefits comfortably offset the larger prompt size. Hence trading cheap tokens for immediate, higher-quality forecasts yields a more favorable cost–performance envelope than the current state of the art, especially when rapid deployment and low engineering overhead are priorities.

4.5 Neighbor Results

In order to understand whether incorporating neighboring time-series into our PatchInstruct approach leads to better performance we compare our results for PatchInstruct and Neighs across multiple datasets (Table 4).

We cap the input at 50k tokens to control latency and cost across datasets. Given our prompt structure (task description, target series patches, and retrieved neighbors), this budget allows at most ($k = 5$) neighbors without truncation on our longest contexts. We thus fix the number of neighbors for all experiments to maximize usable contextual evidence while remaining within the token limit.

Both the Neighs and PatchInstruct+Neighs prompting strategies demonstrate clear improvements over the S^{2}IP-LLM baseline across datasets. As seen in Table 4, using Neighs alone often improves performance over PatchInstruct, particularly in the Weather dataset. For example, at horizon 2, Neighs reduces the MSE from 0.0076 (PatchInstruct) to 0.0039 and MAE from 0.067 to 0.051. At horizon 4, Neighs again performs better with an MSE of 0.0172 compared to 0.0236. These gains suggest that incorporating neighboring series can help the model infer more accurate trends by providing contextual information beyond the target sequence itself.

However, this is not universally true. In some cases, Neighs and PatchInstruct+Neighs underperform compared to PatchInstruct. For instance, in the Traffic dataset at horizon 5, Neighs shows a significant degradation, increasing the MSE from 8.46 (PatchInstruct) to 43.50, and PatchInstruct+Neighs to 36.43. This indicates that when neighbor series are less correlated, they can introduce confusion rather than useful context.

Despite these exceptions, the PatchInstruct+Neighs strategy still achieves the best overall performance in many cases aswell. But these results also highlight the importance of carefully selecting relevant neighbors to avoid negative transfer and ensure consistent forecasting improvements.

These results underline the strength of combining temporal structuring (via patches) with spatial context (via neighbors), enabling the model to learn more holistic representations and deliver significantly more accurate forecasts than our baseline.

Dataset	Horizon	PatchInstruct		Neighs		PatchInstruct+Neighs	
		MSE	MAE	MSE	MAE	MSE	MAE
Weather	1	**0.0014**	**0.029**	0.0024	0.042	0.0032	0.046
	2	0.0076	0.067	**0.0039**	**0.051**	0.0056	0.056
	3	0.0110	0.085	**0.0083**	**0.065**	0.0138	0.087
	4	0.0236	0.113	0.0172	0.091	**0.0114**	**0.075**
	5	0.0159	0.094	**0.0105**	**0.077**	0.0124	0.088
	6	**0.0101**	**0.083**	0.0116	0.084	0.0338	0.108
	12	0.0436	**0.137**	**0.0371**	0.144	0.0393	0.141
Traffic	1	**20.05**	2.76	35.15	3.05	22.09	**2.72**
	2	**9.38**	**1.89**	18.85	2.50	15.40	2.19
	3	**6.47**	**1.78**	26.67	2.74	13.61	2.10
	4	**11.15**	**1.89**	21.41	2.57	13.22	2.15
	5	**8.46**	**1.88**	43.50	3.39	36.43	3.25
	6	25.59	2.75	40.42	3.35	**14.94**	**2.13**
	12	**235.75**	**5.89**	285.82	7.71	269.68	6.88

Table 4: Forecasting Comparison: PatchInstruct vs Neighs vs PatchInstruct+Neighs.

5 Analysis

This analysis compares the performance of S^{2}IP-LLM (baseline) against our approach across different datasets and forecast horizons. The comparison focuses on error metrics (MSE and MAE) where lower values indicate better performance.

5.1 Performance Overview Across Models

The S^{2}IP-LLM baseline consistently shows higher error rates compared to our methods across datasets. Our methods shows remarkable improvements, with percentage reductions in MSE ranging from approximately 13% to 85% depending on the

dataset and method. For the Weather dataset, all our methods achieve over 80% MSE improvement. Our method achieve orders-of-magnitude faster runtimes than S^2IP-LLM with modest token usage growth, making them highly efficient for inference, especially when balanced with patch or neighbor-based prompts.

5.2 Method-Specific Performance

Our approach exhibits distinct strengths across datasets and forecasting horizons. The PatchInstruct framework demonstrates the most balanced performance, delivering substantial improvements on both the Traffic and Weather datasets—achieving up to 83% and 85% improvement over the baseline, respectively. The Neighs variant, which augments prompts with the closest neighboring time series, performs particularly well on the Weather dataset. Meanwhile, the combined PatchInstruct+Neighs strategy outperforms other methods on the Traffic dataset at longer horizons, highlighting the benefit of incorporating both local structure and external context in more challenging settings. These results suggest that method selection can be guided by the characteristics of the dataset and the specific forecasting task, with PatchInstruct offering a robust default across most conditions.

5.3 Dataset-Specific Analysis

For Weather forecasting, all methods substantially outperform the baseline. Overall MSE values are reduced from 0.009–0.0904 (baseline) to as low as 0.0014–0.043 (our methods).

PatchInstruct deliver substantial improvements, reducing MSE from 21–68 (baseline) to 6.47–20.05. However, Neighs and PatchInstruct+Neighs perform poorly in this scenario.

Across both approaches, forecast accuracy generally decreases as the horizon increases, but this pattern varies by dataset and method. For the Weather dataset, the performance degradation with longer horizons is less pronounced, especially for PatchInstruct+Neighs in multivariate settings.

5.4 Key Insights and Implications

The optimal forecasting method varies notably depending on the characteristics of the dataset and the forecasting horizon. For the Weather dataset, the PatchInstruct and Neighs strategy yields the most accurate results, effectively capturing the contextual signals from related series. In contrast, for the Traffic dataset, our main approach PatchInstruct performs best, suggesting that more complex augmentation may not always be beneficial in settings with high variability or less correlated neighbors.

6 Conclusion

The analysis demonstrates that prompt-based methods generally outperform the S^2IP-LLM baseline across most forecasting scenarios, especially at shorter horizons. The optimal method depends significantly on the specific dataset and forecast horizon, with PatchInstruct dominating in the majority of cases. This suggests that while prompt-based strategies offer a lightweight and effective alternative for time series forecasting.

7 Limitations

While our method achieves competitive accuracy compared to the S^2IP-LLM baseline, several limitations warrant consideration. First, the evaluation is limited to two benchmark datasets, which, though diverse, may not fully represent the diversity of real-world time series scenarios, such as irregular sampling or high-frequency patterns. Second, the framework remains heavily contingent upon carefully engineered prompts, introducing a labor-intensive design process that risks overfitting to specific tasks or datasets without systematic adaptation strategies. Future research should prioritize expanding dataset coverage, and developing adaptive prompting mechanisms.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2024a. Tempo: Prompt-based generative pre-trained transformer for time series forecasting.

Defu Cao, Wen Ye, Yizhou Zhang, and Yan Liu. 2024b. Timedit: General-purpose diffusion transformers for time series foundation model. *arXiv preprint arXiv:2409.02322*.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. 2024. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Yuxuan Liang, Yue Wu, Sheng Wang, Xiaoyi Zhou, Wang Yang, Rong Xu, Wen Ye, Weizhi Lin, Zhiguo He, Zongyan Li, et al. 2024. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2024. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4303–4317.

John A Miller, Mohammed Aldosari, Farah Saeed, Nasid Habib Barna, Subas Rana, I Budak Arpinar, and Ninghao Liu. 2024. A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024. S^2ip-llm: Semantic space informed prompt learning with llm for time series forecasting.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. 2023. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*.

Yilong Ren, Yue Chen, Shuai Liu, Boyue Wang, Haiyang Yu, and Zhiyong Cui. 2024. Tpllm: A traffic prediction framework based on pretrained large language models.

Xiaofeng Shao, Soumya Ghosh, and Suhasini Subba Rao. 2020. Time-series analysis and its applications in scientific disciplines. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2174):20200209.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*.

Hao Xue and Flora D. Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting.

G Zerveas, S Jayaraman, D Patel, A Bhamidipaty, and C Eickhoff. 2020. A transformer-based framework for multivariate time series representation learning. arxiv. *arXiv preprint arXiv:2010.02803*.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355.

Xinyu Zhou, Zhengyuan Ding, Shuo Ren, Yutao Chen, Xinhui Huang, Jianhao Shi, and Wayne Xin Zhao. 2024. Ditto: A survey on fine-grained alignments of large language models. *arXiv preprint arXiv:2411.05793*.

A Summary of Forecasting Results Across different datasets

Table 5 represents evaluating five prompting strategies—Basic, Non-Overlapping, STR Decompose, Reverse Patches, and Meta Patches—across the Weather and Traffic datasets, for horizons of 1, 3, and 6. While STR Decompose occasionally shows the lowest error for short-term predictions, Reverse Patch Instruct consistently delivers strong performance across all horizons and datasets. Notably, in long-range forecasts (H=6), where prediction becomes more challenging, Reverse Patch Instruct achieves the lowest or near-lowest MAE and MSE in both datasets, highlighting its stability and generalizability. Although it incurs a slightly higher inference time than the most lightweight methods, the trade-off is minimal when weighed against the accuracy benefits. Overall, the results suggest that Reverse Patch Instruct is the most effective and reliable strategy, outperforming other variants in terms of both robustness and predictive accuracy.

A.1 Basic PatchInstruct

The Basic PatchInstruct method employs overlapping sliding windows (size=3, stride=1) to capture local temporal patterns, followed by strategic sequence reversal to prioritize recent context. Unlike conventional approaches that process time series chronologically, this method reverses the generated patches such that the most recent window $[x_{t-2}, x_{t-1}, x_t]$ appears first in the token sequence. This architectural innovation forces the model to attend to immediate temporal patterns before historical context, combining the local sensitivity of patch-based methods (Nie et al., 2022) with explicit recency prioritization. The approach demonstrates particular efficacy in high-frequency electricity demand forecasting where near-term consumption patterns strongly influence subsequent values.

A.2 Non-Overlapping PatchInstruct

This variant utilizes non-overlapping windows where both window size and stride equal the prediction horizon (typically 3). The method partitions the series into discrete blocks like `[8.35, 8.36, 8.32]` followed by `[8.45, 8.35, 8.25]`, eliminating redundant data coverage while maintaining temporal progression. The design trades off some contextual granularity for computational efficiency, making it suitable for scenarios with pronounced periodic patterns. By processing patches in natural order without sequence reversal, the method preserves strict temporal causality, particularly effective when historical seasonal trends dominate the forecasting signal.

A.3 STR Decompose PatchInstruct

Integrating seasonal-trend-residual decomposition, this method first separates raw values into trend (trend_t) and residual (residual_t = series_t - trend_t) components. Each time step becomes a composite token $[T_t, R_t]$, enabling joint modeling of long-term trajectories and short-term fluctuations. These dual-aspect tokens are organized into overlapping windows:

$$[[T_1,R_1], [T_2,R_2], [T_3,R_3]]$$

preserving both local context and decomposition characteristics. The architecture explicitly captures multi-scale temporal dynamics, particularly beneficial for electricity demand series containing both gradual load changes and sudden consumption spikes.

A.4 Reverse Ordered Patches

Building on basic patch inversion, this method systematically prioritizes recent context through full sequence reversal of overlapping windows. The architectural innovation forces models to process the final patch $[x_{94}, x_{95}, x_{96}]$ first, implementing a "recency-first" attention mechanism. This structural bias proves particularly effective for 10-minute interval forecasting where immediate consumption patterns (last 30 minutes) contain stronger signals than older data. The approach maintains patch-based efficiency while adding temporal prioritization through simple sequence manipulation.

A.5 Meta Tokens Patches

This advanced variant enriches temporal representation through explicit time slot encoding. Each value v_t pairs with its absolute position in the daily cycle (0-143 slots) as $(v_t; \text{slot}_{\text{id}})$, creating hybrid tokens like `(8.35;63)`. These meta-tokens are windowed into overlapping patches:

```
[(v1;slot1), (v2;slot2), (v3;slot3)]
[(v2;slot2), (v3;slot3), (v4;slot4)]
...
[(v94;slot94),(v95;slot95),(v96;slot96)]
```

enabling joint learning of consumption patterns and their absolute temporal positions. The fixed

Table 5: Ablation Study: Comparing Variants of Patch-Based Prompting Strategies Across Datasets.

Dataset	Horizon	Basic			Non-Overlapping			STR Decompose			Reverse Patches			Meta Patches		
		MAE	**MSE**	**Time**	**MAE**	**MSE**	**Time**	**MAE**	**MSE**	**Time**	**MAE**	**MSE**	**Time**	**MAE**	**MSE**	**Time**
WEATHER	1	0.014	0.0003	0.66	0.012	0.0002	1.6151	**0.009**	**0.0001**	1.2553	0.015	0.0005	1.2290	0.020	0.0007	0.7471
	3	0.050	0.0045	0.989	0.055	0.0064	1.3580	**0.045**	**0.0030**	1.0737	0.053	0.0045	0.9732	0.067	0.0073	0.9269
	6	0.078	0.0116	1.146	0.063	0.0079	3.9016	0.100	0.0230	1.1523	**0.056**	**0.0070**	1.5120	0.060	0.0074	0.9760
TRAFFIC	1	1.43	6.27	0.699	1.35	5.60	1.3404	1.34	4.17	1.2677	**1.15**	**3.69**	1.1221	1.32	5.35	1.3846
	3	1.07	3.36	0.940	1.47	7.17	1.1910	0.99	2.53	1.1801	**0.89**	**1.83**	1.2018	0.94	2.38	1.1584
	6	1.14	4.37	0.910	1.14	3.60	1.3010	1.79	8.71	1.5667	**0.89**	**2.26**	1.1137	1.03	2.95	1.1650

slot indices provide crucial circadian context, helping disambiguate similar patterns occurring at different times (e.g., morning vs. evening peaks). This method adapts positional encoding strategies from language models to time series, grounding predictions in both value sequences and absolute time references.

Basic PatchInstruct

You are a forecasting assistant that sees time series data. The sequence represents the total regional humidity measured every 10 minutes.
Task:
(1) Split the series into overlapping patches with window size 3 and stride 1.
(2) Generate the patches in natural order, then reverse the list so the most recent patch appears first.
(3) Use these patch tokens to forecast the next 3 values.

Output format:

```
Patches:
[[latest_patch], ..., [oldest_patch]]

Prediction:
[y1, y2, y3]
```

No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

Non-Overlapping PatchInstruct

Tokenize the given time-series data into non-overlapping patches where a patch is a contiguous subsequence of the time-series. Ensure to use a fixed window size equal to the Horizon size (e.g., 3) and the stride is equal to the window size. This means that each patch starts exactly where the previous one ends and there will be no overlap. Output patches as a list, in order, using square brackets. Each patch becomes a token used to represent local temporal patterns. Use the sequence of patches to predict the next value(s). Below are a few shot examples of non-overlapping patching: Time series data: 8.35, 8.36, 8.32, 8.45, 8.35, 8.25, 8.20, 8.09, 8.13, 8.00, 7.94, 7.86

Patches generated based on Horizon (3), stride = 3:
[8.35, 8.36, 8.32]
[8.45, 8.35, 8.25]
[8.20, 8.09, 8.13]
[8.00, 7.94, 7.86]

Prediction: [7.89, 7.97, 7.94]

STR Decompose PatchInstruct

You are a forecasting assistant that receives STL-decomposed tokens.
Input:
- "series": 96 raw numbers (Humidity demand)
- "horizon" : 3 (fixed)
Task:
1. Decompose the series into
$trend_t$ and $residual_t = series_t - trend_t$
2. For each time-step create a pair token: ($trend_t$, $residual_t$).
3. Split the 96 composite tokens into overlapping patches (window = 3, stride = 1).
4. Use those patches to forecast the next 3 raw values.

Output exactly
`[[T1,R1], [T2,R2], [T3,R3]]`
`[[T2,R2], [T3,R3], [T4,R4]]`
`. . .`
`[[T94,R94], [T95,R95], [T96,R96]]`

`Prediction:`
`[y1, y2, y3]`
No headings or extra words. Decimals $\leq$ 4 places; keep leading zeros (e.g., `0.8032`).

Reverse Ordered Patches PatchInstruct

You are a forecasting assistant that sees time series data. The sequence represents the total regional humidity measured every 10 minutes.
Input:
- "series": 96 raw numbers (Humidity, 10-min cadence) - "horizon" : 3 (fixed)
Task:
1. Split the series into overlapping patches (window = 3, stride = 1).
2. Generate them in natural order, then reverse the list so the most recent patch appears first.
3. Use those patch tokens to forecast the next 3 normalised values.

Output format:

```
Patches:
[[latest_patch],
... ,
[oldest_patch]]

Prediction:
[y1, y2, y3]
```

No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

Meta tokens Patches PatchInstruct

You are a forecasting assistant that sees time series data, where each datapoint is paired with its 10-minute slot index within the day. The sequence represents the total regional humidity measured every 10 minutes.
Input:
- "series": 96 raw numbers (Humidity, 10-min cadence) - "horizon" : 3 (fixed)
Time-slot index - A day is divided into 144 slots (0 $\rightarrow$ 143). - slot = floor((60*HH + MM)/10). Example: 10:30 $\rightarrow$ 63 (because 10*60 + 30 = 630; 630/10 = 63).
Token format (value ; $slot_{id}$)
$slot_{id}$ corresponds to the measurement's clock time
Task:
1. Convert the 96-point series into 96 two-element tokens as above.
2. Split the token stream into overlapping patches (window = 3, stride = 1).
3. Use those patches to forecast the next 3 raw demand values.

Output format:

```
[(v1;slot1), (v2;slot2), (v3;slot3)]
[(v2;slot2), (v3;slot3), (v4;slot4)]
. . .
[(v94;slot94),          (v95;slot95),
(v96;slot96)]

Prediction:
[y1, y2, y3]
```

No headings or extra words. Decimals ≤ 4 places; keep leading zeros (e.g., `0.8032`).

Keeping LLMs from Being Distracted: Grade-Aware Kanji Reading Estimation Fully Executable in Web Browsers for Japanese Education

Yo Ehara
Tokyo Gakugei University / 4-1-1 Nukuikita-machi, Koganei-shi, Tokyo, Japan.
ehara@u-gakugei.ac.jp

Abstract

Automatic reading (ruby or furigana) generation of kanji is a fundamental task in Japanese natural language processing (NLP), particularly for educational purposes. Several web services already provide grade-aware annotations, assigning readings only to characters not included in each elementary school grade list; however, these services usually rely on a server. They upload pupil texts, process remotely, and return with annotations. Such a design has practical problems: server overload caused by simultaneous access to many tablets and security risks due to transmission of text that may contain personal information. We introduce a browser-based system that places the reading estimation model and grade filter inside the JavaScript engine of each client. After a single-page load, all processing occurs locally; therefore, no further network communication is required. In experiments, the proposed system achieved an accuracy comparable to existing server-side approaches while removing the server load entirely. Large language models such as GPT-5, are becoming increasingly popular in education, yet they still make reading estimation errors, even on short examples. When GPT-5 is asked to simultaneously solve elementary math problems and assign kanji readings, its math accuracy degrades significantly, indicating that LLMs are heavily distracted. These results suggest that our method not only offers a practical solution for latency and privacy issues in obtaining high-quality results but also prevents LLMs from being distracted.

1 Introduction

Automatic reading (ruby or furigana) estimation of kanji characters is a fundamental Japanese NLP task with educational applications. Several web services now provide grade-sensitive ruby, adding furigana only to characters that exceed the set of kanji that is officially taught in each grade of an elementary school. However, most of these services transmit user text to a server where readings are inferred and returned. This architecture raises three practical concerns: (i) server bottlenecks when many pupils simultaneously access the system on tablets, (ii) information security risks for texts containing personal data, and (iii) reduced availability under unstable network conditions, such as during natural disasters.

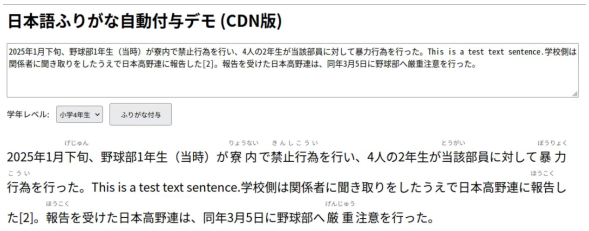


Figure 1: Our System. Once the page that uses JavaScript is loaded, the Japanese reading estimation of the input text is conducted fully on browsers without communicating with servers. When a user specifies the reading of a kanji character, the text with the reading attached appears below the text box. In this example, readings are attached to kanji characters for 4th graders and above. The translation of this example is available in later sections.

We propose a client-side system that performs both reading estimation and grade-sensitive ruby annotation entirely within the browser's JavaScript engine, eliminating the need to upload the input text. After the initial page load, all processing occurs locally, thereby ensuring privacy, resilience, and scalability (Figure 1).

In Japan, the individual kanji characters taught in each elementary school grade are specified by the government. The process of attaching readings to kanji characters at or above the specified grade level is an important process in Japanese education; however, because it is deterministic, the large language model (LLM) often makes mistakes or becomes distracted. Our evaluation measures both

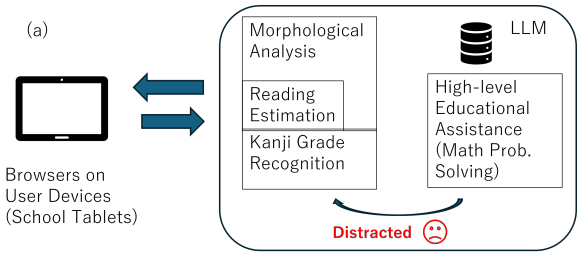


Figure 2: A previous approach combined with LLM educational assistance. LLMs are distracted in assigning grade-aware readings to kanji, which degrades the quality of educational assistance.

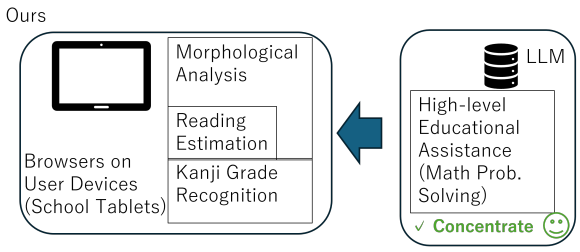


Figure 3: Our approach combined with LLM educational assistance. LLMs concentrate on educational assistance because grade-aware kanji reading assignment is conducted on browsers of the user device (school tablets).

the reading-estimation accuracy and precision of the grade-level filter that determines which of the kanji receive ruby. Because the target characters are predetermined by official gradewise kanji lists, near-perfect accuracy can be achieved in principle. LLMs such as ChatGPT, however, sometimes hallucinate in grade assignments, leading to lower accuracy. Therefore, our experiments highlight the complementarity of traditional rule-based or statistical NLP running in the browser with LLM-based approaches.

This study is the first to create a grade-specific reading estimator for Japanese that works solely on a browser. Additionally, it demonstrates that technical collaboration is necessary when utilizing LLMs for educational purposes, such as entrusting deterministic processing of LLMs to the browser side, in addition to improving LLM performance.

The proposed method is expected to be effective, especially in educational support using LLM. In a recent approach, the LLM performs all the complex tasks (Figure 2). However, with this method, the LLM is distracted by grade-aware kanji reading estimation, which reduces the quality of the essential educational support.

In contrast, in our approach, the task of grade-aware kanji reading estimation, a task that traditional natural language processing is good at, does not require GPUs, but only a browser running on a user device such as a school tablet (Figure 3). The recent tablet browsers are sufficiently powerful to perform morphological analysis, reading estimation, and kanji-level recognition for the amount of text that a user is supposed to read. This allows the LLM to focus on more essential educational support. In this study, we experimentally demonstrated the effectiveness of our approach by considering a specific educational support setting.

The resources such as the demo system can be found at `https://rebrand.ly/okuranukana`.

Translation of Figure 1 The example was taken from a recent Wikipedia article, hence has no copyright issues. It is translated as follows: “In late January 2025, a first-year baseball team member (at the time) engaged in prohibited behavior within the dormitory, and four second-year students committed acts of violence against the said team member. The school conducted interviews with the parties involved and reported the incident to the Japan High School Baseball Federation. Upon receiving the report, the Japan High School Baseball Federation issued a severe reprimand to the baseball team on March 5 of the same year.”

We selected this news because it is domestic to Japan and, as it was a hot topic at the time of submission in Japanese education, elementary school students should be able to read this news.

2 Related Work

Japanese reading estimation underpins tasks, such as text-to-speech synthesis, educational support, and accessibility enhancement (Yasuda et al., 2018; Nagata, 1999; Hoshino and Morise, 2021). Early systems relied on server-side morphological analyzers, such as MeCab and JUMAN++ (Kudo et al., 2004; Kawahara and Kurohashi, 2013), the high accuracy of which come at the cost of network latency and privacy risks when deployed in web applications. Client-side alternatives have emerged with kuromoji.js (Asano, 2014), one of the few analyzers implemented entirely in JavaScript. However, its reading-estimation accuracy has not been rigorously benchmarked against standard corpora.

Methodologically, reading estimation has progressed from statistical to neural approaches. Conditional random fields achieved strong performance when integrated into morphological analyz-

ers (Kudo et al., 2004), and subsequent studies extended segmentation and furigana generation to educational texts (Nagata, 1999). Neural models now exploit contextual embeddings to predict readings with improved prosody for text-to-speech (TTS) (Yasuda et al., 2018) and joint grapheme–phoneme conversion (Hoshino and Morise, 2021). Despite these advances, most neural systems assume a server-side execution, leaving a gap for fully in-browser solutions.

Evaluation typically follows text-input research, consisting of character error rate, first-candidate accuracy, and word error rate, to quantify phonetic conversion errors unified under the framework of Soukoreff and MacKenzie (Soukoreff and MacKenzie, 2003; Morris et al., 2004; Wang and Woodland, 2003). Recent progress in WebAssembly (WASM) and libraries such as TensorFlowjs has reduced the performance gap between browser and native runtimes (Team, 2018; Gardner et al., 2018), thereby enabling complex natural language processing (NLP) models to run locally. The MeCab-WASM and similar ports illustrate this trend; however, a systematic assessment of their reading-estimation quality remains limited. To address this deficiency, the present study offers the first comprehensive benchmark of kuromoji.js on Kyoto University Web Document Leads Corpus (KWDLC) and Advanced Japanese IME Evaluation (AJIMEE)-Benchmark, comparing it with MeCab-WASM, thereby clarifying the extent to which browser-based Japanese NLP has matured.

This study is also connected to research on the interplay between vocabulary and readability. Although not focused on Japanese language education, detailed investigations into the relationship between vocabulary knowledge and readability in English as a second language can be found in (Ehara, 2023, 2022, 2021, 2019, 2018; Ehara et al., 2012).

3 Proposed System

Figure 1 shows our system. The proposal system comprises two components. The first is reading estimation. Japanese, like Chinese, is a language that does not divide words into syllables, requiring that word boundaries be automatically determined. Only Chinese characters require reading estimation, as hiragana and katakana are phonetic characters and do not require reading estimation. In Japanese processing, the system divides Japanese text into words and estimates their readings and parts of speech, conventionally called morphological analysis.

The proposed system requires a morphological analyzer that runs in a browser. Two methods have recently been developed for this purpose. One uses a morphological analyzer written entirely in JavaScript. The other uses Web Assembly, a mechanism that allows virtual machine code to run in a browser, and outputs a morphological analyzer written in C or another language as a Web Assembly binary. We used kuromoji.js to perform a morphological analysis of Japanese text, as it is designed based on JavaScript and has a wealth of usage examples. This is the JavaScript version of Kuromoji, a Java-based open-source Japanese morphological analyzer [1]. We used kuromoji.js version 0.1.2 in combination with the lexicon in the dictionary developed by the Information-Technology Promotion Agency of Japan (IPADIC) (Lab, 2007). We ran the engine under Node.js v18.17.0 and verified identical behavior in chromium-based browsers, ensuring that our results can be generalized to typical client-side deployment.

`kuromoji.js` tokenize Japanese text and estimate readings on the fly. Next, we need an educational kanji database that enumerates characters officially taught in Grades 1–6. For Japanese elementary schools, a clear list of kanji characters taught by grade level is provided by the Japanese government. This list consists of 1,026 characters. These characters are directly loaded into a dictionary-type variable in JavaScript. The surface form and reading of each morpheme are then cross-referenced with the specified grade level in the kanji database; characters that exceed the learner's scope are flagged. Finally, the system injects `<ruby>` elements for the flagged items and streams the fully annotated text back to the DOM, enabling seamless display without network latency or privacy concerns. All user interaction and visualisation are handled by a minimalist HTML/CSS/JavaScript frontend, ensuring platform independence and eliminating the need for server-side processing.

3.1 Implementation Example

The following is an example of furigana annotation for 2nd-grade elementary students:

`Input:` 私は毎日学校に通っています。
`Output:` 私[わ た し]は 毎 日[ま い に ち]学校[がっこう]に通[かよ]っています。

[1] https://www.atilika.com/en/kuromoji/

Here, furigana is not added to kanji learned in 1st grade such as “学,” “校,” and “日.”

4 Evaluation Datasets and Methods

4.1 Evaluation Datasets

We conducted experiments using three corpora that differed in register, genre, and target applications. **KWDLC** (Kyoto University Web Document Leads Corpus; (University, 2011)) supplies 150 web sentences comprising 2,383 morphemes and 6,120 characters, each annotated with morpheme boundaries, part-of-speech tags, and ruby information. **AJIMEE-Bench** (azooKey Project, 2023), built on the Japanese Wikipedia Input Error Dataset v2 (Tanaka et al., 2021), contains 200 kana–kanji pairs, half presented with preceding context and half without, distributed under the CC-BY-SA 3.0 license. Finally, the **Educational Kanji Dataset** was constructed based on the Ministry of Education’ s official grade-level kanji list (Ministry of Education, Culture, Sports, Science and Technology (MEXT), 2017), which covers 1,026 characters across Grades 1–6 (80, 160, 200, 202, 193, and 191 characters, respectively) and serves as the gold standard for grade-aware furigana annotation.

These corpora jointly enable a balanced assessment of general reading estimation accuracy, robustness to input errors, and pedagogical suitability for primary education.

As stated in Section 3, the proposed system is based on kuromoji.js.

MeCab-WASM provides a WebAssembly port for MeCab and functions as a comparative system. Compiled from MeCab 0.996 and linked to the same IPADIC dictionary, this module is executed entirely within WebAssembly-compliant browsers. The shared lexicon and browser-native runtime isolate the differences for the tokenizer, allowing a fair comparison of speed and accuracy with JavaScript-native `kuromoji.js`.

4.2 Evaluation Metrics

Considering the characteristics of reading estimation and kana-kanji conversion, we employed the following metrics:

4.2.1 Character-level Evaluation

$$\text{CER} = \frac{\text{Levenshtein Distance}(r, h)}{|r|} \tag{1}$$

where r is reference reading; h is hypothesis reading; and $|r|$ is reference reading length. Then, Character Accuracy is defined as $(\text{Correct Characters}/\text{Total Characters}) \times 100\%$.

4.2.2 Morpheme-level Evaluation

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP is true positive (correct); FP is false positive (over-detection); and FN is false negative (underdetection).

4.2.3 Perfect Match Evaluation

$$\text{Accuracy@1} = \frac{\text{Perfect Match Count}}{\text{Total Items}} \times 100\% \tag{5}$$

5 Experimental Results

5.1 KWDLC Evaluation Results

We conducted a detailed analysis of 150 sentences containing 2,383 morphemes using KWDLC.

5.1.1 Overall Performance Metrics

Table 1 presents the overall results of the KWDLC evaluation.

5.1.2 Performance Analysis by Sentence Length

Table 2 shows performance according to sentence length category.

Performance degradation is observed with increasing sentence length. This is primarily attributed to increased vocabulary diversity in longer sentences, which makes morpheme boundary identification more difficult.

5.1.3 Error Factor Analysis

Table 3 shows the breakdown of error factors.

The largest error factor was the morpheme boundary mismatch (51.3%), whereas pure reading estimation errors were only 1.3%. This suggests that the reading-estimation function of kuromoji.js is highly accurate.

5.2 AJIMEE-Bench Evaluation Results

The AJIMEE-Bench evaluation by Kuromoji.js demonstrated excellent performance.

Metric	Value	Description
Character Reading Accuracy	**83.17%**	5,090 chars/6,120 chars
Morpheme Perfect Match	**47.38%**	1,129 morphemes/2,383 morphemes
Morpheme Boundary Accuracy	48.72%	(1,129+32)/2,383
Precision	45.16%	1,129/2,500
Recall	47.38%	1,129/2,383
F1 Score	**46.24%**	Harmonic mean

Table 1: Overall Performance Metrics in KWDLC Evaluation

Category	Sentences	Precision	Recall	F1 Score
Medium (10-20 chars)	18	59.3%	64.5%	61.8%
Long (20-30 chars)	59	57.2%	59.3%	58.3%
Very Long (30+ chars)	73	42.0%	43.2%	42.6%

Table 2: Performance by Sentence Length Category (KWDLC Evaluation)

5.2.1 Overall Performance Metrics

Table 4 presents the results of the AJIMEE-Bench evaluation.

5.2.2 CER Distribution Analysis

Table 5 provides detailed CER distributions.

Notably, 89% of the items achieved a CER below 10%, without large errors exceeding 30%.

5.2.3 Context Dependency Analysis

Table 6 shows the performance differences for with and without context.

Interestingly, a higher accuracy was achieved without context. This indicates that kuromoji.js employs dictionary-based methods without considering context, which leads to errors in items requiring context-dependent reading disambiguation.

5.3 Inter-system Comparison

Table 7 presents a performance comparison between the evaluation systems.

As for "*", the value was calculated from the KWDLC evaluation character accuracy of 83.17% (100-83.17=16.83%).

5.4 Performance Evaluation

Table 8 shows the evaluation results of the grade-level furigana annotation system.

In higher grades, appropriate annotation rates improved. This is attributed to the clarification of furigana annotation targets with an increase in the number of kanji learned.

6 Discussion

Reading estimation and kana–kanji conversion require fundamentally different evaluation strategies that diverge in direction, focus, and contextual dependence. Kana–kanji conversion is assessed for the ability to transform purely phonetic input into kanji-kana mixed text, which demands high precision in selecting the correct homophonic alternative among many candidates. In contrast, reading estimation proceeds in the opposite direction, converting mixed text into a phonetic script, and therefore hinges on resolving homographs rather than homophones. Because each kanji character typically maps to a limited set of readings, dictionary information alone can often disambiguate the output, whereas kana-kanji conversion must rely heavily on broader sentential context (e.g. distinguishing 「公園」 *park* from 「講演」 *lecture*). These contrasting requirements necessitate distinct metrics and test suites, underscoring the fact that the performance of an algorithm cannot be extrapolated from one task to another without careful validation.

A methodological mismatch implies that a single algorithm can yield discordant results when evaluated under two paradigms. Empirical evidence from our experiments shows that systems optimized for kana-kanji conversion do not automatically excel in reading estimation and vice versa. Consequently, comparative studies must explicitly frame their evaluation protocols or risk drawing misleading conclusions about model capabilities across tasks that although superficially related, embody different linguistic challenges.

The kuromoji.js node offers a practical browser-

Error Factor	Proportion	Description
Morpheme Boundary Mismatch	51.3%	Different morpheme segmentation
Reading Estimation Error	1.3%	Correct boundary, wrong reading
Perfect Match	47.4%	Both boundary and reading correct

Table 3: Breakdown of Error Factors (KWDLC Evaluation)

Evaluation Metric	Value	Remarks
Average CER	**2.14%**	Extremely low error rate
Accuracy@1	**81.50%**	High perfect match rate
Perfect Match Items	163/200	81.5% perfect reading estimation

Table 4: Overall Performance Metrics in AJIMEE-Bench Evaluation

CER Range	Items	Proportion	Cumulative
0% (Perfect match)	163	81.5%	81.5%
0-10%	15	7.5%	89.0%
10-20%	15	7.5%	96.5%
20-30%	7	3.5%	100.0%
30%+	0	0.0%	100.0%

Table 5: Detailed CER Distribution (AJIMEE-Bench Evaluation)

Category	Items	Average CER	Difference
Without Context	100	1.44%	Baseline
With Context	100	2.84%	+1.40%

Table 6: Context Dependency Analysis (AJIMEE-Bench Evaluation)

native solution that balances accuracy, robustness, and deployment simplicity. On the AJIMEE-Bench dataset, the engine achieved a character error rate of 2.14% and an accuracy @1 of 81.50% while maintaining stable performance without catastrophic errors. Its JavaScript implementation eliminates external API dependencies, enabling high-speed client-side processing that preserves user privacy and ensures availability, even under constrained network conditions. These strengths position `kuromoji.js` as an attractive baseline for educational applications that require lightweight yet reliable reading estimation.

6.1 Limitations Analysis

1. **Lack of context consideration**: Degraded accuracy in context-dependent reading disambiguation

2. **Morpheme boundary differences**: Evaluation difficulties across different corpora

3. **New words and proper nouns**: Challenges in processing words not in dictionary

6.2 Educational Application Potential

Implementation and evaluation of the grade-level furigana annotation system yielded the following insights:

1. **Practical accuracy**: Achieved over 90% appropriate annotation rate

2. **Scalability**: Light operation in browser environment

3. **Customizability**: Flexible configuration based on Course of Study guidelines

These results indicate that kuromoji.js can serve as a practical solution in educational support.

7 LLM Evaluation

7.1 Accuracy of LLM-based reading estimation

We conducted an analysis on how well LLMs can estimate readings by considering Japanese elementary grades. In this short example of Figure 1, the readings are assigned to kanji characters that fourth graders can read, which means that fifth grade and above, can assign the readings to kanji. ChatGPT's GPT-5 assigns excessive readings to six instances. Claude 4.1 and GPT-o3 only assigned excessive reading to one instance. In cases where readings were assigned, all the models assigned accurate readings.

7.2 Measuring LLM Distraction by Multi-tasking

Some may wonder whether it would be better to have LLM perform reading estimations simultaneously. However, studies have shown that when LLMs are tasked with multiple assignments, they become distracted and make errors (Xu et al., 2024).

System	Dataset	CER	Accuracy@1	Remarks
kuromoji.js	AJIMEE-Bench	2.14%	81.50%	This evaluation
kuromoji.js	KWDLC character-level	16.83%*	-	*Estimated value
MeCab-WASM	AJIMEE-Bench	-	-	Under evaluation

Table 7: Inter-system Performance Comparison

Grade Level	Test Sentences	Appropriate Rate	Over-annotation	Under-annotation
Grade 1	50	92.3%	4.1%	3.6%
Grade 3	50	94.7%	2.8%	2.5%
Grade 6	50	96.2%	1.9%	1.9%

Table 8: Grade-level Furigana Annotation System Evaluation Results

Metric	**Count**	**Percentage**
Total Ground Truth Grade-aware Reading Assignments	106	–
Total System Assigned	50	–
Exact Matches	8	–
Mismatches	2	–
Missing Annotations	69	–
Extra Annotations	40	–
Recall	–	7.55%
Precision	–	16.00%
F1 Score	–	10.26%

Table 9: GPT-5 Grade-aware Reading Assignment Predictive Performance

Therefore, if the LLM is tasked with solving problems while performing reading estimation for elementary school students, errors are likely to occur.

To measure the extent to which the LLM is distracted, we conducted the following experiment: The experiment used elementary mathematics from massive multitask language understanding (MMLU) (Hendrycks et al., 2020), which is widely used to evaluate LLM performance. These are simple problems at the elementary school level. MMLU is an English dataset; however, a Japanese version called JMMLU [2] has been published, which includes native Japanese speakers who have confirmed that the problems are valid in Japanese.

Using JMMLU, we had the LLM solve Japanese-translated elementary mathematics problems while estimating the problem text. In our experiments, we used GPT-5, which was released on August 7, 2025. Because GPT-5 automatically switches models during operation, we selected the GPT-5 automatic mode and input the problem statements. The experiment was conducted using an account subscribed to ChatGPT Pro. First, we selected questions 1 through 30 from the elementary-mathematics section of MMLU and provided the following instructions in Japanese: "For the following question, please output the problem statement with furigana for all kanji characters that are at or above the fourth grade level, and then select the correct answer from the four options provided and write your answer immediately after the problem statement. The options are labeled A, B, C, and D for the first, second, third, and fourth options, respectively. “ This was in Japanese, and the results were collected. This is referred to as the ” w/distraction case."

Next, the chat from the ChatGPT account was deleted to prevent it from being recorded. Then, the first instruction in the problem statement was replaced with “please output the problem statement as-is,” and the problems were solved again. The results were then collected. This is referred to as the “w/o distraction case.” In the w/o distraction case, only one out of 30 questions was incorrect, whereas in the w/ distraction case, seven sentences were incorrect. Even with simple elementary school math problems, there is significant distraction, making it impractical for the LLM to perform reading and inference simultaneously. This demonstrates the effectiveness of the proposed method, which performs reading inferences on the browser.

We also measured the reading-assignment accuracy of GPT-5 in Table 9. The results indicated significant challenges in the system's ruby annotation capability, with a recall of only 7.55%, suggesting that most of the required ruby annotations were not generated. A precision of 16.00% indicated that even among the annotations produced, accuracy was limited.

[2] https://github.com/nlp-waseda/JMMLU

7.3 Local LLM Results

We experimented to see what the results would be using a smaller local LLM model. We conducted the experiment using **gpt-oss-20b** [3], which was released by OpenAI on August 5, 2025. The experiment was conducted using the same procedure as that for ChatGPT's GPT-5. The results showed that while the model made 76 reading estimations, only five were correct, resulting in an accuracy rate of 4.72%.

Furthermore, the accuracy rate for elementary school math problems was low, with only 14 out of 30 questions answered correctly, four unanswered, and an overall accuracy of 46%. However, when the reading estimation task was not assigned, there were no unanswered questions, and 18 of the 30 questions were answered correctly, with an accuracy rate of 60%.

These results indicate that the issue of distraction caused by reading estimation remains severe even with a local model, making it infeasible to have LLM perform reading estimation simultaneously while having elementary school students read the text.

Our evaluation demonstrated that the GPT-5 distracted model exhibits significant limitations in Japanese ruby annotation tasks, with an F1 score of only 10.26%. The primary failure mode was low recall (7.55%), indicating that the model failed to identify most of the kanji requiring annotation. These results underscore the continued importance of specialized NLP tools for structured linguistic tasks and suggest that current general-purpose language models are not suitable replacements for dedicated ruby annotation systems in education or accessibility applications.

8 Conclusion

This study presents a comprehensive evaluation of Japanese reading estimation systems that run entirely in JavaScript, with particular focus on the widely used `kuromoji.js`. Our experiments show that the library delivers a performance suitable for real-world deployment, achieving a character-error rate of 2.14% and *Accuracy@1* of 81.50%. In examining the evaluation protocol, we highlight crucial methodological distinctions between conventional kana and kanji conversion benchmarks and the reading estimation task, thereby providing clearer criteria for future assessments. The prototype classroom application further demonstrated that current browser-based Japanese text-processing technologies have matured to a practical stage.

Beyond technical validation, the results indicate four broad areas of impact. First, in education, accurate on-the-fly reading estimation enables automatic furigana annotation and individualized reading support tools. Second, in terms of accessibility, the offline operation of the system offers robust reading assistance to visually impaired or low-connectivity users without risking data leakage. Third, for learners of Japanese as a second language, the same technology can underpin multilingual-level adaptive learning aids. Finally, because all the processing occurs locally, this approach offers privacy benefits that are increasingly demanded by modern web applications.

This study conducted experiments using a single zero-shot prompt, assuming practical use in real educational settings. While from a scientific research perspective, techniques such as Chain-of-Thought, role-based, or sequential prompting could potentially yield higher performance even with LLMs alone, applying such prompting strategies in actual classroom environments remains difficult in practice.

For elementary school students in Japanese, adding furigana is an essential feature in NLP using LLMs, but even LLMs such as GPT-5, which have received significant investment, have been shown to make mistakes. As shown in several studies, such as (Xu et al., 2024), LLMs struggle to perform multiple tasks simultaneously. Therefore, furigana assignments should be performed in a browser, whereas LLM should focus on processing educational content as a desirable approach for future NLP applications in education. In particular, this study demonstrated that simultaneously performing reading inference with an LLM for Japanese elementary school math problems resulted in a seven-fold increase in the error rate of the LLM. As research on the use of LLMs for educational purposes has increased, it has become impractical to have LLMs perform multiple tasks simultaneously. Our results suggest the usefulness of methods such as ours that adapt individually to learners on the browser side.

8.1 Future Work

In this study, we focused on Japanese language processing; however, because it is common for Chi-

[3] https://huggingface.co/openai/gpt-oss-20b

nese characters to have multiple readings, the same mechanism can be applied to Chinese characters. Specifically, after using a Chinese word segmenter that runs on a browser, the Chinese readings of kanji characters can be estimated. In general, since there are fewer variations in Chinese character readings than in Japanese, where each character has two types of readings (an ancient Chinese-derived reading and a Japanese-specific reading), we expect that the word segmenter will operate as smoothly as in this study, even on a browser.

In Japan, the Global Innovation Gateway for All (GIGA) School Project is being implemented to provide elementary school students with tablet computers through government budget subsidies. Similar initiatives are being implemented in regions such as Taiwan and Hong Kong, and efforts to promote tablet use are ongoing in Mainland China. This study provides a useful method for efficiently obtaining effective educational support from LLMs using tablets in elementary schools.

Limitations

This paper's estimation is limited to certain language models. As language models are evolving quickly, other language models may be able to estimate the readings of Japanese texts more correctly or may be able to conduct complicated educational assistance with being less distracted by estimating the readings of Japanese texts.

We did not test calling multiple LLM tasks simultaneously for each task. While this may solve the problem of the performance drop caused by multiple task following, this does not solve the problems caused by client-server network communications.

Claude Code was used to support our coding in experiments. ChatGPT was used for automatic English proofreading.

Ethical Considerations

Since we conducted our experiments using only publicly-available datasets which allows research purpose use, we believe that our study has no ethical issues. We prioritized measuring the performance of kanji reading estimation using the proposed approach and conducting computer experiments to show that the proposed method improves the quality of LLM educational support by eliminating distractions from reading estimation. Therefore, user studies have not been conducted at this stage, and hence this paper does not involve ethical issues related to user studies.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363. We are deeply grateful to the anonymous reviewers for their constructive feedback.

References

Takuya Asano. 2014. kuromoji.js: Japanese morphological analyzer in javascript. https://github.com/takuyaa/kuromoji.js. JavaScript implementation of Japanese morphological analyzer.

azooKey Project. 2023. Ajimee-bench: Advanced japanese ime evaluation benchmark. https://github.com/azooKey/AJIMEE-Bench. GitHub repository.

Yo Ehara. 2018. Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yo Ehara. 2019. Uncertainty-aware personalized readability assessments for second language learners. In *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, pages 1909–1916. IEEE.

Yo Ehara. 2021. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 806–814.

Yo Ehara. 2022. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022, Durham, UK, July 24-27, 2022*. International Educational Data Mining Society.

Yo Ehara. 2023. Innovative software to efficiently learn english through extensive reading and personalized vocabulary acquisition. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky - 24th International Conference, AIED 2023, Tokyo, Japan, July 3-7, 2023, Proceedings*, volume 1831 of *Communications in Computer and Information Science*, pages 187–192. Springer.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Riku Hoshino and Masanori Morise. 2021. Neural reading and pronunciation prediction for japanese text-to-speech synthesis. *IEICE Transactions on Information and Systems*, E104-D(2):285–293.

Daisuke Kawahara and Sadao Kurohashi. 2013. Morphological analysis for unsegmented languages using recurrent neural network language model. *EMNLP 2013*, pages 1292–1302.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

NAIST Computational Linguistics Lab. 2007. Ipadic: Mecab standard dictionary. https://github.com/taku910/mecab/tree/master/mecab-ipadic. Japanese dictionary for morphological analysis.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2017. Curriculum guidelines "zest for life" appendix: Kanji allocation by grade. https://www.mext.go.jp/a_menu/shotou/new-cs/youryou/syo/koku/001.htm. Government of Japan White Paper.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. *Eighth International Conference on Spoken Language Processing*.

Masaaki Nagata. 1999. A new japanese word segmentation method and its application to automatic furigana generation. *Natural Language Engineering*, 5(4):389–409.

R William Soukoreff and I Scott MacKenzie. 2003. Metrics for text entry research: An evaluation of msd and kspc, and a new unified error metric. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 113–120.

Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2021. Construction of japanese wikipedia input error dataset v2. In *Proceedings of the 27th Annual Meeting of the Association for Natural Language Processing*, pages 1001–1006. The Association for Natural Language Processing.

Google Brain Team. 2018. Tensorflow.js: Machine learning for javascript developers. https://www.tensorflow.org/js. JavaScript library for training and deploying ML models in the browser.

Kyoto University. 2011. Kyoto university web document leads corpus. https://nlp.ist.i.kyoto-u.ac.jp/index.php?KWDLC. Annotated Japanese corpus with morphological and dependency information.

Liang Wang and Philip C Woodland. 2003. Word level confidence annotation using combinations of features. In *Eighth European Conference on Speech Communication and Technology*.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548, Mexico City, Mexico. Association for Computational Linguistics.

Yusuke Yasuda, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2018. Neural machine translation with pronunciation prediction for japanese text-to-speech synthesis. In *Proceedings of Interspeech 2018*, pages 3267–3271.

Controlling Emotion Intensity and Blending in Text via Task Vector Composition for Dialog System Personalization

Ryota AMANO, Kazuya MERA, Yoshiaki KUROSAWA, Toshiyuki TAKEZAWA
Graduate School of Information Sciences, Hiroshima City University, Japan
dk65030@e.hiroshima-cu.ac.jp, {mera, kurosawa, takezawa}@hiroshima-cu.ac.jp

Abstract

Recent advances in large language models (LLMs) have accelerated the development of dialog systems, with increasing attention paid to personalization. One key challenge is how to flexibly control emotional intensity and blend multiple emotions in generated text—a crucial component for simulating diverse personalities. Traditional approaches often require training separate models for each emotional configuration. In this study, we propose a method that enables fine-grained control over both emotional intensity and blended emotional states by composing emotion-specific task vectors. Each emotion-specific model is fine-tuned from a base model, and the resulting task vectors are combined and applied to a neutral model to synthesize blended emotional behaviors. Experimental results using LLM-based evaluation demonstrate that our method successfully generates text reflecting specified emotional profiles with controllable intensity and combinations.

1 Introduction

Since the release of ChatGPT in November 2022, the rapid development of LLMs has prompted increased interest in deploying dialog systems in society. Dialog systems are already being used in various contexts, including customer support, counseling, and elderly care. More recently, their use has expanded into domains such as AI characters in the metaverse or human digital twins, where dialog systems are expected to respond with distinct personalities.

While personalization typically focuses on attributes such as memory, temperament, or background, the ability to control emotional expression—particularly its intensity and blend—is a critical yet underexplored aspect. Moreover, compared to personality traits, emotions are more readily perceived and evaluated, both by humans and LLM-based automatic judges. Our study focuses on this gap, leveraging emotions as a proxy for lightweight, controllable personalization.

To endow dialog systems with personality traits such as memory, background, and temperament, methods have been proposed that input these traits as text or embedding vectors. However, using text to represent personality poses challenges in fine-grained control and often requires large volumes of text to cover detailed nuances. Alternatively, embedding-based personality representation typically relies on fine-tuning, which necessitates retraining every time a new personality is needed.

Personalized Soups (Jang et al., 2023) addressed this issue by building models aligned with specific response styles from different perspectives and then merging them to simultaneously express multiple aspects of personality. Extending this idea, it becomes possible to express composite personalities without retraining by creating and merging models that embody typical personality traits. However, building such models requires labeled personality datasets and evaluating the resulting output, both of which are difficult.

To address these limitations, the present study focuses on emotions rather than personality. There are two main reasons for this choice. First, emotional expressions in text are generally easier to recognize and evaluate—both by human judges and by automatic evaluators such as LLM-as-a-judge—compared to personality traits, which are abstract and often require long-term behavioral context. Second, large-scale corpora with explicit emotional annotations (e.g., intensity levels) are more widely available than corpora annotated with personality traits, enabling more robust training and evaluation. We propose a method to express emotion intensity and combinations by merging models that each specialize in a specific emotion.

Furthermore, we introduce an evaluation method using LLM-as-a-judge (Zheng et al., 2023) to assess the emotional expression in generated text, which is challenging to evaluate quantitatively.

2 Related work

Recent studies have shown that merging models through linear interpolation or weighted averaging of parameters can modify a model's capabilities. Notably, Ilharco et al. (2023) introduced the concept of a "task vector," derived from the difference in model parameters before and after training, and showed that adding or subtracting these vectors can alter model behavior accordingly. A task vector can be intuitively understood as a direction in parameter space that represents a specific behavioral change, such as learning a new skill or style, and this concept has since become central in model merging research.

Building on this idea, Huang et al. (2024) proposed the "Chat Vector" concept—capturing the difference between a base model and its instruction-tuned variant—and demonstrated how this enables instruction-following behavior to be transferred to language models in other languages without further training. As illustrated in Figure 1, the conventional pipeline for multilingual adaptation typically begins with continual pre-training (CP) of a pre-trained language model (PLM) on a target language corpus. This is followed by supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF), resulting in a target LM, which is instruction-tuned. In contrast, the Chat Vector approach bypasses these stages entirely: it extracts a parameter vector (referred to as the Chat Vector) from a source-language PLM and its chat-tuned variant, then grafts it onto a continually pre-trained PLM (CP Model)—much like donning a suit of conversational "armor"—to instantly impart dialog capabilities.

Jang et al. (2023) trained specialized models based on expertise, information richness, and response style, and proposed personalizing alignment by merging models according to user preferences. Their merging method uses a weighted sum of parameters under the constraint that weights sum to one. However, their work does not discuss how to determine these weights.

Zhou et al. (2024) formulate smooth attribute-intensity control for text generation and propose an automatic evaluation framework that combines

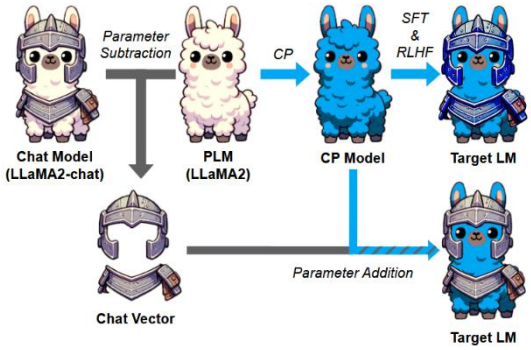


Figure 1: Illustration of the Chat Vector concept (Huang et al., 2024).

GPT-4 pairwise judgments with an Elo-rating aggregation scheme, allowing them to quantify range, calibration and consistency without human annotators.

Their benchmark spans five single attributes—anger, happiness, formality, understandability, and conciseness—covering sentiment, stylistic, and broader linguistic properties. While the authors demonstrate effective control for each attribute in isolation, they explicitly acknowledge that simultaneous manipulation of multiple attributes, though theoretically desirable, is left for future work.

Two of the evaluated attributes (anger and happiness) are clearly emotional; however, the paper's analysis treats them alongside the other attributes and does not offer an in-depth discussion of emotion-specific challenges (e.g., valence diversity or cross-emotion interference). Consequently, issues unique to fine-grained emotional control remain open questions.

These studies collectively demonstrate the feasibility of modifying model behavior via parameter arithmetic. Building upon these foundations, our study applies task vector-based model merging to emotional expression—a domain that allows clearer evaluation and offers richer annotated data resources.

3 Task Vector Composition for Fine Emotional Control and Blending

To enable flexible control over emotional expression in text generation, we propose a lightweight method based on task vector composition. Our approach enables both intensity scaling and emotion blending by linearly combining parameter differences derived from instruction-tuned models. The overall framework

consists of three stages: (1) vector extraction from emotion-specific models, (2) composition of control vectors, and (3) application via parameter addition. We describe each of these components in the following subsections.

3.1 Expressing Compound Emotions via Model Merging

We propose a model merging method based on the linear composition of task vectors to produce compound emotional expressions. As illustrated in Figure 2, Chat Vectors—parameter differences between a neutral emotion model and each emotion-specific model—are combined with scalar weights and added to the neutral model.

In this visualization, Chat Vectors are depicted as armor pieces, where colors symbolize distinct emotional types (e.g., Emotion α and β). When combined with respective weights (w_α, w_β), the resulting armor takes on a blended color, such as purple, reflecting the composite emotion. A stronger weight for Emotion α results in a more reddish purple, visually signifying its dominance.

This metaphor illustrates how weighted blending enables fine-grained emotional control—e.g., emphasizing joy while keeping surprise subtle. Our experiments later confirm that such weighted combinations effectively modulate the emotional tone of generated text.

Assigning a weight of 1.0 to a Chat Vector yields a strongly expressed emotion, while lower weights reduce its influence. Blending with the neutral model allows for mild emotional expression.

To build the models, we fine-tune a base LLM on neutral-emotion data to create a neutral model. Further fine-tuning on emotion-specific corpora produces emotion-specific models, and their differences from the neutral model are used to derive Chat Vectors.

3.2 Dataset

To generate emotionally expressive text, we use the WRIME dataset (Kajiwara et al., 2021), a single-post social networking service (SNS) dataset comprising 43,200 Japanese-language social media posts authored by 80 participants. Although our long-term objective is to apply our approach to dialog systems, ideally using dialog datasets, we focus on WRIME here because it provides a large-scale dataset with reliable emotion annotations. Each post is annotated with two types of emotion ratings: (1) self-reported emotions by the author,

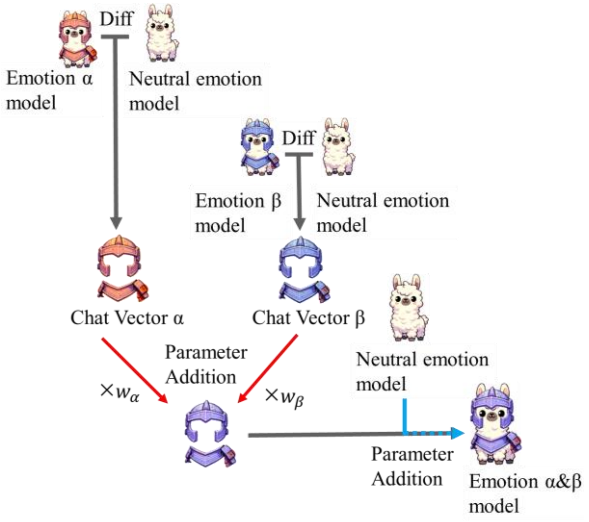


Figure 2: Model of the Proposed method.

and (2) perceived emotions as judged by three independent readers. The annotations are based on Plutchik's (1980) eight basic emotions—joy, sadness, anticipation, surprise, anger, fear, disgust, and trust—and rated on a 4-point scale: none, weak, medium, and strong. We rely on reader-perceived emotion annotations, because our goal is for users to correctly perceive the system's expressed emotions rather than for a system to mimic human internal states; this choice is also supported by prior findings that reader labels are more consistent and predictable than author self-reports (Kajiwara et al., 2021). Since each post is rated by three readers, we average the reader scores as the final emotion score for each emotion.

Because many posts express multiple emotions, we focus exclusively on posts characterized by a single dominant emotion to train emotion-specific models. We define such "single-emotion posts" using the following criteria:

1. All emotion scores are "none" → **neutral data**
2. Emotion α is rated medium or strong and has the highest among all emotions → **strong data for Emotion** α
3. Emotion α is rated weak and all other emotions are "none" → **weak data for Emotion** α
4. Posts not matching 1–3 are excluded from training

After filtering, the class distribution becomes markedly imbalanced; joy, sadness, surprise, and fear remain sufficiently represented for stable training, whereas anger, disgust, and trust become

low-resource. In addition, anticipation is difficult to distinguish from joy because the two occupy a similar region in valence-arousal space, which increases inter-annotator confusion. Accordingly, we restrict our experiments to four emotions—joy, sadness, surprise, and fear—which offer adequate sample sizes and higher annotator consistency, enabling clearer evaluation of intensity control and emotion blending.

3.3 Training Emotion-Specific Models

As our base model, we use llm-jp-3-1.8b-instruct[1], which is an instruction-tuned version of the 1.8-billion-parameter foundation model llm-jp-3-1.8b[2]. We first fine-tune this model using neutral emotion data to construct a neutral emotion model. Based on this, we then fine-tune four emotion-specific models, each corresponding to one of the target emotions.

For training the emotion-specific models, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2023), which fine-tunes the model by contrasting preferred and dispreferred outputs. In our case, we generate preference pairs by matching strong-emotion examples (preferred) with weak-emotion ones (non-preferred) for each target emotion, so that the model learns to favor stronger emotional expression. As a result, the number of training pairs for each emotion is capped at twice the size of the smaller class (weak or strong).

4 Evaluation Experiments

In this section, we verify whether the generated text reflects the specified emotional intensity.

4.1 Experimental Conditions

We use the Transformers library (Wolf et al., 2020), the Transformers Reinforcement Learning (TRL) library [3] for preference-based tuning, and the Parameter-Efficient Fine-Tuning (PEFT) library[4] to apply Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA parameters are set as rank=8 and alpha=16. GPT-4o mini is used as the baseline. Few-shot prompts include three randomly selected posts representing different intensity levels from the training set.

	Training Samples	Validation Samples
Neutral	2,401	93
Joy (pairs)	2,240	83
Sadness (pairs)	1,926	54
Surprise (pairs)	1,873	41
Fear (pairs)	1,466	27

Table 1: Number of Text Samples Used.

	Joy	Sadness	Surprise	Fear
Inter-Annotator Agreement (Average)	0.603	0.393	0.427	0.432
Agreement between Estimator and Annotator Average	0.668	0.539	0.456	0.512

Table 2: Agreement Rates in Human and Estimator-Based Emotion Evaluations.

We train the neutral model on 2,401 posts, and four emotion-specific models on a total of 7,505 preference pairs drawn from the WRIME corpus (joy 2,240; sadness 1,926; surprise 1,873; fear 1,466; see Table 1). The neutral emotion model is trained for one epoch, whereas the emotion-specific models are trained for up to four epochs.

For single emotion testing, 10 weight settings ([0.1, 0.2, ..., 1.0]) are tested with 100 generations each. For compound emotion testing, all combinations of 10 intensity levels for two emotions are tested with 100 generations per combination.

As a comparative experiment with the proposed method, we also generate texts using GPT-4o mini (gpt-4o-mini-2024-07-18) in a prompt-based method, where the desired emotion intensity or combination of intensities is explicitly specified as a numerical value in the prompt. The prompt to use GPT-4o mini is in Appendix A.1.

4.2 LLM-Based Emotion Intensity Estimator

We used the Llama-3.1-70B-Japanese-Instruct-2407 model (Ishigami, 2024), a Japanese fine-tuned variant of Llama-3.1-70B-Instruct (Grattafiori et al., 2024), as an LLM-as-a-judge, which estimates the intensity of emotions expressed in the given input text. When a text is provided along with a few-shot prompt (Appendix A.2), the model predicts the intensity of each emotion as one of four levels: none, weak, medium,

[1] https://huggingface.co/llm-jp/llm-jp-3-1.8b-instruct
[2] https://huggingface.co/llm-jp/llm-jp-3-1.8b
[3] https://github.com/huggingface/trl
[4] https://github.com/huggingface/peft

or strong. The LLM-as-a-judge model is quantized to 4-bit precision for computational efficiency.

To evaluate the reliability of the estimated emotion intensities, we compared the model outputs against the average human rating on 1,980 samples from the WRIME dataset, using Quadratic Weighted Kappa (Cohen, 1968) as the evaluation metric. As shown in Table 2, the estimation quality of the model achieves agreement levels comparable to or exceeding those of human annotations.

4.3 Evaluation Results by Emotion Intensity

4.3.1 Expressed Intensity for Single Emotion

In this experiment, we verify whether the generated texts accurately reflect the emotion weights configured during generation, using the emotion estimator described in Section 4.2.

For single-emotion intensity control, we vary the weight of the target emotion in increments of 0.1 and compute the distribution of estimated intensity levels for 100 generated texts at each setting.

Figure 3 shows the proportion of texts generated by the proposed method that are classified into four levels of joy intensity: None, Weak, Medium, or Strong. Figure 4 presents the corresponding results for the prompt-based method.

The results indicate that with the proposed method, increasing the specified emotion weight leads to stronger emotional expression in the generated texts.

In contrast, the prompt-based method exhibits intensity saturation at medium and higher levels.

4.3.2 Compound Emotion Expression

We conducted an experiment to estimate the expressed intensities of compound emotions using the same method as in Section 4.3.1. When combining Emotion α and Emotion β, we fixed the weight of Emotion α at 0.5 and varied the weight of Emotion β. Figure 5 shows the estimated intensity of the varied emotion, and Figure 6 shows the estimated intensity of the fixed emotion.

Although some variations were observed depending on the emotion pair, the general trend remained consistent: lower emotion weights resulted in weaker emotional expression, while higher weights led to stronger expression, even in the compound emotion setting.

Figures 7 and 8 show the results for the prompt-based method. Similar to the single-emotion experiments, this method tended to produce

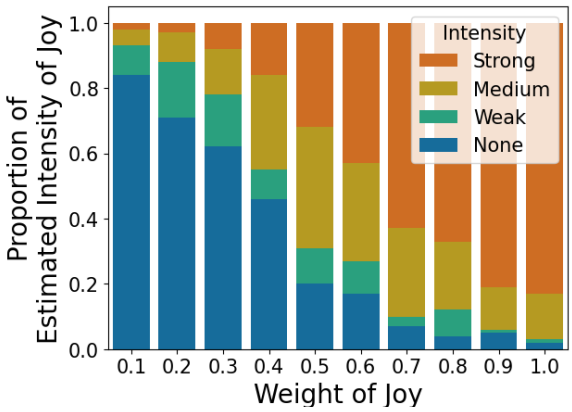


Figure 3: Distribution of predicted joy intensity levels (None–Strong) for texts generated by the **proposed method** across joy weights (0.1–1.0).

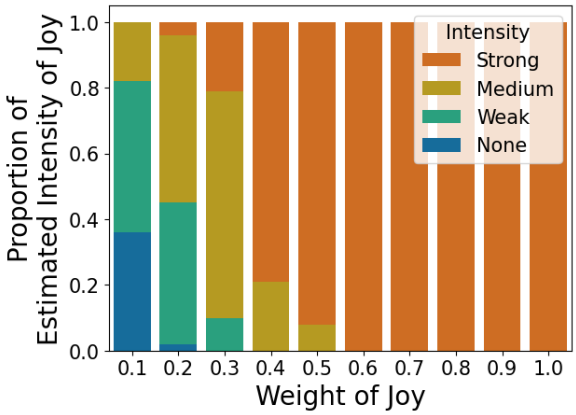


Figure 4: Distribution of predicted joy intensity levels (None–Strong) for texts generated by the **prompt-based method**.

strongly expressed emotions even at medium weight levels. In contrast to the proposed method, the estimated intensity of the fixed emotion also increased with the varied emotion's weight, indicating less stable control over the fixed component.

4.4 Influence of Emotional Combinations on Expression Strength

As shown in Figure 5, for emotion combinations other than fear and sadness, the proposed method successfully adjusted the intensity of the varied emotion in accordance with the specified weight. However, for the pair of fear and sadness, even when one of the emotions was assigned a low weight, the resulting text often expressed that emotion with medium or higher intensity. In psychological terms, *valence* refers to the affective dimension of emotional pleasantness, ranging from negative (e.g., sadness) to positive (e.g., joy). One possible explanation is that both fear and sadness are low-valence negative

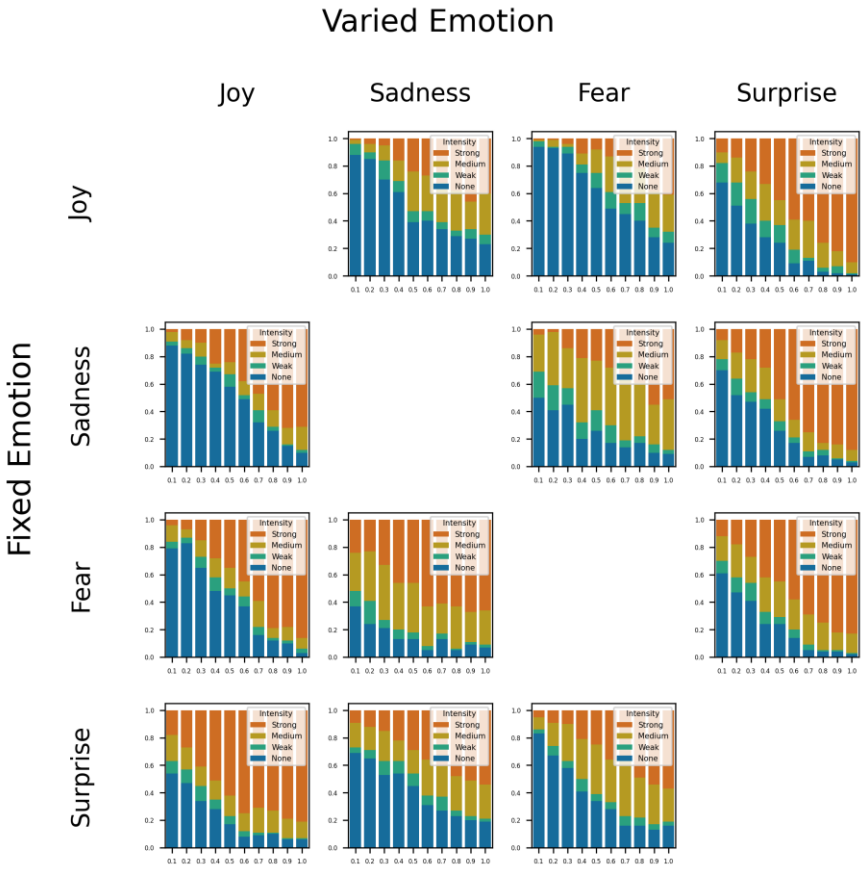


Figure 5: Estimated intensity of the **varied emotion** in texts generated by the **proposed method.**

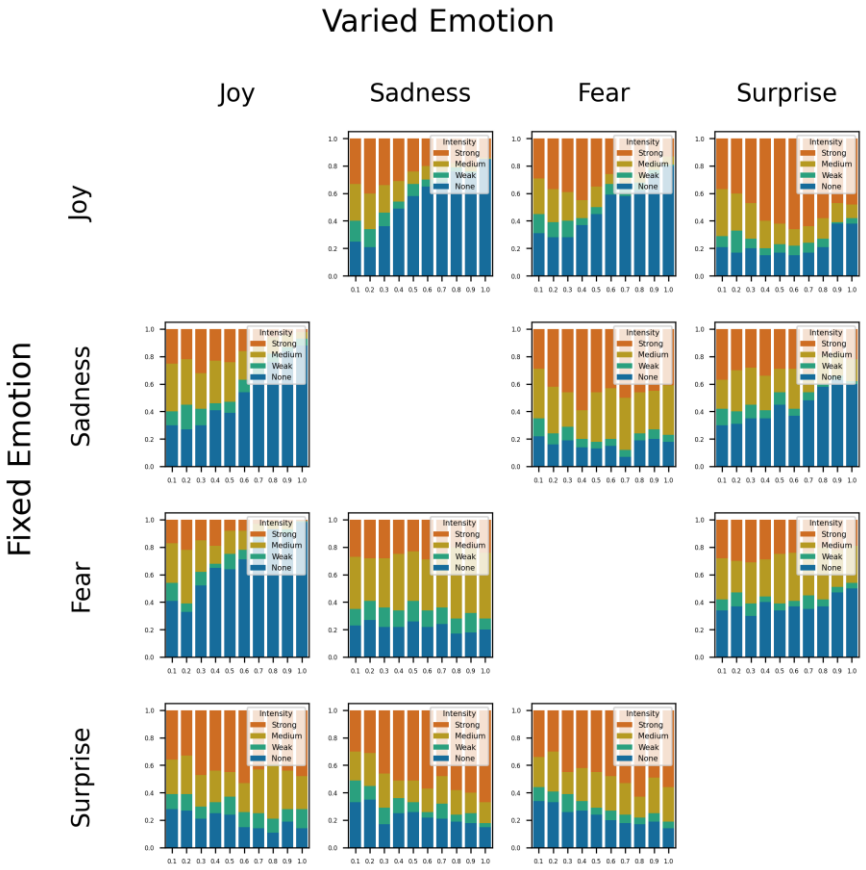


Figure 6: Estimated intensity of the **fixed emotion** in texts generated by the **proposed method.**

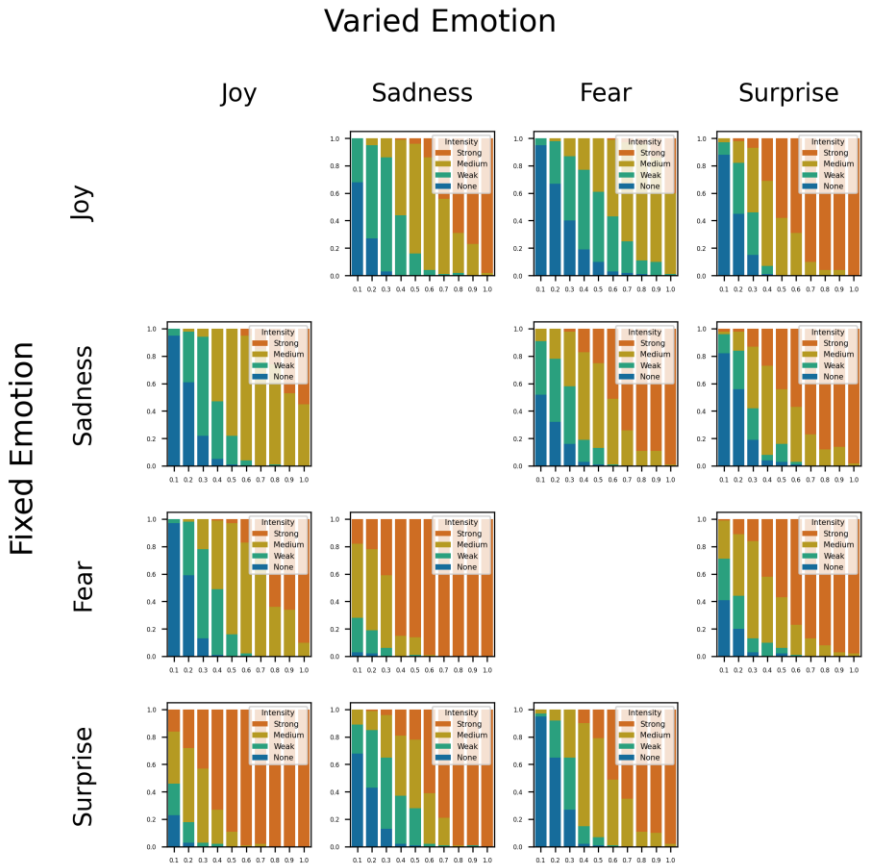


Figure 7: Estimated intensity of the **varied emotion** in texts generated by the **prompt-based method.**

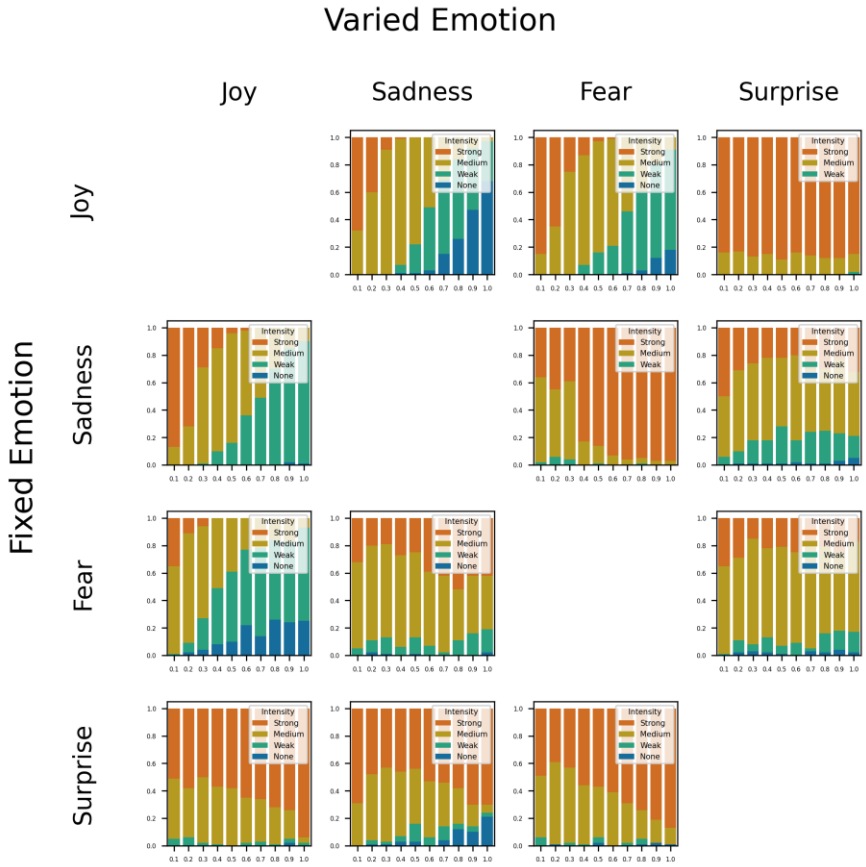


Figure 8: Estimated intensity of the **fixed emotion** in texts generated by the **prompt-based method.**

emotions with similar expressive characteristics, making it likely that expressions generated as fear were also perceived as sadness (or vice versa).
In contrast, Figure 6 reveals that in combinations such as joy–sadness and joy–fear, the fixed emotion's estimated intensity tended to decrease as the weight of the varied emotion increased. This suggests that when combining emotions with opposing valence—such as joy (high valence) and sadness or fear (low valence)—increasing the weight of one emotion can relatively suppress the expression of the fixed emotion.

Based on these results, two key issues emerge for future work. First, in combinations of emotions with similar valence (e.g., fear and sadness), low-weight emotions may be expressed more strongly than intended, indicating the need to suppress such unintended dominance. Second, in combinations of emotions with opposing valence, the strongly weighted emotion may overly suppress the fixed emotion, necessitating better control of this interaction. Addressing these issues could enable more precise control over complex emotional expressions.

These tendencies are consistent with psychological findings. For instance, the circumplex model of affect (Russell, 1980) structurally represents the difficulty of distinguishing between emotions that are close in valence and arousal space. Furthermore, neuroimaging evidence has shown that processing emotions with opposing valence simultaneously—such as joy and anger—activates conflict-monitoring regions like the dorsal anterior cingulate cortex (Wittfoth et al., 2010). These alignments suggest that the proposed method may be beginning to replicate aspects of human emotional cognition.

5 Conclusion

This study proposed a method for generating emotionally expressive text using model merging techniques. We constructed and evaluated models for joy, sadness, surprise, and fear, and confirmed that the method enables fine-grained control over both the intensity and combination of emotions in generated text, without retraining or manual prompt design.

Compared to prompt-based generation, our method achieves more stable and interpretable emotional outputs—especially in compound emotion settings—without requiring extensive retraining or elaborate prompts. The modular nature of task vector composition makes it highly scalable and efficient.

Our evaluation was conducted on single-turn social media posts; we did not test multi-turn dialog or long-range contextual effects on emotional expression. Consequently, applicability to conversational settings with evolving context remains to be verified.

As future work, we plan to compare our approach with alternative task-vector construction methods (e.g., Wang et al., 2025), investigate adaptive weighting strategies, and evaluate multi-turn emotional control in dialog generation. Beyond emotion control, we plan to extend our framework to persona-conditioned generation, and to train and evaluate it in multi-turn dialog settings to assess trait-consistent behaviors across turns.

References

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4):213-220.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. Computing Research Repository, arXiv:2407.21783. Version 1.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations*.

Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. 2024. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10943–10959, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *Proceedings of the 11th International Conference on Learning Representations*.

Ryosuke Ishigami. 2024. cyberagent/Llama-3.1-70B-Japanese-Instruct-2407.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. Computing Research Repository, arXiv:2310.11564. Version 1.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, pages 2095–2104, Online. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of emotion*. Academic press, New York:3-33.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, pages 53728-53741.

Weiqi Wang, Wengang Zhou, Zongmeng Zhang, Jie Zhao, and Houqiang Li. 2025. Controllable style arithmetic with language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15750–15799, Vienna, Austria. Association for Computational Linguistics.

Matthias Wittfoth, Christine Schröder, Dina M. Schardt, Reinhard Dengler, Hans-Jochen Heinze, Sonja A. Kotz. 2010. On emotional conflict: Interference resolution of happy and angry prosody reveals valence-specific effects, *Cerebral Cortex*, 20(2):383–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595-46623.

Shang Zhou, Feng Yao, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. Evaluating the smooth control of attribute intensity in text generation with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4348–4362, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Prompt for Text Generation Using GPT4o-mini

The following is the prompt used in the baseline condition with GPT-4o mini (gpt-4o-mini-2024-07-18). In this setting, the desired emotional intensity—or a blend of multiple emotions—is specified directly as real-valued weights in the prompt.

The prompt is written in Japanese, and the model is expected to generate a sentence that reflects the specified emotional profile.

Japanese Prompt:

System: あなたは一般的な SNS ユーザーです。
Example 1
User: 一回の投稿で{emotion_A}の感情を強 度[{emotion_A_intensity_1}]、{emotion_B}の感情を強度[{emotion_B_intensity_1}]で表現する SNS 投稿を書いてください。
Assistant: {example_post_1}
Example 2
…
Query
User: 一回の投稿で{emotion_A}の感情を強度[{w_A}]、{emotion_B}の感情を強度[{w_B}]で表現する SNS 投稿を書いてください。

Latin Script:

System: Anata wa ippanteki na SNS yu-za desu.
Example 1

User: Ikkai no toukou de {emotion_A} no kanjou wo kyoudo [{emotion_A_intensity_1}], {emotion_B} no kanjou wo kyoudo [{emotion_B_intensity_1}] de hyougen suru SNS toukou wo kaite kudasai.
Assistant: {example_post_1}
Example 2
…
Query
User: Ikkai no toukou de {emotion_A} no kanjou wo kyoudo [{w_A}], {emotion_B} no kanjou wo kyoudo [{w_B}] de hyougen suru SNS toukou wo kaite kudasai.

English Translation:

System: You are a typical social-media user.
Example 1
User: Please write a social-media post that conveys {emotion_A} with an intensity [{emotion_A_intensity_1}] and {emotion_B} with an intensity [{emotion_B_intensity_1}] in a single post.
Assistant: {example_post_1}
Example 2
…
Query
User: Please write a social-media post that conveys {emotion A} with an intensity [{w_A}] and {emotion_B} with an intensity [{w_B}] in a single post."

A.2 Prompt for Estimating the Intensity of Emotion Expressed in Text

The following is the few-shot prompt used for evaluating emotion intensity in the LLM-as-a-judge setting. Given a text input, the model is asked to predict the intensity of each emotion (e.g., joy, sadness, fear) as one of four levels: none, weak, medium, or strong.

This prompt is written in English, as the LLM-as-a-judge (Llama-3.1-70B-Japanese-Instruct) was found to perform more reliably with English instructions, even when judging Japanese input texts.

You will be given a Japanese Social Media post.
Your task is to rate the post on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Emotion of {target_emotion}
Level: (0 - 3)
How much emotion of {target_emotion} can you infer from the text?
0 is the lowest score, 3 is the highest.
Post: {example_post_score_1}
Your Answer(Score Only): 1
Post: {example_post_score_2}
Your Answer(Score Only): 2
Post: {example_post_score_3}
Your Answer(Score Only): 3
Post: {post_to_be_rated}
Your Answer(Score Only):

USA Model: Japanese Universal Sentiment Analysis Model & Construction of Japanese Sentiment Text Classification and Part of Speech Dataset

Chengguang Gan[1] **Qinghao Zhang**[2] **Tatsunori Mori**[1]
[1]Yokohama National University, Japan
ganchengguan@yahoo.co.jp, tmori@ynu.ac.jp
[2]Department of Information Convergence Engineering,
Pusan National University, South Korea
zhangqinghao@pusan.ac.kr

Abstract

Sentiment analysis is a pivotal task in the domain of natural language processing. It encompasses both text-level sentiment polarity classification and word-level Part of Speech(POS) sentiment polarity determination. Such analysis challenges models to understand text holistically while also extracting nuanced information. With the rise of Large Language Models(LLMs), new avenues for sentiment analysis have opened. This paper proposes enhancing performance by leveraging the Mutual Reinforcement Effect(MRE) between individual words and the overall text. It delves into how word polarity influences the overarching sentiment of a passage. To support our research, we annotated four novel Sentiment Text Classification and Part of Speech(SCPOS) datasets, building upon existing sentiment classification datasets. Furthermore, we developed a Universal Sentiment Analysis(USA) model, with a 7-billion parameter size. Experimental results revealed that our model surpassed the performance of gpt-3.5-turbo across all four datasets, underscoring the significance of MRE in sentiment analysis.

1 Introduction

In sentiment analysis, many datasets and models predominantly focus on polarity classification of an entire text. Alternatively, they extract specific words from the text for sentiment polarity categorization. However, these methods often overlook the reciprocal relationship between individual words and the overall sentiment of the text. In other words, while specific words can influence the sentiment classification of an entire text, the overall sentiment can, conversely, shape the sentiment polarity of individual words within the text.

The concept of Mutual Reinforcement Effect(MRE) was initially introduced in the realms of sentence categorization and Named Entity Recognition(NER) tasks(Gan et al., 2023). Here, it was employed to simultaneously enhance the accuracy of both classifications by amalgamating the sentence classification and NER tasks. Drawing inspiration from this, our study aims to investigate the possible presence of MRE between the sentiment polarity categorization of texts and individual words.

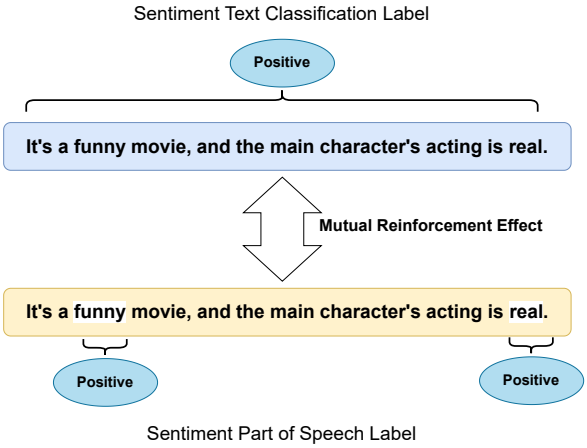


Figure 1: Example of Mutual Reinforcement Effect in Sentiment Text Classification task and Part of Speech task.

As illustrated in Figure 1, the upper section represents the sentiment polarity categorization of the entire text, while the subsequent section delves into the sentiment polarity extraction and labeling of adjectives within the text. The examples in the figure demonstrate that if a text is classified as positive, then certain adjectives within the text are likely to lean towards a positive sentiment. Conversely, when a text contains a plethora of adjectives or nouns with positive polarity, the overarching sentiment of the text likely leans towards the positive. Recognizing this interplay provides an opportunity to merge these two tasks, enhancing the model's granular understanding of the text and thus improving the performance of both tasks simultaneously. This synergistic approach underpins our proposed Sentiment Text Classification and Part of

Speech(SCPOS) dataset.

Moreover, there is a notable scarcity of datasets pertinent to sentiment analysis in the Japanese language. Currently, the MARC-ja dataset, part of the Japanese General Language Understanding Evaluation(JGLUE)(Kurihara et al., 2022), stands as the sole resource available for Japanese text sentiment classification. It is a binary classification dataset derived from user reviews on shopping websites. Remarkably, there is no dataset that offers sentiment polarity categorization at the word level for Japanese. The japanese itself presents unique characteristics distinct from other global languages. Not only is the Japanese script a combination of kanji, hiragana, and katakana, but its linguistic features—such as word transformations, sentence structures, and the nuanced use of honorifics—also set it apart. Given these intricacies, there's a compelling need to develop a dedicated dataset for Japanese sentiment analysis. This endeavor, therefore, addresses the existing deficiency in word-level sentiment polarity datasets for Japanese and also pioneers a fresh research trajectory in the SCPOS mixture task.

Otherwise, the advent of Large Language Models(LLMs) has heralded new avenues in sentiment analysis. When GPT-3(Brown et al., 2020) was launched, it exhibited a remarkable zero-shot capability alongside an innate ability to predict subsequent sentences based on provided input samples. This feature is often referred to as "In-context learning". Such a capability is particularly advantageous for languages like Japanese that have limited datasets. Remarkably, this method allows for the benefits of fine-tuning on expansive datasets to be achieved with minimal manually labeled data. Furthermore, the introduction of ChatGPT(Ouyang et al., 2022) showcased its robust generalization capabilities. LLMs possess extensive generalized knowledge and a depth of text comprehension that is challenging for smaller models to attain. This raises the intriguing question: Can the unique attributes of LLMs be leveraged to cultivate a generalized sentiment analysis model capable of handling both text sentiment analysis and Part-of-Speech(POS) sentiment polarity classification? Enhancements in word-level sentiment categorization and text-level sentiment classification could provide LLMs with heightened granularity and accuracy in discerning textual sentiments, thereby boosting their overall performance.

The remainder of this paper delineates the related work(§2), introduces the construction of the SCPOS dataset(§3), explains the training of the USA model(§4), presents an evaluation of both the baseline and USA models on the SCPOS dataset(§5), and analysis of MRE(§B).

2 Related Work

2.1 Initial studies in sentiment analysis

Sentiment classification emerged as a specialized subset of text classification tasks(Pang et al., 2002), with the primary objective of discerning the overall sentiment of a text as either negative or positive. In above study, employed machine learning methods such as naïve bayes, maximum entropy, and Support Vector Machines(SVM) to categorize the sentiment of texts. Notably, the research first time introduced movie reviews as a dataset for sentiment classification. Over the following decade, a plethora of datasets were developed for sentiment classification(Medhat et al., 2014). These encompassed a range of sources including News Articles(Bai, 2011), Hotel Reviews(Wu and Tan, 2011), Restaurant Reviews(Robaldo and Di Caro, 2013), Tweets(Go et al., 2009)(Agarwal et al., 2011), Blogs(Yu et al., 2013), and more. As a result of these expanded datasets, sentiment classification algorithms have undergone significant evolution. They progressed from initial rule-based systems to statistical models, and further to Hidden Markov Models(HMM) and Conditional Random Fields(CRF). Such advancements have robustly shaped the foundation for future sentiment analysis endeavors.

2.2 Deep Learning for Sentiment Analysis

Over the past decade, there has been a significant surge in the use of neural networks in AI. This has ushered sentiment categorization into the realm of deep learning(Zhang et al., 2018). Initially, researchers employed word embeddings and rudimentary neural network models for sentiment analysis. Subsequently, the advent of transformer models(Vaswani et al., 2017) marked the transition into the era of Pre-trained Language Models(PLMs), with BERT(Devlin et al., 2018) and T5(Raffel et al., 2020) models enhancing the accuracy of text sentiment categorization considerably. Presently, we are in the era of Large Language Models(LLMs) with the introduction of models like GPT-3(Brown et al., 2020), ChatGPT(Ouyang et al., 2022), and GPT-4(OpenAI, 2023), which claim to address a myriad

	SRW	NVA	N	VA
Label	positive, Xnegative, neutral, Xpositive, negative	positive, neutral, negative	positive, neutral, negative	positive, negative
Count	5	3	3	2
Total	2000	187528	187528	187528

Table 1: Statistical data of SCPOS. **NVA** dataset represents a dataset consisting of nouns, verbs and adjectives. **N** then represents consisting of nouns only. **VA** represents consisting of verbs and adjectives.

of NLP challenges. However, when it comes to sentiment categorization accuracy, these LLMs often lag behind fine-tuned smaller models. We aspire for LLMs to leverage their extensive pre-training knowledge and comprehension when training a universal sentiment analysis model. This is crucial in ensuring high sentiment classification accuracy, especially in 0/1-shot scenarios.

(Li et al., 2023) introduced UniSA, a unified generative framework to integrate sentiment analysis subtasks, addressing challenges like modality alignment, varied input/output formats, and dataset bias. They also curated SAEval, a benchmark that consolidates various sentiment subtask datasets. Their results highlighted UniSA's competitive performance across all subtasks in sentiment analysis. In contrast to previous research, this study emphasizes the relationship between individual words and the overall text. Specifically, it trains LLMs for sentiment analysis using a multi-task approach, distinguishing it from earlier studies.

2.3 Resource of Japanese Sentiment Analysis

The field of sentiment analysis in Japanese currently faces a significant dearth of resources. Existing datasets, such as one that categorizes shopping site reviews, offer only binary classifications(Kurihara et al., 2022). Remarkably, there is no specialized dataset available specifically for word-level sentiment polarity classification in Japanese. However, two sentiment polarity dictionaries do exist, focusing on common Japanese nouns and verb & adjectives(Kobayashi et al., 2005)(Higashiyama et al., 2008). Despite these dictionaries, word-level polarity classifications have not been applied to broader textual contexts. Thus, there is an evident and pressing need for a comprehensive dataset tailored for sentiment analysis in this language. Moreover, the endeavor to train generalized LLMs specifically for sentiment analysis remains uncharted territory. Previous efforts have largely centered around training universal Named Entity Recognition(NER) LLMs using open-domain NER datasets(Zhou et al., 2023b). The outcomes of such 0-shot have proven to be commendable.

3 Sentiment Text Classification and Part of Speech Dataset

In this chapter, we present the methodology employed in the creation of the SCPOS dataset. This elucidation facilitates a clearer comprehension when one proceeds to the training of the USA model. This chapter is organized into two subsections: the first addresses manual construction of the SCPOS datasets, while the second discusses its rule-based construction. All four SCPOS datasets are annotated based on the MARC-ja dataset from JGLUE. Table 1 provides statistical data regarding the SCPOS dataset, and Figure 2 offers a visual representation of its composition and labeling process, encompassing both test set and train corpus.

3.1 SCPOS dataset of Manually Annotated and LLM generated

In this section, we delineate the process of constructing the Sentiment Related Word (SRW) dataset and outline the method for utilizing manually annotated data to train the LLM, enabling it to autonomously annotate the dataset. Within the manually annotated dataset, emphasis is placed on words that significantly influence the overarching sentiment polarity of the text. Consequently, individual words are not labeled for their specific sentiment polarity. This approach was adopted as adjectives play a predominant role in determining the text's overall sentiment. As a result, all adjectives in the SRW dataset were annotated. Furthermore, nouns and verbs associated with either a negative or positive sentiment were annotated, while those deemed neutral were excluded from annotation.

A conventional POS sentiment polarity categorization typically encompasses 2-3 labels, such as Positive, Neutral, and Negative. To this taxonomy,

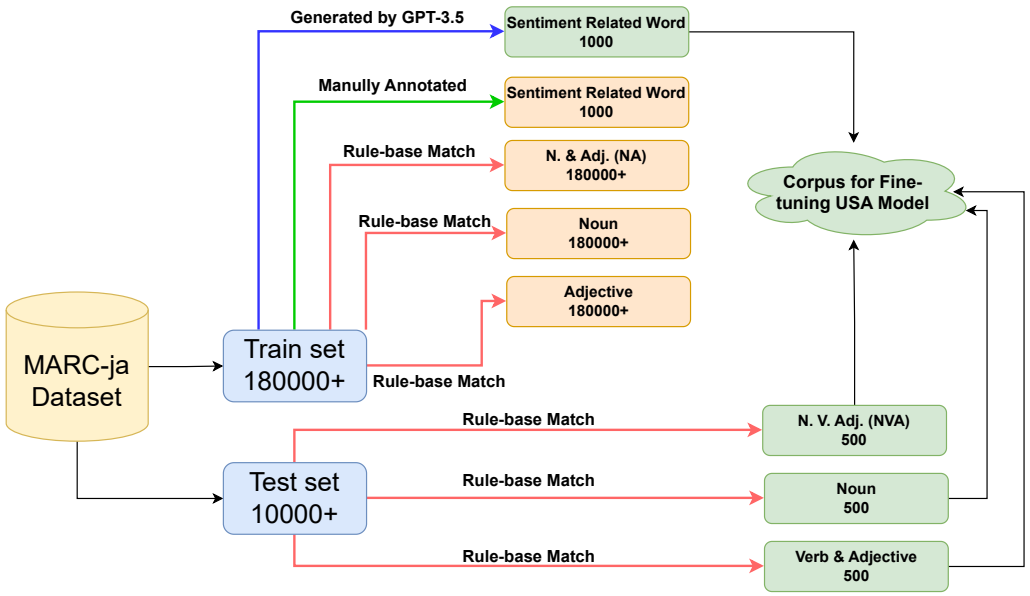


Figure 2: Overview of SCPOS dataset construction process. Different colored lines correspond to different methods of construction. The number below the sub-dataset name is the dataset sample size.

we introduce two additional labels: Xnegative and Xpositive. As illustrated in Appendix A Figure 5, the sentiment conveyed by a positive adjective, like "happy," is inverted when preceded by a negative word, turning the otherwise positive adjective into a negative context. Conversely, a negative adjective, such as "boring," when preceded by a negative word, can be transformed to convey a positive sentiment. Recognizing these nuances is crucial in accurate sentiment polarity determination. This is particularly salient in Japanese where the linguistic structures and words indicating negation are both diverse and markedly distinct from many other languages. Such complexities may lead the model to misinterpret the sentiment polarity of words or entire texts. Consequently, by introducing these two labels, we aim to enhance the model's proficiency in understanding the lexical shifts of adjectives in the presence of negation, ensuring more accurate sentiment polarity assessments. It's also worth noting that, with the addition of negative words or phrases, multiple word might be required for a single label, extending to two words or even a phrase.

Following the aforementioned annotation rules, we obtained 1,000 samples, constituting a high-quality, manually annotated dataset. We utilized this dataset to fine-tune the GPT-3.5 model[1]. Subsequently, the trained GPT-3.5 model was employed to automatically annotate an additional 1,000 unlabeled data samples.

To fix our results, we selected the GPT-3.5-Turbo-0613 version for the fine-tuning process. The fine-tuning incorporated a blend of In-context Learning (ICL) and Instruction Learning (IL) methodologies for the input format. For ICL, a moderately lengthy sample was randomly chosen as a sequence. For IL, an instructive prompt was strategically placed between the sample and the sentence set for categorization. This prompt guides the model to classify and extract subsequent text based on the preceding sample. Further details on ICL and IL will be elaborated upon in the upcoming "USA Model Training" section.

Upon completing this process, the model successfully auto-generated 1,000 accurately labeled datasets. We subsequently conducted a manual review of these 1,000 samples, rectifying any incorrect labels. This streamlined approach significantly reduced the manual labor and time investment.

3.2 SCPOS dataset of Rule-based Matching

In this section, we outline the method employed to use a Japanese word polarity dictionary[2] for matching corresponding words within a sentence. We subsequently constructed three SCPOS datasets, each containing different parts-of-speech (POS) classifications.

[1] https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[2] https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html

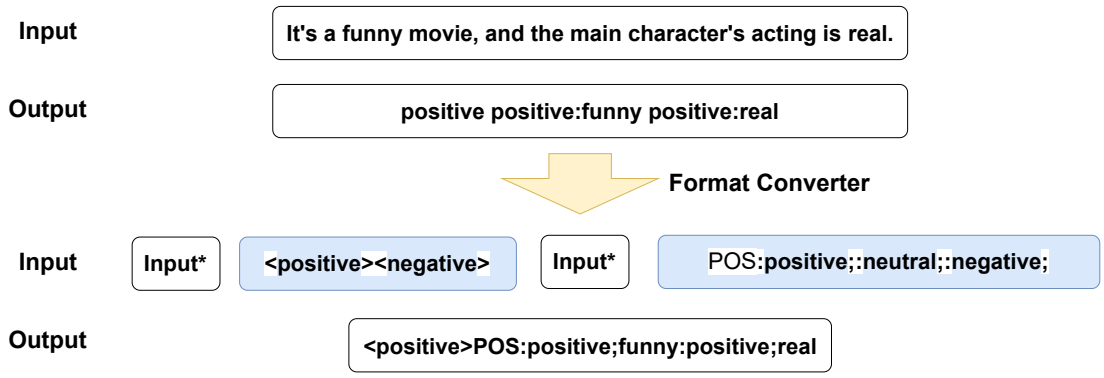


Figure 3: Example of datasets input and ouput format.

Initially, we utilized two distinct word polarity dictionaries. The first dictionary comprises 13,264 prevalent Japanese nouns, categorized into three polarity classes: positive, neutral, and negative. In contrast, the second dictionary has 5,280 frequently used Japanese verbs and adjectives, classified into two polarity groups: positive and negative. Figure 2 illustrates the procedure of using these dictionaries for rule-based matching against the training set of the MARC-ja dataset. It's worth noting that if identical words are repeated within the text, each occurrence is identified and listed sequentially. Following this process, we derived two SCPOS datasets labeled according to their respective POS: Noun (N) and Verb & Adjective (VA). Furthermore, by amalgamating both dictionaries, we crafted a comprehensive polarity thesaurus covering three POS: Noun, Verb, and Adjective. This merged thesaurus was then matched against the MARC-ja dataset, resulting in the creation of the SCPOS dataset encompassing Noun, Verb, and Adjective (NVA).

Consequently, all the derived datasets not only retain the original text sentiment classification label but also incorporate word polarity classifications.

4 Universal Sentiment Analysis Model

In this chapter, we present the Universal Sentiment Analysis Model (USA). The discussion is organized into three subsections to elucidate the process systematically. First, we address the preparation of the training corpus, followed by the conversion of input and output formats. We conclude with the implementation of ICL and IL techniques, which are essential for training an LLM to achieve robust performance in sentiment analysis tasks.

4.1 Constrcution of Train Corpus

To fine-tune the LLM, we initially prepared a specialized corpus. As depicted in Figure 2, the green block represents the corpus constructed for this purpose. We extracted three sub-datasets, each containing 500 samples, from the MARC-ja dataset's test set. A consistent rule-based method was employed for both matching and generation. When combined with the 1,000 samples previously generated during the fine-tuning of the GPT-3.5 model, we amassed a composite dataset of 2,500 samples, encompassing four distinct subtasks.

Our primary objective was to maximize the model's learning from the SRW sub-dataset. This emphasis on the SRW dataset is because it solely concentrates on words that align with the overall sentiment polarity of the text, making it an optimal choice for teaching the model about words that significantly influence the text's sentiment. Additionally, the SRW dataset boasts exceptionally high labeling quality. Consequently, we doubled the weight of the SRW sub-dataset in the training corpus compared to the other datasets.

Previous research on large model fine-tuning suggests that only a minimal amount of high-quality data is required to effectively fine-tune an LLM(Zhou et al., 2023a). In alignment with this understanding, we used a mere 2,500 samples to train the LLM. Utilizing fewer samples for fine-tuning also ensures better preservation of knowledge acquired by the LLM during its pre-training phase.

4.2 Format of Input and Output

After preprocessing the training corpus, it's essential to address the input and output data formats. We bifurcate the role of the Format Converter (FC) into two primary functions. First, the FC standard-

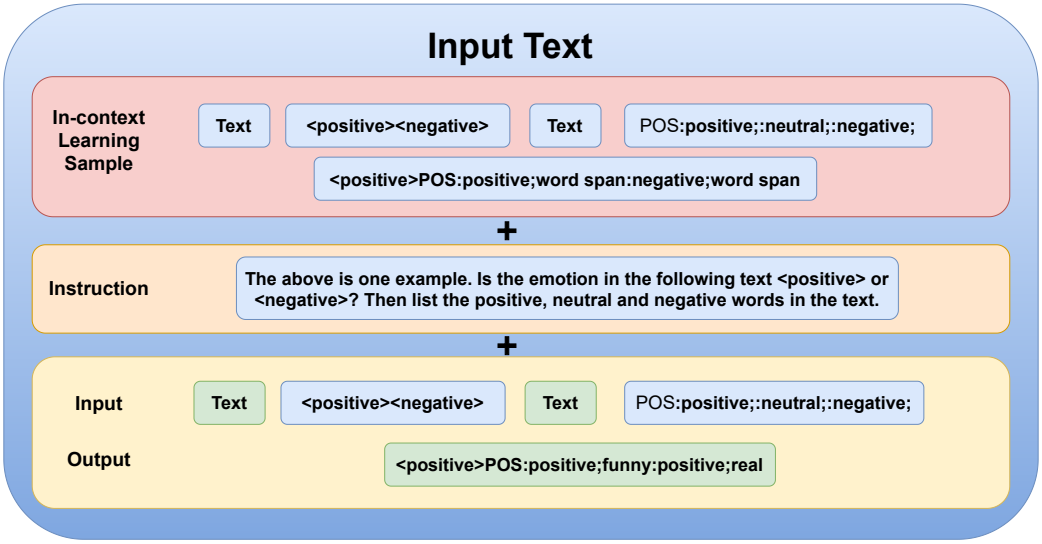


Figure 4: Example of In-context Learning and Instruction Learning of input format.

izes the input and output formats across all tasks, ensuring that both the text sentiment polarity classification and the word sentiment polarity classification tasks utilize a consistent format. Secondly, the FC serves the function of providing a prompt. The concept of a prompt was introduced to guide the model in generating desired labels with higher accuracy in few-shot scenarios. In our research, we consistently structure both input and output using fixed symbols and words. This strategy aims to employ these symbols and words as cues, guiding the model to produce tokens corresponding to the desired labels following these cues.

According to the two roles proposed earlier, we refer to the structure of FC in SLG Framework(Gan et al., 2023). Figure 3 illustrates the transformation of the input and output format, using the original dataset. The input comprises the text and its corresponding overall sentiment polarity categorization labels. Additionally, the output provides polarity labels for individual words and their respective word spans. For simplicity in subsequent discussions, we will refer to these as 'PW pairs'. These PW pairs maintain a one-to-one correspondence between the labels and word pairs.

In the lower segment of Figure 3 and Appendix D Figure 6, the FC-converted input and output formats are depicted. The blue block represents a fixed sequence of tags and words in the input. Initially, the text for categorization is input, followed by the addition of two text polarity classification labels. This text is then re-input, and the cue word "POS" is concatenated with the next three word polarity classification labels. Together, these elements form a comprehensive input sequence for the model.

In the output section, text sentiment polarity classification labels are enclosed by pink background symbols. The trailing "POS" serves as an indicator for the model to begin producing PW pairs. The characters ":", set against a dark blue background, and ";" signify the start and finish of word sentiment polarity labeling, respectively. With this, the transition of the input and output format is effectively completed.

4.3 In-context Learning and Instruction Learning

In this section, we elucidate the process by which the 'transformed format' dataset is packaged by ICL and IL, and subsequently fed into the model. Figure 4 illustrates the sequence: the ICL sample is at the top, followed by the IL instruction problem, and lastly, the input in its transformed format. Blue blocks denote sequences consistent across all input samples, while green blocks indicate segments replaced based on specific input samples.

Focusing first on the ICL component: a medium-length text was randomly selected from the MARC-ja test set, ensuring avoidance of the 1500 training phrases previously extracted. We intentionally did not choose a lengthy text encompassing a wide array of POS samples. The chosen samples were labeled following the conventions of the prior four datasets. This labeling was then formatted into the final ICL sample using the FC format.

Turning to the instructional segment, it's divided into three distinct sentences. First, in the first sentence we hint that the sequence in front of the

model is an example. The second sentence asks whether the overall sentiment polarity classification of the text given next is positive or negative. Here again, marks are placed on both sides of the positive and negative words. Finally, the third sentence asks the model to list the words in the text that are related to the polarity of the sentiment. This instruction requires modifications based on the specifics of four datasets. As an instance, the version portrayed in Figure 4 corresponds to the SRW sub-dataset. For the NVA dataset, "word" should be replaced with "noun, verb, and adjective".

In summation, the combination of the ICL sample problem and Instruction, along with the Input, establishes the sequence for model input, leading to the output generation. Employing IL serves as a prompt, aligning the downstream task with the instruction fine-tuning phase in LLM pre-training. Simultaneously, the integration of ICL potentially extends instructional length, thereby refining the generative outcomes of LLMs.

5 Evaluation

In this chapter, we delineate the experimental framework and define the evaluation metrics. Subsequently, we assess the chosen models employing the SCPOS dataset.

5.1 Experiment Set

To begin, due to constraints in time and resources, we opted for a sample size of 1,000, randomly selected from all SCPOS sub-datasets. Each sample was tested three times, and the average value was taken as the final result. For model selection that the LLaMA2-7B model served as the base model for the USA model. Under the USA-7B model umbrella, we trained two distinct versions: 1.USA-7B(ICL+IL) using the combined ICL and IL data format . 2.USA-7B(IL) solely utilizing the IL data format. For comparative purposes, we incorporated the T5-base-Japanese model, which operates on the SLG framework. Additionally, the original gpt-3.5-turbo model was chosen as a benchmark for comparison. It's imperative to note that our selection of comparative models wasn't arbitrary. We rigorously tested a multitude of LLMs trained on Japanese corpora, as well as the LLaMA2-70B model. These models were subjected to both 1-shot ICL and IL evaluations. Regrettably, none of the models produced satisfactory results. They frequently generated text echoing the input or text irrelevant to the task. Consequently, our final comparison was limited to the T5 and gpt-3.5-turbo models.

The T5 model lacks the capability for few-shot learning in tasks with long sequences. Therefore, for training the T5 model, we utilized 1,000 randomly selected samples, while another set of 1,000 samples served as the test set. In the SRW task, we used the GPT-3.5-generated data for training and a manually labeled dataset for testing.

Upon evaluating the GPT-3.5 model, we observed that simultaneous use of ICL and IL during testing essentially prevents the generation of the desired text. Consequently, for testing GPT-3.5, only ICL was employed to input the dataset into the model.

5.2 Evaluation Metrics

In the selection of evaluation metrics, because of the uncertainty of the results generated by the generative model. It leads to the length of generated sequences is not fixed like sequence labeling models. So it is difficult to use metrics such as F1 to evaluate the task. Especially in this task, some texts contain as many as 20-30 PW pairs. Here accuracy is used as the evaluation metrics. Three different accuracies are categorized according to the type of task.

The first one is the text sentiment polarity classification accuracy(i.e. ACC_{SC}). the characters in the text classification part of the generated sequence and the actual text are intercepted and compared, and all of them are considered to be correct if they are equal (e. g. Generated: <positive> = Actual: <positive>).

Next is the calculation of the accuracy of the remaining part of the PW pair. All the PW pairs of the generated and actual sequences are partitioned according to the ":" and ";" notations. Let the set P contain n PW pairs (X_1, X_2, ..., X_n). Two sets can be obtained, PW pair $P_{generated}$ for the generated sequence and PW pair P_{actual} for the actual sequence, and then the sets$P_{generated}$ and P_{actual} are matched. The number of matched pairs is divided by the total number of PW pairs of the actual sequence to get the ACC_{pos}.

Finally the SCPOS accuracy is calculated. When both the ACC_{text} and the ACC_{pos} of the same sample are equal to 1, it is computed as a correct sample. The number of all correct samples is divided by the total number of samples to get the ACC_{SCPOS}.

	SRW			NVA		
Accuracy	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}
T5-base(SLG)	88.21	55.57	17.28	87.30	26.22	1.60
USA-7B(ICL+IL)	**89.60**	**56.32**	**18.10**	**90.20**	**60.09**	**3.97**
USA-7B(IL)	88.60	53.24	17.50	88.43	55.28	3.60
GPT-3.5	53.6	14.99	1.60	73.20	10.34	0.13
	Nous			**V & Adj**		
	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}
T5-base(SLG)	89.50	27.62	3.00	83.00	**73.84**	**52.47**
USA-7B(ICL+IL)	**91.50**	**62.41**	**6.83**	**92.17**	64.94	50.90
USA-7B(IL)	90.80	57.74	4.73	89.33	69.83	52.43
GPT-3.5	73.83	10.44	0.23	78.83	15.45	9.87

Table 2: Results for the four SCPOS subdatasets on the four models. T5-base was fine-tuned. USB-7B(ICL+IL) was assessed with a 1-shot approach. USB-7B(IL) utilized a 0-shot approach. GPT-3.5 was tested using the 1-shot ICL method.

5.3 Results

As illustrated in Table 2, the four models were tested on the four sub-datasets. It is evident that USB-7B(ICL+IL) outperforms the others by achieving the highest accuracy in 10 out of 12 evaluations across these datasets. This underscores its robust capability in the SCPOS task. Notably, both USB-7B(ICL+IL) and USB-7B(IL) models significantly surpass the accuracy of the GPT-3.5 model.

Historically, in the extraction and classification of constructed data, Large Language Models (LLMs) have been overshadowed by smaller Pre-trained Language Models (PLMs), as evidenced by the results of GPT-3.5. Nevertheless, after fine-tuning with the SCPOS mixed training corpus, the USA models demonstrate an improved ability to outshine their counterparts. Leveraging the extensive pre-training corpus and parameter set of LLMs, they can surpass PLMs that are exclusively fine-tuned with the entirety of sub-tasks in 0-shot scenarios. This also indicates that the finely-tuned USB-7B model, even with fewer samples, can exceed the performance of the full-volume fine-tuned T5-base.

In the VA sub-datasets, the USA model did not surpass the performance of the fine-tuned T5 model. The reason for this disparity lies in the fact that the average input and output lengths in the VA dataset are considerably shorter than in the other three datasets. Consequently, the fine-tuned T5 model demonstrates superior results compared to the 0/1-shot USA model when handling short sequences.

Furthermore, the performance of the USA-7B model augmented with ICL is superior to its counterpart using IL exclusively. This suggests that the combined use of ICL and IL enhances the model's comprehension of each sub-dataset. Moreover, distinct ICL and IL were employed for each of the four sub-datasets. This approach enhances the model's ability to differentiate and comprehend the various tasks.

6 Conclusion

In this paper, we introduces a novel sentiment categorization task termed the SCPOS task and details the creation of four distinct sub-datasets. Utilizing the SCPOS task, we trained the USA-7B model, achieving exemplary performance in 0/few-shot sentiment classification scenarios. We anticipate that the proposed SCPOS dataset and the USA-7B model will pave the way for fresh research avenues and areas of focus in sentiment analysis.

In future research, we plan to evaluate additional LLMs using the SCPOS dataset. Our objectives are to ascertain the presence of MRE in both the text sentiment classification and the POS sentiment polarity classification tasks. Furthermore, we aim to determine whether LLMs leverage MRE to enhance the performance of these tasks.

7 Limitations

In this work, the Sentiment Text Classification and Part of Speech (SCPOS) dataset was developed, comprising four sub-datasets labeled according to different lexical rules. Additionally, two USA models were trained using the IL and ICL methods. While the USA models outperformed the baseline, this study focused exclusively on the Japanese language and did not extend the analysis to other languages. Furthermore, the ablation experiments did not result in significant improvements in sentence classification accuracy. These limitations highlight areas for further investigation, which we plan to explore in future research.

Acknowledgements

This research was supported in part by JSPS KAKENHI Grant Numbers JP24K15084, JP23H00491 and JP22K00502.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

Xue Bai. 2011. Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4):732–742. Enterprise Risk and Security Management: Data, Text and Web Mining.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023. Sentence-to-label generation framework for multi-task learning of japanese sentence classification and named entity recognition. In *Natural Language Processing and Information Systems*, pages 257–270, Cham. Springer Nature Switzerland.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Learning sentiment of nouns from selectional preferences of verbs and adjectives. *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, 4:584–587.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, and Kenji Tateishi. 2005. Collecting evaluative expressions for opinion extraction. *Journal of Natural Language Processing*, 12(3):203–222.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. Jglue: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. *Preprint*, arXiv:2309.01339.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Livio Robaldo and Luigi Di Caro. 2013. Opinionmining-ml. *Computer Standards & Interfaces*, 35(5):454–469.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiong Wu and Songbo Tan. 2011. A two-stage framework for cross-domain sentiment classification. *Expert Systems with Applications*, 38(11):14269–14275.

Yang Yu, Wenjing Duan, and Qing Cao. 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision support systems*, 55(4):919–926.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023b. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Example of Xpositive and Xnegative Label

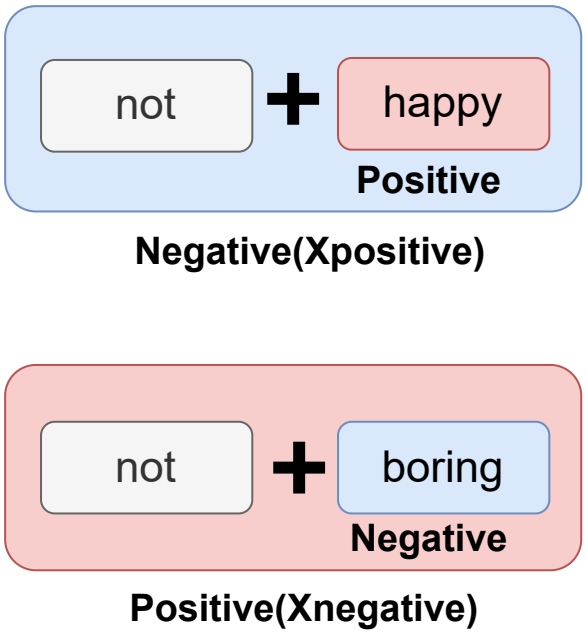


Figure 5: Example of Xpositive and Xnegative label.

B Analysis of Mutual Reinforcement Effect

To determine if MRE is prevalent in both the SC and POS sentiment polarity classification tasks, we carried out distinct evaluations for each task. To ensure consistency and control for other variables, we employed the T5-base model across all four datasets. As evidenced in Table 3, when the SC task is executed independently, the performance across all datasets surpasses that of the SCPOS task. This is likely because the SC task necessitates generating only a brief text categorization label, leading to enhanced accuracy. In contrast, when the POS task is conducted independently, there's a marked decline in accuracy across all datasets. This further substantiates the presence of MRE in the sentiment polarity classification task.

C Details of Experiment

To train the USA-7B model, we employed four NVIDIA A800 GPUs, each with 80GB, running for approximately 20 hours. We trained two models for durations of 2 and 3 epochs, respectively, using a learning rate of 1e-5. Collectively, they consumed between 160-200GB of GPU memory.

For the dataset auto-annotation task, we utilized the fine-tuned GPT-3.5 model. Our dataset comprised 1,000 data points, amounting to 4.5 million tokens, and the associated cost was approximately $57. We conducted 3 epochs of training for this task. Unless specified, we maintained the default settings provided by OpenAI for all other training parameters. In subsequent training and testing experiments involving the T5 model, we used an RTX 3090 GPU with 24GB. Regarding the generation parameter settings for the models. The USA-7B model was set with the following parameters:

- max new tokens: 2000
- repetition penalty: 1.3
- temperature: 1.0
- top_p: 0.7
- top_k: 40

For both the T5 and GPT-3.5 models, we only adjusted the "max new tokens" setting to 400. It's worth noting that the average output length of the SCPOS dataset, which boasts the longest average, is 144 tokens. All other parameters remained at their default values.

D Input and Output Example of USA model

	SRW			NVA		
Accuracy	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}
SCPOS	88.21	**55.57**	**17.28**	87.30	**26.22**	**1.60**
SC Only	**92.91**	-	-	**91.40**	-	-
POS Only	-	30.69	-	-	24.65	-
	Nous			**Verb & Adjective**		
	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}	ACC_{SC}	ACC_{POS}	ACC_{SCPOS}
SCPOS	89.50	**27.62**	**3.00**	83.00	**73.84**	**52.47**
SC Only	**91.40**	-	-	**91.40**	-	-
POS Only	-	24.05	-	-	52.04	-

Table 3: The results of the SC and POS tasks were tested separately using the T5-base model.

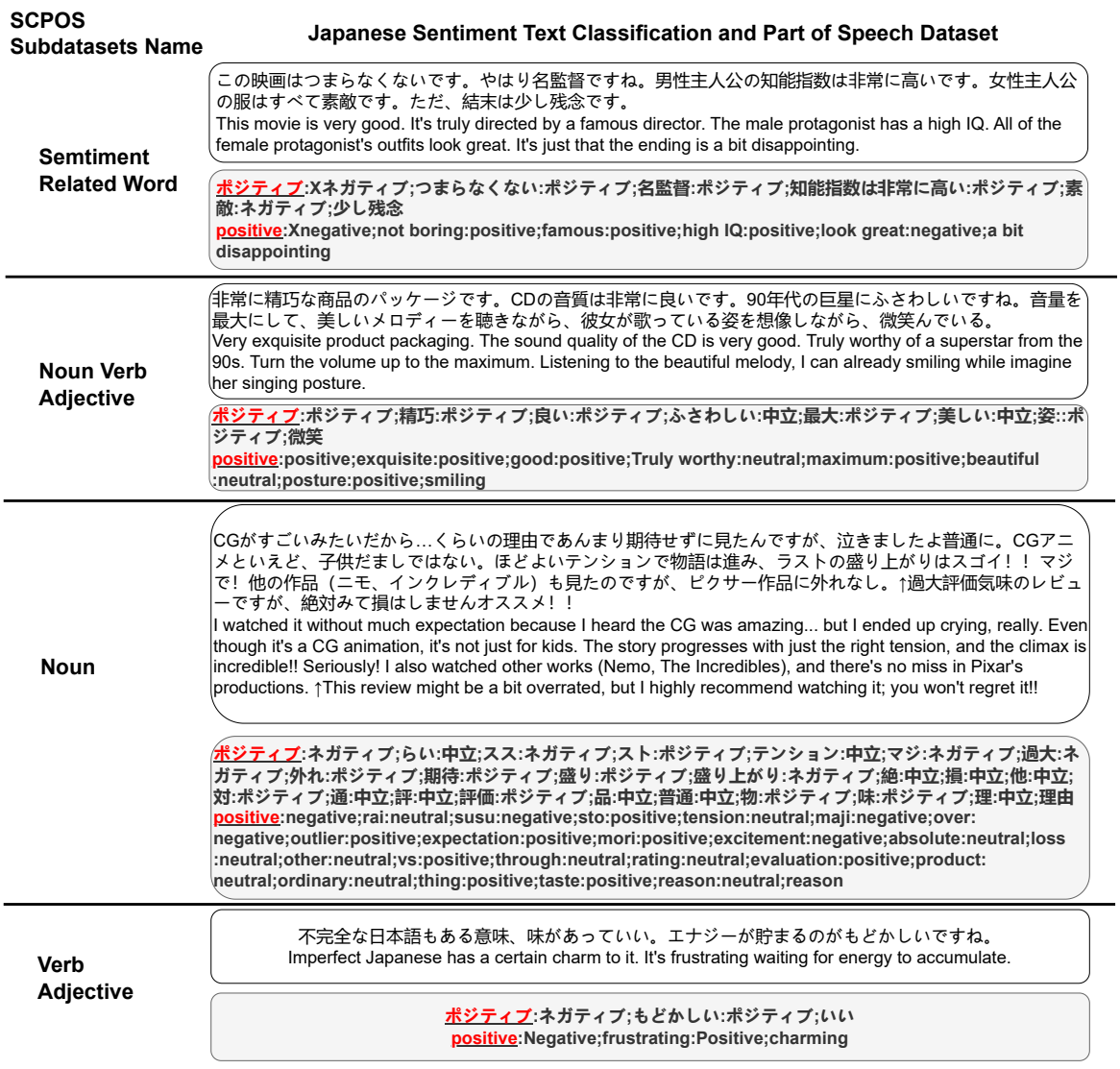


Figure 6: Input and Output Example of USA model.

SiDiaC: Sinhala Diachronic Corpus

Nevidu Jayatilleke and Nisansa de Silva
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{nevidu.25, NisansaDdS}@cse.mrt.ac.lk

Abstract

SiDiaC, the first comprehensive *Sinhala Diachronic Corpus*, covers a historical span from the 5th to the 20th century CE. SiDiaC comprises 58k words across 46 literary works, annotated carefully based on the written date, after filtering based on availability, authorship, copyright compliance, and data attribution. Texts from the *National Library of Sri Lanka* were digitised using Google Document AI OCR engine, followed by post-processing to correct formatting and modernise the orthography. The construction of SiDiaC was informed by practices from other corpora, such as *FarPaHC*, particularly in syntactic annotation and text normalisation strategies, due to the shared characteristics of low-resourced language status. This corpus is categorised based on genres into two layers: primary and secondary. Primary categorisation is binary, classifying each book into Non-Fiction or Fiction, while the secondary categorisation is more specific, grouping texts under Religious, History, Poetry, Language, and Medical genres. Despite challenges including limited access to rare texts and reliance on secondary date sources, SiDiaC serves as a foundational resource for Sinhala NLP, significantly extending the resources available for Sinhala, enabling diachronic studies in lexical change, neologism tracking, historical syntax, and corpus-based lexicography.

1 Introduction

Sola lingua bona est lingua mortua[1]; given that all languages that are in use, evolve with a gradual process of linguistic change over time, proposed to have originated from an initially gestural communication system (Corballis, 2017). Factors affecting this complex evolution include cultural influences, which drive irregular word meaning shifts through phenomena such as new technologies (e.g., *cell* to mean *cell phone* in addition to *prison cell*) or community-specific vernaculars (e.g., *gay* to mean *homosexual* in addition to *carefree*). These cultural shifts often impact nouns more significantly. Conversely, regular linguistic processes, such as subjectification (e.g., *actually* shifting from objective to subjective usage) or grammaticalisation (e.g., *promise* undergoing rich changes), cause more predictable semantic changes and tend to affect verbs, adjectives, and adverbs more readily (Hamilton et al., 2016).

The Sinhala language is an Indo-European language, which possesses a rich and diverse literary heritage that has developed over the course of several millennia, with its origins tracing back to between the 3rd and 2nd centuries BCE (de Silva, 2025). This language has undergone significant evolution and transformation throughout its history, resulting in the form of modern Sinhala that we engage with today. Sinhala is spoken as L1 by approximately 16 million people, primarily located on the island of Sri Lanka (de Silva, 2025). The Sinhala script, which is unique to the language, descends from the Indian Brahmi script (Fernando, 1949; De Mel et al., 2025). Sinhala is classified as a lower-resourced language (Category 02) according to the criteria presented by Ranathunga and de Silva (2022).

In this study, we introduce a novel diachronic Sinhala dataset, SiDiaC[2], which covers the period from 426 CE to 1944 CE. This dataset is based on distinct identifications of written years or specific time frames of the recognised Sinhala literature.

2 Existing Work

The development of historical corpora has garnered significant attention due to its importance beyond the creation of general-purpose corpora. It enables researchers to investigate the evolution of language, taking into account changes in semantics, lexicon,

[1] **Latin:** The only good language is a dead language.

[2] https://github.com/NeviduJ/SiDiaC

morphology, and syntax. As a result, studies have been conducted to develop diachronic corpora for different languages.

2.1 LatinlSE

McGillivray and Kilgarriff (2013) introduced LatinlSE, a 13-million-word historical Latin corpus developed for the Sketch Engine[3], a leading corpus query tool. Covering an extensive 22-century period from the 2nd Century BCE to the 21st Century CE, LatinlSE is equipped with detailed metadata, including author, title, genre, era, date, and century.

The methodology for creating this corpus involved gathering texts from various online digital libraries, such as *LacusCurtius*[4], *IntraText*[5], and *Musisque Deoque*[6]. This process ensured a broad classification of genres as prose and poetry, and the texts were converted into a verticalized format while preserving their metadata. A significant aspect of the creation process was the automatic linguistic annotation using advanced NLP tools. This included lemmatisation with the PROIEL[7] project's morphological analyser, complemented by *Quick Latin*[8] for unrecognised forms. Part-of-Speech (POS) tagging was achieved by training TreeTagger (Schmid, 1999) on existing Latin treebanks, including the *Index Thomisticus Treebank*[9], the *Latin Dependency Treebank* (Bamman and Crane, 2006), and the PROIEL project's Latin treebank (Haug and Jøhndal, 2008). This training helps disambiguate analyses and assign the most likely lemma and POS to each token in context. This comprehensive dataset allows users to perform sophisticated searches based on lemmas, POS, and context, facilitating the study of shifts in word meanings over time.

2.2 IcePaHC and FarPaHC

The *Icelandic Parsed Historical Corpus* (IcePaHC) (Rögnvaldsson et al., 2012) is a one-million-word parsed historical corpus of Icelandic, spanning from the late 12th century to the early 21st century. But more relevant to our work in this study is the *Faroese Parsed Historical Corpus* (FarPaHC), a syntactically annotated corpus of Faroese historical texts, that is presented as a *spin-off* of IcePaHC. The reason for this relevance is that, according to Ranathunga and de Silva (2022), Faroese also belongs to Category 02, similar to Sinhala. The FarPaHC corpus has 53,000 words.

It's given that FarPaHC is an extension of IcePaHC; the primary sources included narrative and religious texts that have parallel texts in IcePaHC. A key step in the process was the conversion of all texts to modern spelling using the IceNLP package[10] (which includes a tokeniser, POS tagger, and lemmatiser), which was necessary for preprocessing and for facilitating searches. The annotation process involved manually dividing clauses, semi-automatically preprocessing texts with IceNLP and CorpusSearch[11] for partial annotations, and extensive manual parsing carried out by one annotator using a custom-developed visual tree editor, Annotald[12].

2.3 Other Historical Corpora

Pettersson and Borin (2019) provides a comprehensive survey on existing diachronic and historical corpora. The work by Keersmaekers and Van Hal (2024) presents a case study demonstrating how large-scale automated parsing of Greek papyri can create richly annotated diachronic resources. Chen and Liu (2025) have created a Chinese corpus from the last 30 years of news articles on land usage. Even the corpora in higher-resourced languages such as DIAKORP (Kučera et al., 2015) (Czech), ARCHER (Biber et al., 1994), COHA (Davies, 2012)[13] (English), DTA (Geyken et al., 2011) and GerManC (Scheible et al., 2011) (German) differ in size, balance, annotation depth, and access models. DIAKORP offers seven centuries of Czech texts, though it lacks linguistic annotation. ARCHER samples English registers across four centuries in 50-year intervals, while COHA spans two centuries of American English with lemmatisation and POS tagging. *Penn Parsed Corpora of Historical English* (Taylor and Kroch, 1994) (PPCHE) and SRCMF (Stein and Prévost, 2013) (Old French) are similar to FarPaHC in the sense that they, too, are manually annotated corpora which provide syntactic analyses suitable for structural studies.

[3] https://www.sketchengine.eu/
[4] https://penelope.uchicago.edu/Thayer/E/Roman/Texts/
[5] https://www.intratext.com/
[6] https://www.mqdq.it/
[7] https://www.hf.uio.no/ifikk/english/research/projects/proiel/
[8] https://www.quicklatin.com/
[9] https://itreebank.marginalia.it/
[10] https://sourceforge.net/projects/icenlp/
[11] https://corpussearch.sourceforge.net/
[12] https://github.com/Annotald/annotald
[13] https://www.english-corpora.org/coha/

PPCHE, in particular, has influenced corpora in other languages through its *Penn-Helsinki* annotation scheme, facilitating cross-linguistic comparison. Similarly, the PROIEL treebank family (Eckhoff et al., 2018) extends such comparisons to some Indo-European languages via aligned New Testament translations.

ReM (Klein and Dipper, 2016) (Middle High German), RIDGES (Odebrecht et al., 2017) (German-Science), and the *Swedish Culturomics Gigaword* corpus (Eide et al., 2016), offer layered annotation or harmonised spellings for OCR quality control. While all corpora use some metadata scheme to provide critical contextual information such as date, genre, region, and authorship, beyond that, the metadata coverage varies widely. However, it can be noted that TEI-based[14] metadata schemes are popular among European language corpora.

3 Methodology

In this section, we describe the methodology used to create this dataset from the ground up. The process involved careful attention to detail at every stage, from planning to the final presentation, ensuring that the data are valid and of high quality. The procedure included addressing copyright laws in Sri Lanka, acquiring data, extracting text, and performing post-processing and formatting of the data as shown in Figure 1.

3.1 Dataset Assembly

At first, we began acquiring Sinhala literature, including both fiction and non-fiction books, from the *Internet Archives*. However, the amount of data we were able to gather was quite limited. As a result, we decided to turn to the primary institution dedicated to Sinhala literature: the National Library (Natlib) of Sri Lanka[15], which has its own digital repository[16].

In the digital repository, we were able to organise all available content chronologically by issue date, allowing us to see publications printed dating back to 1800 CE. We carefully selected the book title, author name, identifier number, and collection name for each book from that point onward. This process required careful filtering, as most of the available content consisted of gazettes and police reports.

[14] Text Encoding Initiative (TEI) guidelines
[15] https://www.natlib.lk/
[16] https://diglib.natlib.lk/

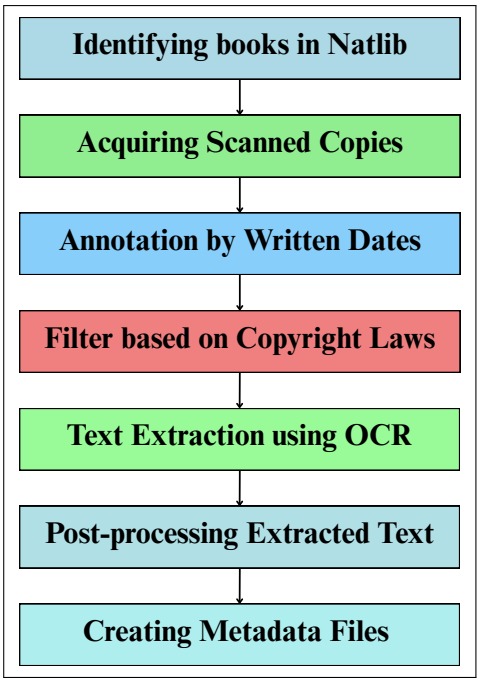


Figure 1: Summary of the Methodology Used in the Creation of SiDiaC.

We identified 233 unique books printed between 1800 CE and 1955 CE in the Natlib digital repository (this is based on the issued date, not the written date). Of these, only 12 books were available for open access; we had to request access to the remainder. Our initial plan was to collect 100 sentences per year. To achieve this, we estimated that obtaining five pages of text from each book, excluding content pages and the preface, would provide us with more than 100 sentences, which amounts to approximately 1500 to 2000 word tokens, assuming there are about 15 to 20 words per sentence. Therefore, for the closed-access books, we requested five pages from each one. The process of obtaining access to these books was very difficult because most of them were part of the Rare Books Collection.

3.2 Annotation by Written Date

The issue date of the identified books was clearly stated in the digital repository at Natlib. However, this does not imply that the books were actually written during those specified dates. In fact, a book could have been written centuries earlier, while the printed version was released much later.

Document dating has become extensively recognised in computational sociology and studies within digital humanities (Ren et al., 2023; Baledent et al., 2020; Hellwig, 2020). When compared to other dating tasks, historical text dating is more

complex due to the absence of explicit temporal indicators (such as time expressions) that aid in determining the date a document was written (Toner and Han, 2019; Baledent et al., 2020; Hellwig, 2020). It is clear that text dating, or the process of annotating the written date of a document, is an important task in diachronic studies (Ansari et al., 2023; Ren et al., 2023; Favaro et al., 2022).

Therefore, a comprehensive analysis was conducted to ensure that the written year of each book was accurately represented, ensuring that the resulting SiDiaC dataset accurately represents a proper diachronic corpus.

Upon the recommendation of experts in Sinhala linguistics, we identified a comprehensive book on Sinhala literature that claims to encompass literature information from its inception until 1994 CE (Sannasgala, 2015). This text served as the primary reference for the establishment of the respective date anchors, employing both time periods and specific years as outlined. The date ranges identified in Sannasgala (2015) correspond either to the period during which the book was authored or to the time period in which the author lived.

The process of determining the written dates for the books became more complicated because some books in the dataset include commentaries and discourses on earlier works. In this version of the dataset, these cases are tied to the original earlier book's written date, as they contain both the information from the original book and its corresponding commentary (often written centuries prior, with extensive sections given as direct quotes without paraphrasing).

3.3 Challenges from Copyright Laws

During the planning stage, one of the biggest challenges we faced was managing copyright issues. To address this, we conducted a thorough analysis of copyright laws in Sri Lanka, which are governed by the Intellectual Property Act No. 36 of 2003[17].

According to this act, copyright in Sri Lanka is generally protected for the life of the author, plus an additional 70 years after their death. In cases where the author is unknown, copyright protection lasts for 70 years from the date of first publication. As a result, we focused on literature where the author passed away before 1955, as well as works by unknown authors that were published before 1955.

3.4 Data Filtration

We initially identified 233 unique books, but after careful consideration of several factors, we ultimately selected only 46. Our selection process was influenced by the availability of scanned copies, the written dates of the works, and compliance with copyright laws as illustrated in Figure 2.

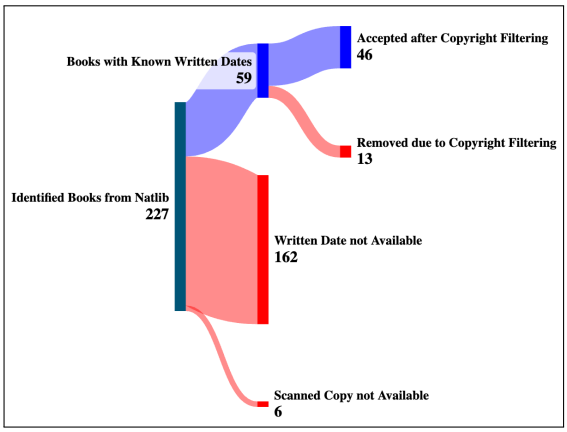


Figure 2: Sequential Data Filtration Procedure

Our first limitation was the availability of scanned copies at the digital repository of Natlib, which restricted us to 221 books. Additionally, we were only able to determine the written dates or periods for 65 of the 233 books. Taking both the availability of scanned copies and the accessibility of written dates into account, the number of selected records was reduced to 59. We then further refined this selection following the copyright law based filtering process we discussed, resulting in a final total of 46 books as presented in Table 3 in Appendix A.

3.5 Text Extraction using OCR

As the digital repository of Natlib shared the scanned copies of the requested books, we were forced to extract text information from the documents. Therefore, we selected an Optical Character Recognition (OCR) engine to get this task done.

In the comparative study conducted by Jayatilleke and de Silva (2025), five different OCR engines were analysed for text extraction using a synthetically created image-text dataset for Sinhala. Based on this study, we identified two standout OCR engines: Google Document AI[18] and

[17] https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

[18] https://cloud.google.com/document-ai/

Surya[19], with Surya being reported to outperform all other systems compared. However, during our text extraction process, we found that, under realistic conditions (unlike the synthetic conditions used in the study), Google Document AI provided more accurate results as shown in Appendix B.

Document AI is a service provided by Google Cloud Platform (GCP)[20]. In this platform, we created a processor and utilised its Application Programming Interface (API) key to conduct OCR. The processor can handle a maximum of 15 pages at a time; however, this was not an issue since all of our scanned copies contained 5 to 8 pages, as shown in Figure 3. Throughout the procedure, we ensured that we obtained the model confidence for every page of each processed document. We then calculated the average confidence score, which is included in the metadata file of each book folder.

This OCR processor has demonstrated that it can perform text recognition that goes beyond simple extraction. It adapts effectively and generates words in modern Sinhala spelling while also taking into account morphology, where morphemes are formed accordingly, as explained in the section 4.1.

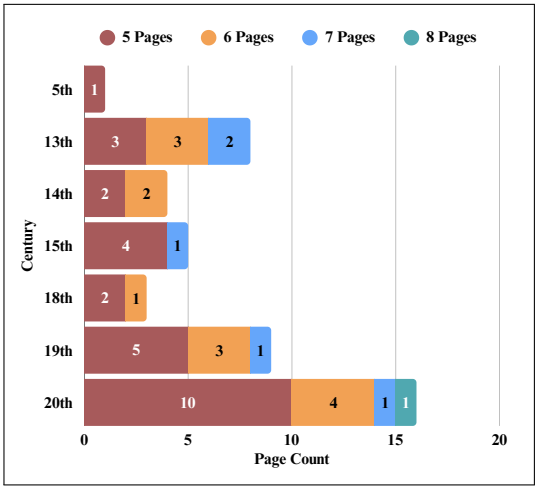


Figure 3: Distribution of Page Counts of Scanned Copies

3.6 Post-Processing Extracted Text

Although the OCR accuracy averaged 96.84% across all documents, the formatting issues were significant enough to require manual adjustments. The Document AI's advanced performance helped to streamline manual post-processing tasks, significantly reducing the time required for that work.

[19] https://github.com/VikParuchuri/surya
[20] https://cloud.google.com/

The post-processing includes correcting the following text formatting issues:

- **Spacing Errors**: This involves fixing incorrect or inconsistent spacing between words and sentences, such as missing spaces, uneven spaces and extra spaces.
- **Multi-column Text**: This refers to texts containing errors where two columns were treated as a single column, resulting in entire horizontal lines being extracted without properly traversing each column separately.
- **Misplaced Words/Phrases**: This addresses instances where words or phrases are out of order, leading to illogical text.
- **Paragraph and Line Indentation**: This involves standardising the indentation of paragraphs and individual lines. This could mean adding consistent indents to new paragraphs (e.g., poem blocks) and removing incorrect indents.
- **Removal of Seal Context**: This involves identifying and eliminating specific phrases or watermarks that represent seals or official stamps.
- **Page Number Removal**: This focuses on identifying and deleting page numbers that appear within the body of the text, as they are part of the document's structure rather than its content.

While language understanding was not a strict requirement for addressing these formatting-related factors, all manual post-processing procedures were carried out by the authors using a human-in-loop strategy (Lamba and Madhusudhan, 2023). This approach involved correcting formatting errors within a single window that contained both page scans and editable transcripts (Christy et al., 2017). The authors responsible for these corrections are native Sinhala speakers. The post-processing steps applied, along with examples, are further discussed and illustrated in Appendix C.

An important finding of this study was the presence of Pali, Sanskrit, and minimal English in certain records of SiDiaC. The inclusion of Pali and Sanskrit can be attributed to the fact that most historical texts are related to religion. Notably, both Pali and Sanskrit are written in the Sinhala script in all cases. In this study, we chose not to remove the content in these languages to avoid losing context.

3.7 Creation of Metadata Files

The dataset consisted of folders, each dedicated to a specific book. Within each folder, there is a text file along with a metadata file. The metadata files contain information such as the title and author names in both Sinhala and romanised forms, as well as the genre, issue date, written date, and the OCR confidence level for each particular book. Most of these information fields were identified through the LatinlSE corpus (McGillivray and Kilgarriff, 2013). Following the conventions of Davies (2012) and Rögnvaldsson et al. (2012), we maintain a consistent metadata annotation method throughout the corpus without changing it across the centuries. The overall composition of the SiDiaC corpus, including folder and file level examples, is illustrated in Figure 4.

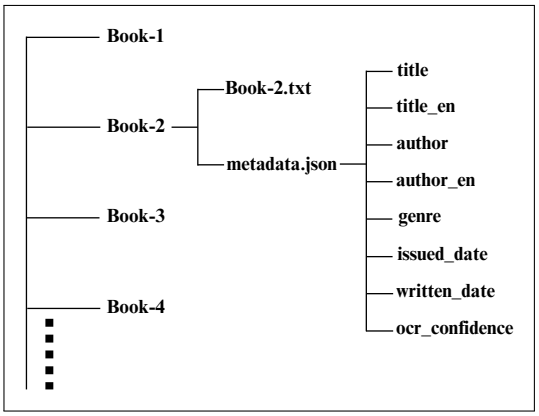


Figure 4: Composition of the SiDiaC Corpus

The title of each book was consistently provided. When known, the authors' names were included; if the authors were unknown, they were labelled as *unknown*. The issued date corresponds to the published year as listed in the digital repository of Natlib, while the written date was determined by referring to Sannasgala (2015), as explained in the section 3.2.

The genres of the books were selected based on the details provided by Sannasgala (2015), as well as the content evaluated by authors who are native Sinhala speakers. The classification process occurs at two levels. The primary level is broad and divides the books into two categories: 'Fiction' and 'Non-Fiction.' The secondary level is more specific, categorising the content of the books into five distinct classes: religious, history, poetry, language, and medical. This approach was inspired by the methodologies followed in IcePaHC and DIAKORP (Rögnvaldsson et al., 2012; Kučera et al., 2015) corpora to ensure diverse genres. It is important to note that the first level of categorisation was applied to all documents, while the second level of categorisation was applied only to the books that fell into the selected specific categories.

Furthermore, the average OCR confidence per page is included, as mentioned in the section 3.5. In addition, we romanised the titles and author names for each book, which involved transliterating Sinhala content into the Latin alphabet. An example of a metadata record is shown in Figure 5.


```
{
  "title": "අධිමාස දීපනය",
  "title_en": "Adhimasa Dheepanaya",
  "author": "මාදම්පේ ධම්මතිලක හිමි",
  "author_en": "Madhampe Dhammathilaka Himi",
  "genre": "Non-Fiction; Religious",
  "issued_date": "1896",
  "written_date": "1850 - 1896",
  "ocr_confidence": 0.9984
}
```


Figure 5: An example of a metadata record in SiDiaC

4 Analysis of SiDiaC

4.1 OCR Performance

As previously noted, the performance of Document AI extended beyond simple text extraction as shown in Figure 6. These improvements helped to streamline manual post-processing tasks, significantly reducing the time required for that work.

Two significant additional steps were clearly observable to be performed by Document AI.

1. **Text Modernisation**: The historical development of the Sinhala language has resulted in various eras of syntax being linguistically equivalent but different in grapheme representation (Nandasara and Mikami, 2016). It is clear that Document AI adjusts to generate text in modern Sinhala syntax, ensuring a consistent and unified syntax throughout the dataset. Given that the change is only at the grapheme level, this does not violate the syntactic or semantic properties of the word.
2. **Morpheme Segmentation**: In Historical Sinhala, certain words are combined without spaces, forming a closed compound (Gaikwad and Saini, 2024). This phenomenon was accurately identified by Document AI, which effectively performs morpheme segmentation.

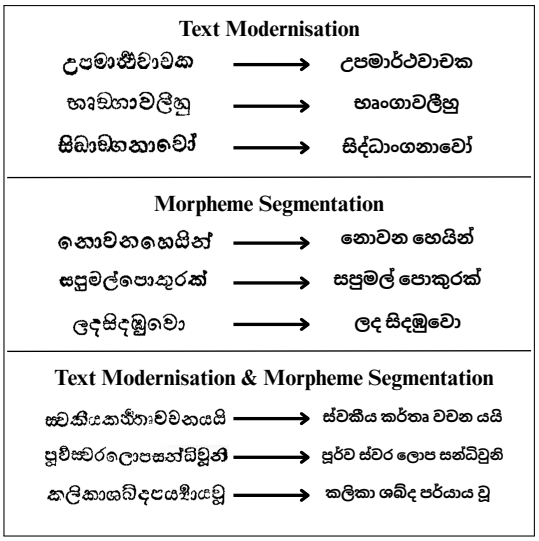


Figure 6: Examples of Sinhala Text Modernisation and Morpheme Segmentation in Document AI

Some characters in SiDiaC literature do not exist in the Sinhala Unicode. Therefore, mapping old characters to modern ones was an essential task performed by Document AI. Morpheme segmentation is crucial for maintaining consistency among words. If this issue is not addressed, combined multi-word expressions may be treated as unique words during the word embedding process, which can lead to significant differences in results when analysing semantic meaning.

4.2 Evaluation of Metadata

The dataset spans from the 5th to the 20th century CE, making it the longest continuous diachronic Sinhala corpus created to date. It covers many significant time periods, from the Anuradhapura era (377 BCE – 1017 CE) to just after Sri Lanka gained independence from Britain in 1948. This extensive timeframe allows for a representation of various changes in the language over the centuries.

Assuming that the books with specified date ranges are attributed to the upper bound year, an analysis of the number of books per century was conducted, as illustrated in Table 1. The analysis reveals that the distribution of books in the corpus is heavily skewed toward the 20th century, with 28 out of 46 records originating after the 18th century. This trend may largely be attributed to the introduction of the printing press to Sri Lanka by the Dutch in 1737, which thereafter popularised book printing in the country (Wickremasuriya, 1978; Nandasara and Mikami, 2016).

In the first level of genre classification between fiction and non-fiction, it is evident that there are more non-fiction books than fiction books in the corpus. At the second level of genre classification, religious texts and poetry dominate among the five categories. This predominance is largely due to the close relationship between Sinhala literary culture and Theravada Buddhism, which provided both subjects and a framework for preserving texts. Additionally, the influence of Sanskrit *kavya* traditions and courtly patronage, which valued literary artistry and prestige, also played a significant role (Hallisey, 2003).

The author of the book is known for 32 out of 46, while the remaining books are labelled as "Unknown." Only three authors have published more than one book in this dataset: two authors each have two books, and one author has four.

The OCR confidence levels are extremely high, with an average of 96.84% across all books and a minimum confidence score of 85.53%. Despite these encouraging figures, it is clear that Document AI encountered various types of errors, which we largely addressed during the post-processing phase as discussed in Appendix C. The accurate identification of characters and words likely contributes to these strong confidence scores; however, most errors appear to arise from the challenges presented by complex content formats.

4.3 Evaluation of the Corpus

SiDiaC consists of 58,027 word tokens that were filtered using regex, retaining only Sinhala and Latin characters, and subsequently tokenised by whitespace. The corpus contains 833 words in Latin script, which accounts for just 1.42% of the entire dataset. Also, the complete dataset comprises 22,837 unique word tokens in Sinhala script, which accounts for 39.36% unique word coverage of all words. This total word token count, while not in the range of millions, such as the COHA corpus for English, comfortably passes the 53,000 token count of FarPaHC for Faroese, which is in the same language resource category as Sinhala according to Ranathunga and de Silva (2022).

In the 5th century, 72.98% of words were unique, while the 13th century had 52.12% unique words. The 14th century saw 54.1%, the 15th century 55.97%, and the 18th century 55.05%. The 19th century featured 54.57%, but by the 20th century, the percentage dropped to 44.49%. As illustrated in Figure 7, generally a higher word count correlates with a lower percentage of unique words across the

	Primary Category		Secondary Category					Total
	Non-Fiction	Fiction	Religious	History	Poetry	Language	Medical	
5th	1	0	0	0	0	0	1	1
13th	7	1	5	0	1	2	0	8
14th	2	2	3	1	0	0	0	4
15th	1	4	2	0	3	0	0	5
18th	3	0	1	0	0	1	1	3
19th	6	3	2	2	3	2	0	9
20th	12	4	5	2	5	3	0	*16
Total	32	14	18	5	12	8	2	46

Table 1: Distribution of Books Across Centuries and Genres. *The total count for the secondary category in the 20th century amounts to 15, while the overall number of books is 16. This discrepancy arises because the book 'Hithopadhesha Sannaya', which offers advice, was not classified under any of the five secondary categories.

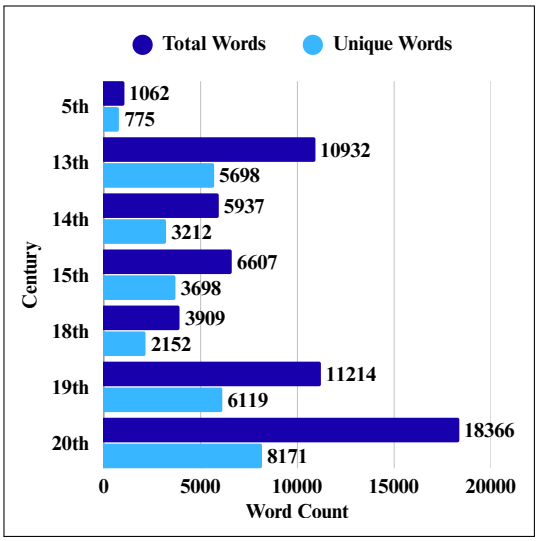


Figure 7: Total vs Unique Word Counts per Century

centuries.

Additionally, we conducted an analysis to identify the stopwords in the corpus by examining word tokens at the century level. Following the method described by Wijeratne and de Silva (2020) for their contemporary Sinhala corpus, we used a combination of word frequency analysis and manual vetting to identify appropriate stopwords from the corpora. To achieve this, we first counted the frequency of each unique word for each century and then converted these frequencies into z-scores.

$$z_{w,c} = \frac{f_{w,c} - \mu_c}{\sigma_c} \quad (1)$$

where, $z_{w,c}$ is the z-score of the word w in the century c. The term $f_{w,c}$ refers to the frequency of the word w during that century, μ_c denotes the mean frequency of all words in century c, and σ_c indicates the standard deviation of the frequencies of all words in that century.

Next, we calculated $-\infty < Z < 6.1027$ for a 99.80% threshold. In the 20th century, we observed the highest number of word tokens, which was used to establish the threshold, assuming that it provided adequate coverage of stop words throughout the entire corpus. After sorting the words in the 20th century by z-score in descending order, we manually inspected the words to determine the ideal z-score as the upper limit.

Over the centuries, the number of identified stopwords has varied accordingly. In the 5th century, there were 47 stopwords, followed by 42 in the 13th century, 39 in the 14th century, 28 in the 15th century, 44 in the 18th century, 65 in the 19th century, and finally, 61 in the 20th century.

The stopword that ranked highest across six of the seven centuries was ' වූ \ʊu: '(was, became), while ' හෝ \hɔ: '(or, either) ranked highest in the fifth century. Additionally, other frequently occurring stopwords include ' නම් \nʌm '(that, if), ' යන \ jʌnʌ '(that[named]), ' ඒ \e: '(that, those), ' මේ \me: '(this, these), ' ද \ðʌ '(or, interrogative particle), ' කොට \ kɒtʌ '(while), ' විසින් \ʊisin '(by), and ' ඇති \æθɪ '(be [exists]). However, it was observed that certain words, such as ' වහන්සේ \ʊahɛnse: '(honorific for a revered person) and ' ලක්ෂණය \ lʌkʃʌnʌjʌ '(Quality) were also receiving high z-scores, potentially due to the corpus's strong connection to Buddhist literature. The methodology for identifying the top 48 stopwords in the entire corpus and their depiction is presented in Appendix D.

5 Future Work

SiDiaC, being the first resource of its kind, paves the way for diachronic linguistic studies of the Sinhala language. We encourage researchers to explore areas such as lexical semantic change, tracking neologisms, grammatical change, historical language modelling, and corpus-based lexicography. Additionally, since the data is annotated by genre, it also allows for synchronic studies focusing on domain-based differences.

The corpus, as mentioned earlier, has comfortably surpassed the 53,000-token count of FarPaHC for Faroese, which falls into the same language resource category according to Ranathunga and de Silva (2022). However, the token count of SiDiaC, currently at 58,027, could certainly be increased by adding more literary works and additional pages from the existing 46 books. This enhancement would support more accurate research studies.

It is important to note that the OCR post-processing conducted in this study focuses only on formatting. However, as discussed in Appendix E, there are clear issues present at the word and character levels that need to be meticulously addressed. Furthermore, as mentioned earlier, the corpus is code-mixed with Pali, Sanskrit, and English. This highlights the need for a processing step to identify and remove irrelevant text, ensuring that the corpus is entirely focused on Sinhala.

We also identified that books called 'සන්න \sʌnnʌ' (meaning *commentaries*) may include two dates: one for the original (quoted) text and another for the commentary. However, this study did not consider this phenomenon, and such instances were attributed to the well-known original version. Therefore, in future studies, it would be beneficial to identify the two dates and include the text sections originating from the different time periods at their correct positions in the corpus.

The SiDiaC corpus did not undergo any lexical annotations during this study. Typically, the most recognised method for creating diachronic corpora involves parsing the entire corpus using POS tagging. However, this approach was not feasible with Sinhala POS taggers due to their limited performance. In future research, it would be highly beneficial to have the entire corpus parsed manually by Sinhala linguists who understand the evolution of language structure in Sinhala.

6 Conclusion

In this study, we introduced SiDiaC, a diachronic corpus of the Sinhala language. The corpus contains approximately 58k tokens, categorised into genres at two levels, and spans from the 5th century to the 20th century. This makes it the first diachronic Sinhala corpus ever created, which can serve as a foundational dataset for enhancing historical corpora in the Sinhala language. The entire process involved carefully identifying literature from the Natlib of Sri Lanka, followed by data filtering, date annotation, text extraction from PDF images, and post-processing. We also created metadata files containing important information about each book.

The complete corpus was thoroughly analysed, highlighting the powerful OCR performance of Document AI beyond simple text extraction. This was followed by a detailed evaluation of the dataset based on the metadata of all the books. Additionally, a comprehensive analysis was conducted at the word token level to ensure the identification of important findings within the corpus. Finally, we discussed potential future studies and approaches that could enhance the dataset, as well as the research opportunities that this corpus provides.

Limitations

The creation of the corpus went through different types of limitations due to various challenges we faced.

Literature Identification: While we recognised the *Department of National Archives*[21] of Sri Lanka also as a credible source, data acquisition was conducted only from the *National Library of Sri Lanka* due to permission constraints.

Data Filtration: Out of the 221 scanned copies acquired, we were able to identify the written dates or periods for only 59 of them. The written dates of the books were annotated based on the lifespans of well-known authors, while the majority of the remainder were annotated relying heavily on the work by Sannasgala (2015), which represents an over-reliance on a single source.

Post-Processing after OCR: Under this process, while corrections were initiated to address identified formatting issues, possible identification errors at the word or character level discussed in Appendix E were not addressed.

[21] https://websnew.lithium.lk/archives/

Code-Mixed Data: It was noted that the corpus contains code mixing of Pali, Sanskrit, and English languages, but the removal of text in these languages from the corpus has not been done.

Commentary Books: The identified books primarily named with the term ' සන්න \sʌnnʌ ' (meaning commentaries) will include two written dates for the original and the commentary. However, these instances were anchored to the well-known original version without removing the commentary.

Lexical Annotation: Unlike the LatinlSE, IcePAHC, COHA, and Google N-gram corpora, which have undergone lexical annotations specifically for POS tagging, we were unable to conduct similar annotations due to the unavailability of Sinhala POS taggers (de Silva, 2025).

Acknowledgments

The creation of the SiDiaC corpus was made possible through the valuable contributions of several individuals. We extend our sincere gratitude to Padma Bandaranayake, *Director of the National Library & Documentation Centre*, for her assistance with data acquisition. We also acknowledge Uthpala Nimanthi and Charani Palangasinghe for their efforts in the post-processing of the data, and the expertise of Nalaka Jayasena, a Sinhala Linguist, which was important in the identification of the book by Sannasgala (2015). Finally, we would like to thank Jayath de Silva, Savin Madapatha, and Thushan Bawantha for their dedicated work on the written date annotation.

References

Marjan Ansari, Bahram Hadian, and Vali Rezaei. 2023. Diachronic study of information structure in Persian. *Journal of Researches in Linguistics*, 15(2):65–76.

Anaëlle Baledent, Nicolas Hiebel, and Gaël Lejeune. 2020. Dating ancient texts: an approach for noisy French documents. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 17–21, Marseille, France. European Language Resources Association (ELRA).

David Bamman and Gregory Crane. 2006. The design and use of a latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. *Creating and using English language corpora*, pages 1–14.

Cheng Chen and Renping Liu. 2025. How administrative powers have impacted land-use development in China during the last 30 years: A diachronic corpus-based news values analysis. *Cities*, 159:105786.

Matthew Christy, Anshul Gupta, Elizabeth Grumbach, Laura Mandell, Richard Furuta, and Ricardo Gutierrez-Osuna. 2017. Mass digitization of early modern texts with optical character recognition. *Journal on Computing and Cultural Heritage (JOCCH)*, 11(1):1–25.

Michael C Corballis. 2017. The evolution of language.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Yomal De Mel, Kasun Wickramasinghe, Nisansa de Silva, and Surangika Ranathunga. 2025. Sinhala transliteration: A comparative analysis between rule-based and Seq2Seq approaches. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 166–173, Abu Dhabi. Association for Computational Linguistics.

Nisansa de Silva. 2025. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v25*.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Proceedings of the From Digitization to Knowledge workshop at DH*, pages 8–12.

Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. Towards the creation of a diachronic corpus for Italian: A case study on the GDLI quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–100, Marseille, France. European Language Resources Association.

P E E Fernando. 1949. Palaeographical Development of the Brahmi Script in Ceylon from 3rd Century BC to 7th Century AD. *University of Ceylon Review*, 7(4):282–301.

Hema Gaikwad and Jatinderkumar R. Saini. 2024. Identification of closed compound words in devanagari scripted and non-devanagari scripted corpora. In *Proceedings of Fifth Doctoral Symposium on Computational Intelligence*, pages 411–418, Singapore. Springer Nature Singapore.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. *Digitale Wissenschaft*, 157.

Charles Hallisey. 2003. Works and persons in sinhala literary culture. *Literary cultures in history: Reconstructions from South Asia*, pages 689–746.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Oliver Hellwig. 2020. Dating and stratifying a historical corpus with a Bayesian mixture model. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 1–9, Marseille, France. European Language Resources Association (ELRA).

Nevidu Jayatilleke and Nisansa de Silva. 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. *arXiv preprint arXiv:2507.18264*.

Alek Keersmaekers and Toon Van Hal. 2024. Creating a large-scale diachronic corpus resource: Automated parsing in the Greek papyri (and beyond). *Natural Language Engineering*, 30(5):1035–1064.

Thomas Klein and Stefanie Dipper. 2016. Handbuch zum Referenzkorpus Mittelhochdeutsch. Technical report, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut.

Karel Kučera, Anna Řehořková, and Martin Stluka. 2015. DIAKORP: diachronic corpus of Czech, version 6. *Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague*.

Manika Lamba and Margam Madhusudhan. 2023. Exploring ocr errors in full-text large documents: a study of lis theses and dissertations. *Library Philosophy and Practice (e-journal)*, 7824.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New methods in historical corpus linguistics*, 1(3):247–257.

S T Nandasara and Yoshiki Mikami. 2016. Bridging the digital divide in Sri Lanka: some challenges and opportunities in using Sinhala in ICT. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 8(1).

Carolin Odebrecht, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, 51(3):695–725.

Eva Pettersson and Lars Borin. 2019. Characteristics of diachronic and historical corpora. *Features to consider in a Swedish diachronic corpus.[online].[cit. 29. 1. 2022]. Dostupné z.*

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-aware language modeling for historical text dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656, Singapore. Association for Computational Linguistics.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association (ELRA).

Punchibandara Sannasgala. 2015. *Sinhala Sahithya Wanshaya*. S. Godage saha Sahodarayo.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA. Association for Computational Linguistics.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: the syntactic reference corpus of medieval french (srcmf). *New methods in historical corpora*, 3:275.

Ann Taylor and Anthony S Kroch. 1994. The pennhelsinki parsed corpus of middle english. *MS. University of Pennsylvania*, page 30.

Gregory Toner and Xiwu Han. 2019. *Language and chronology: text dating by machine learning*, volume 84. Brill.

Sarathchandra Wickremasuriya. 1978. The beginnings of the sinhalese printing press. In *Senarat Paranavitana commemoration volume*, pages 283–300. Brill.

Yudhanjaya Wijeratne and Nisansa de Silva. 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. *arXiv preprint arXiv:2007.07884*.

A Utilized Literary Works

During this study, we collected 46 literary works from the Natlib of Sri Lanka, including 32 by recognised authors and the rest by 'unknown' authors. Munidhasa Kumarathunga contributed four books, Madhampe Dhammathilaka Himi and Hikkaduwe Sri Sumangala Himi contributed two each, and the remaining authors each had one book, totalling 27 unique authors.

The metadata includes the title in Sinhala, the romanised title, the author's name, the romanised author's name, the genre, the issue date, the writing date, and the OCR confidence level.

The complete metadata for each literary work used in this compilation of the dataset can be found in Table 3 with the titles and authors' names presented in romanised Sinhala. This diachronic spread ensures coverage of Sinhala evolution across medieval, pre-modern, and modern stages. Religious texts dominate, reflecting both preservation biases and the centrality of Buddhism in Sinhala literary culture.

B Comparison of Document AI & Surya

The quantitative analysis conducted by Jayatilleke and de Silva (2025) indicates that Surya outperforms Document AI when evaluated on a synthetically created Sinhala dataset. However, during our text extraction process, we found that Document AI actually surpasses Surya. We believe this discrepancy stems from the synthetic data used in their study, which does not accurately reflect the challenges presented by real scanned documents.

Table 4 highlights three examples that clearly demonstrate why Document AI is the superior OCR engine. The errors indicated in red boxes for both systems demonstrate that Document AI excels in character identification, particularly with diacritics and similar-looking letters. Additionally, Document AI appears to be the only system effectively implementing morpheme segmentation, which is crucial for maintaining consistent word forms over time. Lastly, Document AI's text modernisation feature provides another significant advantage, making it the ideal choice for integration into the OCR pipeline used in this study.

C Post-Processing Extracted Text

During this phase, we addressed six types of formatting issues. This careful task was carried out by the authors of this study, who are native Sinhala speakers.

Certain literary works contained unwanted text referred to as seal context, which did not relate to the books' actual content. As a result, this text was identified and removed from the book files in the dataset. Some examples of these seal contexts can be seen in Table 2.

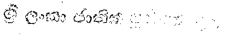
Seal Contexts	OCR Extractions
	ශ්‍රී ලංකා ජාතික ක
	ප්‍රමමිහි කෙළ මමහිසෙ
	වර ඇනි. වික්‍රමණි,
	BATAXABARA Y.C.T.
	RRalanazára

Table 2: Examples of seal contexts found in the PDF image documents and the corresponding extractions by Document AI that were removed during post-processing.

It was notable to see errors in spacing in poems, especially when the last word or letter of a word is separated by multiple spaces that were not correctly detected by Document AI, as shown in Table 5. Although correcting these errors may not hold significant semantic importance, accurately replicating the structure of the poems is crucial for preserving the original form of the books in case any downstream task using our dataset requires it. But for any task that does not require the original structure and would work on the lexical or semantic properties of the writing, the Document AI output is adequate. There were also multi-column texts, particularly evident in poetry books such as *'Kavya Wajrayudhaya - Palamu Kotasa'*. These texts contained errors where two columns were treated as a single column, resulting in entire horizontal lines being extracted without properly traversing each column, as shown in Table 6. To address these errors, corrections were made by mimicking the

Title	Author	Genre		Issued Date	Written Date	OCR Confidence↑
		Primary	Secondary			
Adhimasa Dheepanaya	Madhampe Dhammathilaka Himi	Non-Fiction	Religious	1896	1850 - 1896	0.9984
Adhimasa Winishchaya	Walikande Sri Sumangala Himi	Non-Fiction	Religious	1904	1850 - 1904	0.9971
Adhimasa Sangrahawa	Madhampe Dhammathilaka Himi	Non-Fiction	Religious	1903	1850 - 1903	0.9692
Anagathawanshaya: Methe Budu Siritha	Watadhdhara Medhanandha Himi; Siri Parakumabahu; Wilgammula Sangaraja Himi	Fiction	Religious	1934	1325 - 1333	0.9992
Ashoka Shilalipi saha Prathimakarana Winishchaya	D. E. Wickramasuriya	Non-Fiction	History	1919	1916	0.9989
Okandapala Sannaya hewath Balawathara Liyana Sanna	Don Andhris Silva	Non-Fiction	Language	1888	1760 - 1778	0.9884
Kavya Wajrayudhaya - Palamu Kotasa	Engalthina Kumari	Fiction	Poetry	1889	1825 - 1893	0.9254
Kavyashekaraya	Thotagamuwe Rahula Himi	Fiction	Poetry	1872	1408 - 1491	0.9813
Kudusika	Unknown	Non-Fiction	Poetry	1894	1270 - 1293	0.9988
Kusajathaka Wiwaranaya (Prathama Bagaya)	Munidhasa Kumarathunga	Non-Fiction	Religious	1932	1887 - 1932	0.9701
Gadaladeni Sannayai Prasidhdha wu Balawathare Purana Wyakyanaya	Hikkaduwe Sri Sumangala Himi	Non-Fiction	Language	1877	1827 - 1911	0.9970
Jubili Warnanawa	John de Silva	Non-Fiction	Language	1887	1857 - 1922	0.9957
Dhaham Sarana	Unknown	Fiction	Religious	1931	1220 - 1293	0.9891
Dhaladha Pujawaliya	Unknown	Non-Fiction	History	1893	1325 - 1333	0.9978
Dhurwadhi Hardhaya Widharanaya	Sri Dhanudhdharacharya	Non-Fiction	Religious	1899	1854 - 1899	0.9919
Dhampiya Atuwa Gatapadaya	D.B. Jayathilaka	Non-Fiction	Religious	1932	1868 - 1932	0.9241
Dharma Pradheepikawa hewath Mahabodhiwansha Parikathawa	Unknown	Non-Fiction	Religious	1906	1187 - 1225	0.9682
Dharmapradeepikawa	Gurulu Gomeen	Non-Fiction	Religious	1951	1187 - 1225	0.9786
Nikam Hakiyawa	Munidhasa Kumarathunga	Fiction	Poetry	1941	1887 - 1941	0.8932
Nikaya Sangrahaya hewath Shasanawatharaya	Unknown	Non-Fiction	Religious	1922	1390	0.9754
Nidhahase Manthraya	S Mahinda Himi	Non-Fiction	Poetry	1938	1901 - 1938	0.8997
Pansiya Panas Jathaka Potha	Unknown	Fiction	Religious	1881	1303 - 1333	0.9987
Parawi Sandheshaya	Unknown	Fiction	Poetry	1873	1430 - 1440	0.9902
Parani Gama	Galpatha Kemanandha Himi	Non-Fiction	History	1944	1944	0.9846
Budhdha Sikka hewath Kudu Sika	Unknown	Non-Fiction	History	1898	1270 - 1293	0.9699
Mage Malli	G. H Perera	Non-Fiction	Poetry	1938	1886 - 1938	0.8659
Mahawansa Teeka	Hikkaduwe Sri Sumangala Himi	Non-Fiction	History	1895	1827 - 1895	0.9978
Muwadew da Wiwaranaya	Munidhasa Kumarathunga	Non-Fiction	Religious	1949	1887 - 1944	0.8710
Moggalalayanawyakaranan	Moggallana Himi	Non-Fiction	Language	1890	1070 - 1232	0.9051
Moggallana Panchika Pradeepaya	Unknown	Non-Fiction	Language	1896	1070 - 1232	0.9918
Liyanora Nadagama	Unknown	Fiction	Poetry	1936	1852 - 1927	0.9935
Wibath Maldhama	Kirama Dhammarama Himi	Non-Fiction	Language	1906	1821	0.9986
Waidya Chinthamani Baishadhya Sangrahawa	Unknown	Non-Fiction	Medical	1909	1706 - 1739	0.9965
Wyakarana Wiwarana hewath Sinhala Bashawe Wyakaranaya	Munidhasa Kumarathunga	Non-Fiction	Language	1937	1887 - 1937	0.9029
Sadhdharma Rathnawaliya -Prathama Bagaya	Dharmasena Himi	Non-Fiction	Religious	1930	1220 - 1293	0.9962
Sanna sahitha Abhisambodhi Alankaraya	Waliwita Saranankara Sangaraja Himi	Non-Fiction	Religious	1897	1698 - 1778	0.9989
Sanna sahitha Salalihini Sandheshaya	Unknown	Fiction	Religious	1859	1450	0.9909
Sanskrutha Shabdhamalawa hewath Sanskrutha Nama Waranagilla	Rathmalane Dharmaloka Himi	Non-Fiction	Language	1876	1828 - 1887	0.9671
Sarartha Sangrahawa: Prathama Bhagaya	Srimadh Budhdhadhasa Rajathuma	Non-Fiction	Medical	1904	398 - 426	0.9997
Sithiyam sahitha Mahiyangana Warnanawa	Unknown	Fiction	Poetry	1898	1878	0.9989
Sithiyam sahitha Sadhdharmalankaraya	Unknown	Non-Fiction	Religious	1954	1398 - 1410	0.9810
Sithiyam sahitha Siyabas Maldhama	Kirama Dhammanandha Himi	Fiction	Poetry	1894	1820	0.9256
Sithiyam sahitha Sinhala Mahawanshaya	D.H.S Abhayarathna	Non-Fiction	History	1922	1874	0.9549
Sinhala Wyakaranaya enam Sidath Sangarawa	Hikkaduwe Sri Sumangala Himi	Non-Fiction	Language	1884	1827 - 1911	0.9886
Hansa Sandheshaya	C.E. Godakumbure	Fiction	Poetry	1953	1457 - 1465	0.8553
Hithopadhesha Sannaya	Waligama Sri Sumangala Himi	Non-Fiction	-	1884	1825 - 1905	0.9871

Table 3: The metadata information for all the literature used in the creation of this dataset.

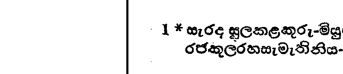
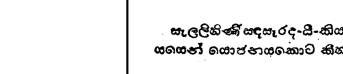
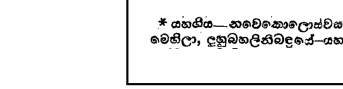
Image Example	Surya	Document AI
	1 * සැරද සුලකළකුරු-මියුරුහෙපලෙන් රදනා, රජකුලරහසැමැනිනිය-සියනිහිසැලලිහිණි සඳ.	1 * සැරද සුලකළකුරු - මියුරු හෙපලෙන් රදනා, රජකුල රහසැමැනිනිය - සියනිහි සැලලිහිණි සඳ.
	සැලලිහිණි සඳසැරද-යි-කියා ධනුර්ගුණසම්බන්ධනා න යගෙන් ගොස්පායනොට කීනයි දහසුතු.	සැලලිහිණි සඳ සැරද - යි - කියා ධනුර්ගුණ සම්බන්ධන්‍යා යයෙන් යොජනය කොට කීහයි දන යුතු.
	* යහගිය ... නවෙකොලොස්වසමු... බැගි නෙකොලොස්සෙ මෙහිලා, ලුහුබහලිනිබලුයේ-යහගිය (මේනිදසුන්) යි.	* යහගීය - නවෙකොලොස් වස්මි - බැගි නෙ කොලොස්සෙ මෙහිලා, ලුහු බහලිනි බඳුයේ - යහ ගිය (මේ නිදසුන්) යි.

Table 4: Examples of sentences along with their corresponding text extractions from `Surya` and `Document AI` for comparison. Note that the characters and phrases highlighted in red boxes contain errors. † This character is known as ‘ කුණ්ඩලිය \ kuːndʌlijʌ ’, a punctuation mark that indicates the end of a text or section in historical Sinhala.

original structure of the books to preserve their intended format.

The corrections provided in the table demonstrate that OCR outputs can be restructured faithfully only through column-aware preprocessing (for example, layout analysis, region detection, or image segmentation).

The misplaced words and phrases appeared mul-

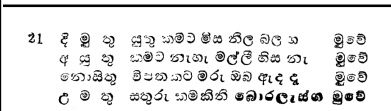
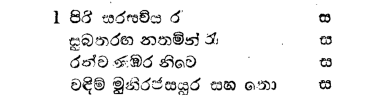
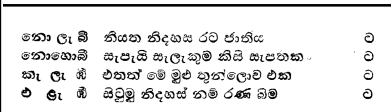
Image Example	OCR Extracted Text	Corrected Extracted Text
	21 දිමුතු යුතු කමට මිස නිල බල හ මුවේ අයුතුකමට නැහැ මල්ලි හිස නැ මුවේ නොසිතු විපතකට මරු ඔබ ඇද දැ මුවේ උ ම තු සතුරු කමකිනි බොරලැස්ග මුවේ *	21 දිමුතු යුතු කමට මිස නිල බල හ මුවේ අයුතුකමට නැහැ මල්ලි හිස නැ මුවේ නොසිතු විපතකට මරු ඔබ ඇද දැ මුවේ උමතු සතුරු කමකිනි බොරලැස්ග මුවේ
	1 පිරි සරසවිය ර ස සුබතරඟ නතමින් රැ ස රන්වණඟර නිවෙ ස වඳිමි මුනි රජසයුර සහ තො ස *	1 පිරි සරසවිය ර ස සුබතරඟ නතමින් රැ ස රන්වණඟර නිවෙ ස වඳිමි මුනි රජසයුර සහ තො ස
	නොලැබී නියත නිදහස රට ජාතිය ට නොහොබී සැපැයි සැලැකුම කිසි සැපතක ට කැලැඹී එතත් මේ මුළු තුන්ලොව එක ට ළැඹී සිටුමු නිදහස් නම් රණ බිම ට *	නොලැබී නියත නිදහස රට ජාතිය ට නොහොබී සැපැයි සැලැකුම කිසි සැපතක ට කැලැඹී එතත් මේ මුළු තුන්ලොව එක ට ළැඹී සිටුමු නිදහස් නම් රණ බිම ට

Table 5: Examples of spacing errors after OCR using `Document AI` on images and their corresponding corrections. *Note that the OCR extractions depicted were not exact; some final words were completely unidentified, which were added manually, and some had line breaks in awkward places.

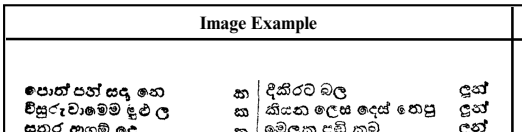
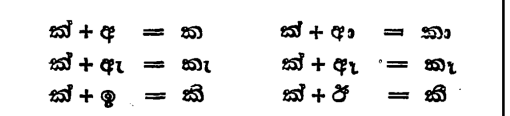
Image Example	OCR Extracted Text	Corrected Extracted Text
	පොත් පත් සදා නෙ ක දී කිරට බල ලුන් විසුරුවාමෙම මුළු ල ක කියන ලෙස දෙස් නෙපු ලුන් සතර ආගම් දෙ ක මෙලක පඬි කඩ ලුන් නසන්නට තැන්කරති නොවැසැ වැණුව හැටි බල විලස සලෙලුන් *	පොත් පත් සදා නෙ ක දී කිරට බල ලුන් විසුරුවාමෙම මුළු ල ක කියන ලෙස දෙස් නෙපු ලුන් සතර ආගම් දෙ ක මෙලක පඬි කඩ ලුන් නසන්නට තැන්කරති නොවැසැ වැණුව හැටි බල විලස සලෙලුන්
	ක් + අ = කා ක් + ආ = කා ක් + ඇ = කැ ක්+ ඈ = කෑ ක් + ඉ = කි ක් + ඊ = කී *	ක් + අ = කා ක් + ආ = කා ක් + ඇ = කැ ක් + ඈ = කෑ ක් + ඉ = කි ක් + ඊ = කී

Table 6: Examples of errors in multi-column text after OCR using `Document AI` on images and their corresponding corrections. *Note that the OCR extractions shown here are not exact, as we could not fully represent an entire page that experienced this type of error in real case scenarios.

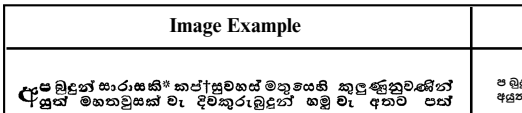
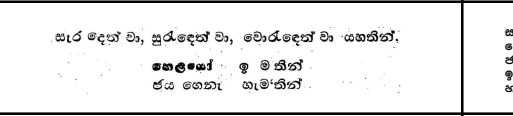
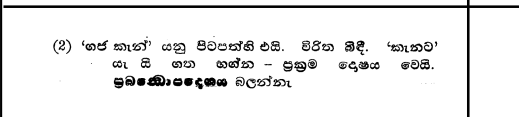
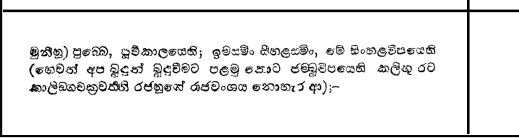
Image Example	OCR Extracted Text	Corrected Extracted Text
	ප බුදුන් සාරාසකි* කප්සුවහස් මතුයෙහි කුලුණු නුවණින් අයුත් මහත වූ සක් වැ දිවකුරු බුදුන් හමු වැ අතට පත්	අප බුදුන් සාරාසකි* කප්සුවහස් මතුයෙහි කුලුණු නුවණින් යුත් මහත වූ සක් වැ දිවකුරු බුදුන් හමු වැ අතට පත්
	සැර දෙන් වා, සුරැදෙන් වා, වොරැදෙන් වා යහනින්, * හෙළයෝ ඉමනින් ජය ගෙනැ ඉමනින් හැමතින්.	සැර දෙන් වා, සුරැදෙන් වා, වොරැදෙන් වා යහනින්, † හෙළයෝ ඉමනින් ජය ගෙනැ හැමතින්.
	(2) 'ගජ කැන්' යනු පිටපත්හි එසි. * විරිත බිඳී. 'කැනට' ප්‍රකුම දොෂය වෙසි. යැ සී ගත භග්න ප්‍රබන්ධොපදේශය බලන්නැ	(2) 'ගජ කැන්' යනු පිටපත්හි එසි. † යැ සී ගත භග්න - ප්‍රකුම දොෂය වෙසි. ප්‍රබන්ධොපදේශය බලන්නැ
	මුනිහු) පුබ්බේ, පූර්ව කාලයෙහි; ඉම සම් (හෙවත් අප බුදුන් බුදු වීමට පළමු කාලිංග චක්වර්ති රජහුගේ රාජ වංශය සිහළ ස්ම්, මේ සිංහළ ද්වීපයෙහි කොට ජම්බුද්වීපයෙහි කලිඟු රට නොහැර ආ);-	මුනිහු) පුබ්බේ, පූර්ව කාලයෙහි; ඉම සම් (හෙවත් අප බුදුන් බුදු වීමට පළමු කොට ජම්බුද්වීපයෙහි කලිඟු රට කාලිංග චක්වර්ති රජහුගේ රාජ වංශය නොහැර ආ);-

Table 7: Examples of misplaced words and phrases, along with †errors in paragraph and line indentation that had occurred after using `Document AI` for OCR on images. Additionally, the corrections for these errors are provided. *Note that the OCR extractions displayed were not precise, as the errors were shown together rather than individually. The other errors were corrected to highlight the specific error being focused on.

Stopword	Meaning [in Context]	5th	13th	14th	15th	18th	19th	20th
වූ \ʊu:	[that which came to] be	X	X	X	X	X	X	X
නම් \nʌm	that, if	X	X	X	X	X	X	X
යන \ jʌnʌ	that [named]	X	X	X	X	X	X	X
ඒ \e:	that, those		X	X		X	X	X
මේ \me:	this, these	X	X	X	X	X	X	X
ද \ðʌ	or, interrogative particle	X	X		X	X	X	X
කොට \kɒtʌ	while		X	X	X	X	X	X
විසින් \ʊisin	by	X	X	X		X	X	X
ඇති \æθɪ	be [exists]		X	X	X	X	X	X
ලද \lʌðʌ	that [which was]	X	X	X	X		X	X
වේ \ve:	be [affirmed]		X		X	X		X
හෝ \hɔ:	or	X	X					X
හා \hɑ:	and		X	X	X	X	X	X
න \nʌ	by		X				X	X
යනු \ jʌnʊ	is [so named]		X			X	X	X
ච \tʃʌ	with, by		X				X	X
ය \ jʌ	be [affirmed]		X				X	X
වහන්සේ \ʊahɐnse: *	Honourable [suffix]		X	X			X	
කළ \kʌlʌ	do [end of action]					X	X	X
හට \hʌtʌ	to	X				X		
හෙයින් \hɛjɪn	because		X	X		X		X
ලක්ෂණය \lʌkʃʌnʌjʌ *	Quality	X						
මහා \mʌhɑ: *	Great [Prefix]		X	X			X	X
බව \bʌvʌ	[the fact] that					X	X	X
යයි \ jʌjɪ	so [called]			X		X	X	X
යි \ jɪ	be [reported]		X	X				X
කියා \kɪjɑ:	is [reported]					X	X	X
කරණ \kʌrʌnʌ	by [means of]				X		X	X
කල්හි \kʌlhɪ	while, when		X	X		X		
සේ \se:	like [in manner of]		X					X
ස \sʌ	[poetic suffix]				X		X	
නි \θɪ	be [affirmed]		X				X	
යැ \ jæ	be [reported]		X					X
මහ \mʌhʌ *	Great [Prefix]		X	X				
ර \rʌ	[poetic suffix]				X			X
වන \vʌnʌ	be [known to exist]		X					X
නන්හි \θʌnhɪ	at, after [so declared]		X			X		
එක \ɛkʌ	a/the/one						X	X
තෙම \θɛmʌ	him	X				X		
නො \nɒ	not		X					X
මෙන් \mɛn	also [in similar manner]				X			X
එහි \ɛhɪ	therein			X				X
පිණිස \pɪnɪsʌ	for						X	
බුද්ධ \bʊððʰʌ *	Buddha			X		X		
ගුණ \gʊnʌ *	Quality				X			X
වැ \væ	[after] be		X					X
මෙහි \mɛhɪ	herein							X
කර \kʌrʌ	do [end of action]				X			X

Table 8: The top 48 stopwords and their presence in each century after applying the threshold $-\infty < Z < 6.1027$. *Note that some words, which are not technically stopwords, were included in this analysis. This may have occurred due to the limited availability of literary works in certain centuries and a bias toward Theravada Buddhism in the selected literature.

-tiple times due to varying spacing and text positioning styles used in different books. Unlike simple character substitutions, these errors dislocate words or phrases from their expected syntactic and semantic slots. For example, text fragments are merged across lines, while key markers or paragraph boundaries vanish. This was particularly noticeable at the beginning of the first paragraph in texts with drop caps. Similarly, in poetry, irregular line breaks caused by extra spaces before the last letter or word were among these issues. A misplaced suffix or dislocated phrase could invert rhetorical emphasis or obscure reference chains. Importantly, these errors also impact computational parsing, where algorithms expecting consistent lineation and phrase boundaries will misinterpret discourse structure. Another one of the most common issues addressed during post-processing was the indentation errors in paragraphs and lines. These errors were often found in the first line of paragraphs, with certain poems being misaligned, and page headings being centred incorrectly. A few examples of these issues are presented in Table 7.

`Document AI` also captured meta information on the physical book, such as page numbers. During the text extraction, these were removed from the books because they do not add any value to the presented corpus. These numbers were typically placed at the top or bottom, and they were centred or right-aligned in different literary works.

D Analysis of Stop Words in `SiDiaC`

The stopword analysis was conducted using the z-score calculation, as detailed in section 4.3. During this analysis, we identified a union set of words that exceeded the established threshold, resulting in a total of 194 unique words. The words were sorted by their average z-score for each century in descending order. The list of words that were above the set threshold was cross-checked with the union set, illustrating their availability in each century. The top 48 words with the highest mean z-scores are displayed in Table 8.

The continued inclusion of particles and suffixes shows continuity of core grammatical function words. Their presence across the 5th–20th centuries suggests remarkable diachronic stability in Sinhala morphosyntactic scaffolding. However, the table also reveals anomalies: certain lexical items not traditionally classified as stopwords appear in the stopword list. This distortion is likely due to corpus composition (religious texts with Buddhist themes dominate some centuries, inflating the relative frequency of doctrinal terms). Another observation is the persistence of poetic suffixes in older centuries, gradually tapering in modern texts.
This diachronic shift may point to a movement away from poetry and towards prose. The presence/absence patterns also reveal data sparsity in some centuries, as marked by gaps where certain stopwords do not appear due to limited surviving texts. To wit, it is as interesting (mayhap more) to note what is missing in the 20th century, given that it seems to count a majority of the candidate words among its stopwords. Note how the archaic forms of honorifics and some direct references to Buddhism have dropped out of the list.

In earlier centuries, the preponderance of monastic authorship and the dominance of canonical or exegetical works meant that words indexing reverence and religious entities were unavoidable high-frequency items. Their disappearance, or at least their reduced prominence, in the most recent century may be reflecting how the subjects covered in the text have shifted from esoteric religious communication to comparatively more secular discourse. Equally important is the fact that the persistence of other grammatical function words across all centuries stands in stark contrast to this attrition of religiously marked lexemes. This points to a kind of lexical stratification: the unmarked syntactic scaffolding of Sinhala remains stable over time (only being replaced by synonyms when they do), while the culturally bound vocabulary tied to ritual, doctrine, or honorific practice is more vulnerable to historical change.

E Word & Character Level Errors

During the post-processing of extracted text from scanned PDF files, we carefully conducted formatting level corrections as mentioned in section 3.6. However, these adjustments did not resolve all the necessary corrections at the word and character levels.

It became clear that character identification issues persisted throughout the documents. As illustrated in Table 9, diacritics in Sinhala posed significant challenges. Some characters were completely unrecognised, resulting in character deletions. Additionally, some identified diacritics were incorrect substitutions for different diacritics. There were also instances where a diacritic was erroneously

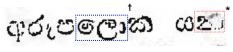
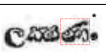
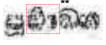
Image Example	OCR Extracted Text
	අරූපාලක යන්
	ලක්ෂණ.
ස න් න ස හි ත	සන්න සහිත ත
	කණිමුලෙ
	පූර්වග

Table 9: Examples of word deformation caused by character-level identification errors, including *incorrect identifications of diacritics/letters and **a complete deletion of a character at the marker. † Note that this is not an error, but rather a text modernisation step.

added even when no corresponding character existed, indicative of character insertion errors. In simpler terms, most of the errors in this category are related to spelling issues.

The Evolutionary Mechanisms of Transitivization in Mandarin VO Compounds: A Corpus-Driven Study of Competing Alternations

Menghan Jiang
Shenzhen MSU-BIT University
Shenzhen, China
menghan.jiang@connect.polyu.hk

Chu-ren Huang
The Hong Kong Polytechnic University
Hong Kong, China
churen.huang@polyu.edu.hk

Abstract

This study investigates the evolutionary mechanisms of transitivization in Mandarin verb-object (VO) compounds, addressing the core question of why this change diffuses at uneven rates across different lexical items. Based on a large-scale corpus analysis of 102 VO compounds, we reveal a significant statistical pattern: the transitivity frequency of a VO compound is positively correlated with the presence of a competing Verb-Complement (VC) construction ([VO1 Prep O2]) and negatively correlated with an Adverbial-Verb (AV) construction ([Prep O2 VO1]). To explain these findings, this study proposes a cost-based evolutionary framework. We argue that these correlations reflect different evolutionary pathways, embodying different evolutionary costs. The VC pathway represents a low-cost, direct route driven by grammaticalization, where the reanalysis of a semantically bleached preposition facilitates rapid diffusion. In contrast, the AV pathway is a high-cost, indirect route inhibited by a dual cost: its output violates the "Dependency Length Minimization" (DLM) principle and neutralizes the source construction's crucial information-structuring function. This framework provides a principled explanation for the observed statistical patterns, linking synchronic variation to diachronic mechanisms. It frames the uneven transitivization of VO compounds as a gradual lexical diffusion shaped by the competition between evolutionary pathways of differing cognitive and functional costs.

1 Introduction

The Verb-Object compound (hereafter 'VO compound') in Mandarin Chinese consists of two constituents with a syntactic/grammatical relation of a verb and its direct object (Li and Thompson, 1981), such as *touzi* invest-capital 'to invest' and *qianyue* sign-contract 'to sign a contract'. These compounds usually function as verbs. For instance, the compound *touzi* is formed from *tou* 'to invest' and *zi* 'capital', and the entire unit means 'to invest'. Previous studies (e.g., Huang, 1984; Li, 2012) proposed that because the verb in a VO compound already assigns case/theta roles to its object—such as *tou* 'invest' and *qian* 'sign' taking the objects *zi* 'capital' and *yue* 'contract'—the compound cannot take an additional direct object. However, recent studies have observed that, although their numbers are relatively small to begin with, an increasing number of VO compounds can now take another external object and yield the [VO1+O2] construction (e.g., Diao, 1998; Wang, 1997). Attested examples include *touzi fangdichan* invest-real estate 'invest in real estate'. In this construction, O1 refers to the compound's internal object (i.e. *zi* 'capital'), while O2 is the new, external object (i.e. *fangdichan* 'real estate').

Notably, while an increasing number of VO compounds are showing a tendency toward transitivization, the diffusion of this grammatical change is highly uneven across lexical items. For instance, some verbs have a very high frequency of taking an object, such as *guanxin taren* 'care about others', while the transitive usage of other verbs is less common, such as *pimei Aozhou de Huangjin hai'an* 'rival the Gold Coast of Australia'. Some transitive examples, like *guanguang Yidali* 'sightsee Italy', are almost exclusively found in specific registers like news headlines (Jiang and Huang, 2022, 2024).

Previous studies on the VO1+O2 construction have either focused on static descriptions of grammatical features or pointed out possible source structures (Rao, 1984; Xu, 1988; Liu, 1993; Liu, 1998; Liu and Li, 1998; Gao, 1998; Yang, 2001). However, a key 'explanatory gap' remains: few studies have yet been able to systematically explain the internal mechanisms behind the uneven diffusion of this grammatical change across different lexical items. This study aims to fill this gap through large-scale corpus analysis and theoretical framing to answer the core question: Why does the

transitivization of VO compounds spread at such uneven rates across different lexical items?

The central argument of this study is that the diffusion rate of transitivization for a given VO compound fundamentally depends on the syntactic and cognitive costs involved in the evolutionary pathway from its competing source constructions.

2 Theoretical Framework

This study integrates two theoretical perspectives: Lexical Diffusion Theory and Constructionalization Theory. Lexical Diffusion Theory (Wang, 1969, 1977, 1979) posits that language change proceeds gradually on an item-by-item basis and at different rates (Iyeiri, 2010; Yue-Hashimoto, 1993; Zhang, 2000; Cheng, 1990, 1998; Tottie, 1991; Nevalainen, 2006; Ogura, 1993). Synchronic variation observed at any given point in time is essentially a snapshot of a stage in the process of linguistic evolution (Wang, 1979). In language change, new and old forms coexist and compete for an extended period, with the new form eventually replacing the old one (Yue-Hashimoto, 1993).

Construction Grammar (CG) holds that the basic units of language are constructions, i.e., form-meaning pairings, which can cover linguistic phenomena at all levels from morphemes to complex sentence patterns (Goldberg, 2006). Constructionalization Theory, as a diachronic extension of Construction Grammar, specifically investigates the process by which new constructions emerge (Traugott and Trousdale, 2013, 2014).

Lexical Diffusion Theory and Constructionalization Theory are not mutually exclusive but complementary. The former describes the diffusion patterns and temporal trajectories of language change at the lexical level, while the latter explains the specific grammatical mechanisms of the change. In this integrated view, the emergence of the [VO1 O2] construction can be regarded as the emergence (constructionalization) of a new argument structure construction, with its origins in competing alternative forms. Through micro-steps such as reanalysis and analogy, [VO1 O2] gradually emerges from older constructions, completing the creation of a new one. Within this integrated framework, these synchronic competing alternations provide a window into the diffusion mechanisms of transitivization. Based on previous research and corpus observations, the main competing formats include (Zhang, 2010; Li and Wu, 2017):

(1) **Verb-complement construction with postverbal preposition (VC)**: This involves transforming structures like [VO1 + preposition + O2] into [VO1 O2]. For example, *zhili yu keyan* 'devote to research' evolves into *zhili keyan*.

(2) **Adverbial-Verb construction with preverbal preposition (AV)**: This involves reordering structures from [preposition + O2 + VO1] into [VO1 O2]. For example, *wei guojia jingji bamai* 'check for the country's economy' changes into *bamai guojia jingji*.

(3) **Separation construction (SEP)**: This involves condensing structures from [V + O2 (DE) + O1] into [VO1 O2]. For instance, *hao yibaiwan de zi* 'spend one million funds' condenses to *haozi yibaiwan*.

(4) **No alternative construction (NO)**: Some VO compounds lack alternative forms and directly appear as [VO1 O2], such as *huozeng yi ben shu* 'receive a book'.

The theoretical framework of this study is based on this integrated diffusion-constructionalization perspective: [VO1 O2], as a nascent argument structure construction, evolves from various source constructions (i.e., competing alternations) along different constructionalization pathways. The degree of syntactic restructuring and cognitive processing cost required for each pathway directly influences the differential diffusion rates of the construction across different VO compounds. In other words, the unevenness in the speed of transitivization diffusion among different VO compounds stems from the varying syntactic and cognitive costs of the constructionalization pathways they each undergo.

3 Methodology

This study investigates a set of 102 Mandarin VO compounds. This list was compiled by aggregating all compounds identified as capable of transitive usage in previous key studies, primarily Gao (1998) and Qian (2011).

The corpus data for this study were primarily drawn from the Annotated Chinese Gigaword corpus (Huang, 2009), which contains over 1.1 billion characters from formal news texts. To ensure a comprehensive identification of competing alternations, we also consulted several other large-scale corpora, such as the BCC Corpus and the CCL Corpus, to verify and supplement the annotation of possible patterns for each compound.

The annotation process involved two main stages. First, two experts, both native Mandarin speakers with linguistic training, manually annotated the presence or absence of the competing alternation patterns (VC, AV, SEP, NO) for each of the 102 compounds. A compound was marked as having a particular alternation if at least one clear instance of that pattern was found across the consulted corpora. Disagreements were resolved through discussion.

Second, to quantify the transitivity of each VO compound, we extracted a random sample of its occurrences from the Gigaword corpus. For each of the 102 VO compound, 1,000 tokens were extracted. We quantified the transitivity of each compound using its relative frequency. The transitivity frequency was calculated with the following formula: Transitivity Frequency = Number of transitive tokens/Total numbe of extracted tokens. For example, the transitivity frequency of *qianyue* sign-contract 'to sign a contract' in the corpus was calculated by dividing its number of transitive instances (13 tokens) by the total number of its extracted instances (1,000 tokens), yielding a frequency of 0.013.

The statistical analysis procedure included descriptive statistics to characterize the distribution of transitivity frequencies, Spearman rank correlation to assess the relationship between transitivity and alternation types, and the Kruskal-Wallis H test to compare group differences. Furthermore, a Beta regression analysis was conducted to validate the core findings, and a k-means clustering analysis was used to explore the typological patterns of the compounds.

4 Data Analysis

4.1 Descriptive Statistics

The analysis of 102 VO compounds shows that the mean transitivity frequency (0.303) is substantially higher than the median (0.182), and the distribution is highly right-skewed (see Figure 1). This indicates that the transitivization of most compounds is still in its early stages, a typical characteristic of uneven lexical diffusion.

In terms of alternation patterns, 74.5% of the compounds exhibit a VC format, and 76.5% exhibit an AV format, while fewer have SEP and NO formats (See Figure 2).

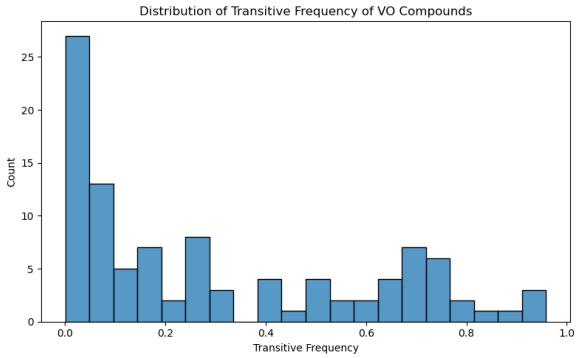


Figure 1: Distribution of Transitive Frequency of VO Compounds.

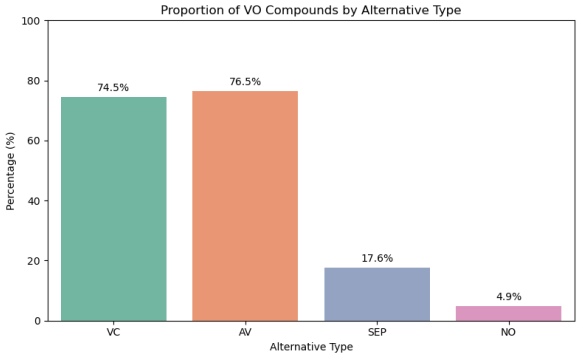


Figure 2: Proportion of VO Compounds by Alternative Type.

4.2 Core Relationship Findings

Spearman rank correlation analysis reveals a clear pattern of association (see Figure 3): VO compounds that possess a VC alternation exhibit significantly higher transitivity frequencies (= 0.233, p < 0.05). In contrast, those with an AV alternation show significantly lower transitivity frequencies (= –0.361, p < 0.001). There was no significant correlation between word frequency and transitivity frequency (p = 0.18), suggesting that the transitivization tendency of VO compounds is not driven by usage frequency but is more likely constrained by their internal syntactic features.

While the correlation analysis revealed general trends across the dataset, it does not isolate the effects of these competing alternations, especially since many compounds exhibit both VC and AV patterns. To more directly test our hypothesis of opposing evolutionary forces—a "pull" from the low-cost VC pathway and an "anchoring effect" from the high-cost AV pathway—we conducted a targeted group difference test.

For this purpose, we categorized the compounds into discrete groups based on their specific combi-

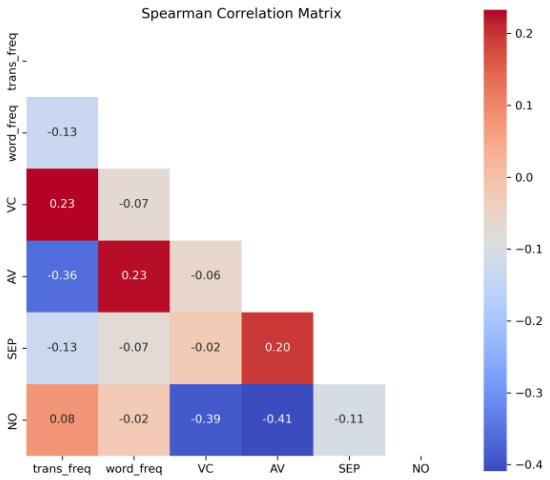


Figure 3: Proportion of VO Compounds by Alternative Type.

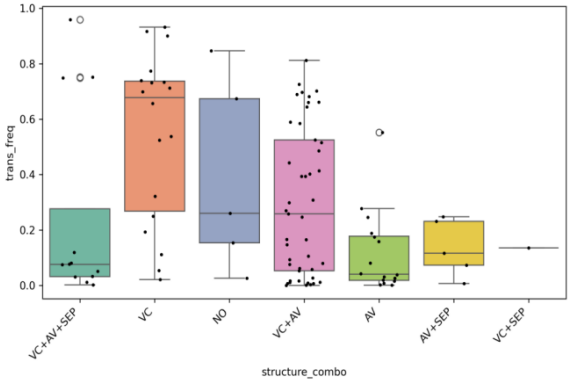


Figure 4: Proportion of VO Compounds by Alternative Type.

nation of alternation patterns (e.g., 'VC-only', 'AV-only', 'VC+AV', etc.). This grouping strategy allows for a crucial comparison: the 'VC-only' group represents a "pure" promoting pathway, while the 'AV-only' group represents a "pure" inhibitory one. If our cost-based framework is correct, we would expect to see the highest transitivity frequency in the 'VC-only' group and the lowest in the 'AV-only' group.

The Kruskal-Wallis H test confirmed a significant difference in transitivity frequency among these groups ($H = 19.27$, $p < 0.01$). As predicted, a post-hoc Dunn's test revealed that the median transitivity of the 'VC-only' group was indeed significantly higher than that of the 'AV-only' group ($p = 0.00094$), a result clearly visualized in Figure 4. This finding provides strong, direct support for the opposing roles of the VC and AV constructions in the transitivization process.

Synthesizing the above analyses, the core empirical finding of this study is: the Verb-Complement (VC) structure significantly promotes the transitivization of VO compounds, whereas the Adverbial-Verb (AV) structure significantly inhibits this process.

4.3 Supplementary Analysis

Subsequent statistical modeling provided further validation for these findings. We first conducted a Beta regression analysis, which is specifically designed for proportional data, to model transitivity frequency. The analysis confirms that the presence of an AV alternation is a significant negative predictor of a compound's transitivity frequency (= –1.041, $p = 0.039$). The model also indicated a positive trend for the VC alternation, which aligns with our hypothesis, although this effect did not reach the conventional level of statistical significance ($p = 0.117$). For completeness, a standard linear regression was also performed, which found the effects of both AV ($p < 0.001$) and VC ($p = 0.010$) to be significant.

Furthermore, exploratory clustering and classification analyses supported this general pattern. K-means clustering identified three types of VO compounds with different "structure-use" characteristics: a "VC-dominant" cluster showed a high tendency for transitivization, while an "AV-dominant" cluster showed the opposite, as shown in Table 1. A supplementary classification model analysis clarified the role of word frequency: while syntactic structure (VC/AV) is the core mechanism, word frequency itself can serve as an effective but non-mechanistic heuristic for distinguishing between high and low transitivization tendencies.

5 Discussion

This section aims to integrate the theoretical framework and statistical findings to propose a model that explains the divergent transitivity tendencies among VO compounds. The core idea of the model is that different competing alternation constructions represent different evolutionary starting points, and the pathways leading from these sources to the transitive construction [VO1 O2] involve varying syntactic and cognitive costs. These differences in cost ultimately determine the diffusion rate of the change.

Cluster	VC	AV	SEP	NO	T. Freq	W. Freq
0	0.6441	0.9831	0.2881	0.00	0.1213	27.9192
1	0.0000	0.0000	0.0000	1.00	0.3922	32.4420
2	1.0000	0.5263	0.0263	0.00	0.5734	18.5489

Table 1: K-means Cluster Centroids. T. Freq and W. Freq refer to transitivity frequency and word frequency, respectively.

5.1 Direct pathway: Low-cost grammaticalization based on the [VO1 [Prep O2]] construction

VO compounds with VC alternation patterns exhibit the highest degree of transitivity because they follow a low-cost grammaticalization pathway, with the core mechanism being a reanalysis of constituent boundaries driven by high-frequency usage. This pathway has a low cognitive cost, thus a high diffusion rate. The evolution begins with a syntactic structure containing a post-verbal prepositional phrase, [VO1 Prep O2], as in "*zhili* (VO1) *yu* (Prep) *keyan* (the external object O2)" ('devote to research'). The post-verbal preposition (such as *yu*) undergoes semantic bleaching and phonological weakening, its function gradually shifting from an independent contentful preposition to a functional word marking the object (Zhang, 2010).

As the preposition becomes grammaticalized, speakers cognitively reinterpret the syntactic structure: the original [VO1[Prep O2]] structure is reanalyzed as [VO1 O2], with the preposition such as *yu* being omitted. This step represents a typical micro-change described in constructionalization theory (Traugott and Trousdale, 2013, 2014).

Since this evolution only involves deleting a semantically bleached element and does not alter the core word order, it aligns with the principle of linguistic economy and prosodic requirements, thus being low-cost. The low-cost pathway greatly facilitates the diffusion of the new transitive construction. This also explains why the VO compounds in the VC group statistically exhibit the highest transitivity frequency.

5.2 Indirect pathway: Dual high cost based on the [Prep O2 VO1] construction

VO compounds with AV alternation patterns show a significantly lower degree of transitivity because their evolutionary pathway is inhibited by a dual high cost. The transformation from [Prep O2 VO1] (e.g., *wei guojia jingji bamai* 'feel the pulse for the national economy') to [VO1 O2] (*bamai guojia jingji*) must overcome two major obstacles: (1) According to the DLM principle, language users tend to minimize the linear distance between syntactically dependent elements to reduce working memory load (Gibson, 1998, 2000; Temperley, 2007). This principle is supported by large-scale, cross-linguistic corpus evidence showing that dependency lengths in natural languages are universally shorter than would be expected by chance (Futrell et al., 2015). The transitive construction evolved from the AV structure has a high cost under this measure because it forces a non-local dependency: the logical object of the core verb (V) is separated from it by its internal object (O1), lengthening their dependency distance and thus increasing the cognitive processing load for the listener/reader. (2) Functional cost: The source construction [Prep O2 VO1] is not pragmatically neutral in Chinese; it is a specialized information structure construction for topic or focus fronting. The shift to [VO1 O2] abandons this specialized pragmatic function, constituting a 'functional cost'.

This dual high cost, composed of both processing and functional pressures, forms a strong 'evolutionary resistance' that suppresses the occurrence and diffusion of the change.

5.3 Hybrid pathway: Dynamic competition between low-cost and high-cost

When a VO compound possesses both VC and AV alternation patterns, its degree of transitivity tends to be intermediate. This is because such compounds are simultaneously influenced by two evolutionary pathways with divergent costs: the low-cost VC pathway provides a "pull" that promotes transitivization, while the high-cost but functionally entrenched AV pathway provides an "anchoring effect" that slows the diffusion of the change.

5.4 Lexicalization pathway

For VO compounds without any prepositional alternation patterns (i.e., the NO group), especially those exhibiting strong transitivity, their transitivity arises not from a syntactic pathway but through

lexicalization. This pathway mainly applies to two types of VO compounds: (1) O1 is a verbal morpheme: e.g., *huozeng* 'receive-gift', where the internal object O1 is itself a transitive verbal morpheme. The transitivity of O1 is thus transferred to the entire compound, naturally equipping it with the ability to govern an external object. (2) O1 is a metaphorical or bleached morpheme: e.g., *chuxi* appear-seat 'attend' *jieshou* take-hand 'take over'), where the meaning of the internal object O1 is highly metaphorical or bleached. This leads to a high degree of semantic integration for the entire VO compound, causing it to be processed by users as an indivisible, single transitive verb [V-O]. Once VO1 is reanalyzed as a single verb at the lexical level, it can directly govern the object O2 without needing an intermediate prepositional construction phase. This is a form of lexical constructionalization, rather than a syntactic one.

6 Conclusion

Through systematic statistical analysis of a large-scale corpus, this study demonstrates that the transitivization tendency of modern Mandarin VO compounds is not random but is closely related to the types of competing alternation constructions these compounds can participate in. Specifically, the presence of a post-verbal preposition (VC) construction is a strong promoting factor for the transitivization of VO compounds, while the presence of a pre-verbal preposition (AV) construction significantly inhibits it.

To explain this core finding, this study proposes a model containing four evolutionary pathways. This model posits that the transitivization process of VO compounds is in fact the emergence of the new [VO1 O2] construction evolving from its respective competing source constructions through different constructionalization pathways. The syntactic and cognitive costs involved in each pathway determine the ease or difficulty of the change, thereby affecting the rate at which transitivization diffuses across different lexical items. It is these differences that ultimately lead to the statistical patterns and typological clustering patterns observed in the corpus.

Future research can be further deepened in the following aspects: (1) Diachronic analysis: Utilize diachronic corpora to trace the emergence and developmental trajectories of different evolutionary pathways, directly testing the evolutionary model proposed in this study and verifying the evolutionary processes and relative speed differences of each pathway from a historical perspective. (2) Psycholinguistic experiments: Directly measure the cognitive processing cost differences of various alternation constructions through sentence acceptability judgments, eye-tracking, or ERP experiments, to provide evidence for the psychological reality of the hypothesis in this model that the level of 'cost' affects the difficulty of evolution. (3) Comparison with other syntactic structures: Investigate whether similar evolutionary pathways and cost-constraint mechanisms exist in other similar compound structures to test the explanatory power of the model. (4) Multi-factor model: Future research could integrate the structural factors found in this study with other known factors into a more comprehensive multi-factor model. This would allow for exploring the interaction of various factors in the transitivization process of VO compounds, thereby more accurately describing and predicting their transitivization trends.

Acknowledgments

This work is supported by the Guangdong Philosophy and Social Science Foundation 2025, titled "A Comparative Study of Syntactic Choice in Cross-Strait Mandarin Based on a Large-Scale Comparable Corpus" (Project No. GD25YZY08), and by the Humanities and Social Science Youth Foundation of Ministry of Education of China, " Research on the Optimization and Evaluation of LLM-Generated Resources for International Chinese Language Education" (Grant No. 25YJC740020).

References

Y. B. Diao. 1998. Ye tan "dongbin shi dongci + binyu" xingshi. *Yuwen Jianshe*, 6:39–41.

R. Futrell, K. Mahowald, and E. Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

G. S. Gao. 1998. "dongbin shi dongci + binyu" de dapei guilü. *Yuwen Jianshe*, 6:36–38.

E. Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

E. Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain*, pages 95–126. MIT Press, Cambridge, MA.

C. R. Huang. 2009. Tagged chinese gigaword version 2.0.

J. C.-T. Huang. 1984. Phrase structure, lexical integrity, and chinese compounds. *Journal of the Chinese Language Teachers Association*, 19(2):53–78.

Y. Iyeiri. 2010. *Verbs of Implicit Negation and their Complements in the History of English*. John Benjamins Publishing, Amsterdam, Netherlands.

M. Jiang and C. R. Huang. 2024. Potential objects and transitivity variations: A comparable corpus-driven study of mandarin chinese verb-object compounds. *Lingua*, 311:103814.

M. H. Jiang and C. R. Huang. 2022. Hanyu dongbin fuhe ci de jiwu xing ji qi yongfa chayi——jiyu yuliaoku qudong fangfa de duibi yanjiu. *Zhongguo Yuwen*, 1:39–47.

Audrey Y. H. Li. 2012. *Order and constituency in Mandarin Chinese*. Springer Science & Business Media, New York, NY.

C. N. Li and S. A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. University of California Press, Berkeley, CA.

Y. Z. Li and Y. C. Wu. 2017. On the evolutionary mechanism of disyllabic transitive verbs in chinese. *Journal of Chinese Linguistics*.

D. W. Liu. 1998. Guanyu dongbin dai bin xianxiang de yixie sikao (shang). *Yuwen Jianshe*, 1:22–26.

Y. Liu and J. X. Li. 1998. "dongbin shi dongci + binyu" de bianhuan xingshi ji binyu de yuyi leixing. *Journal of Jianghan University*, 5:11.

Y. J. Liu. 1993. Dongbin shi dongci yu suo dai binyu zhi jian de yuyi guanxi. *Hanyu Xuexi*, 4:48–53.

T. Nevalainen. 2006. Syntactic structures. In *Introduction to Early Modern English*, pages 103–117. Edinburgh University Press, Edinburgh, UK.

M. Ogura. 1993. The development of periphrastic do in english: A case of lexical diffusion in syntax. *Diachronica*, 10(1):51–85.

C. Y. Qian. 2011. Xiandai hanyu "dongbin shi fuhe ci dai binyu" jiegou fenxi. Master's thesis, Fudan University.

C. R. Rao. 1984. Dongbin zuhe dai binyu. *Zhongguo Yuwen*, 6:413–418.

D. Temperley. 2007. Minimization of dependency length in written english. *Cognition*, 105(2):300–333.

G. Tottie. 1991. Lexical diffusion in syntactic change: frequency as a determinant of linguistic conservatism in the. In *Historical English syntax*, volume 2, page 439.

E. C. Traugott and G. Trousdale. 2013. *Constructionalization and constructional changes*. Oxford University Press, Oxford, UK.

E. C. Traugott and G. Trousdale. 2014. Contentful constructionalization. *Journal of Historical Linguistics*, 4(2):256–283.

H. D. Wang. 1997. "dongbin shi dongci + binyu" guilü hezai? *Yuwen Jianshe*, 8:30–31.

W. S.-Y. Wang. 1969. Competing changes as a cause of residue. *Language*, 45(1):9–25.

W. S.-Y. Wang, editor. 1977. *The lexicon in phonological change*. Mouton de Gruyter, Berlin, Germany.

W. S.-Y. Wang. 1979. Language change: A lexical perspective. *Annual Review of Anthropology*, 8(1):353–371.

D. N. Xu. 1988. Shuang bin tong zhi yu shuang bin yi zhi. *Yuyan Jiaoxue yu Yanjiu*, 2:35–45.

H. M. Yang. 2001. "vo+n" yu yuyi, jiegou de jianrong yu chongtu——hanyu dongbin zuhe dai binyu jiegou zhong de yuyi wenti. *Hanyu Xuexi*, 1:28–34.

A. Yue-Hashimoto. 1993. The lexicon in syntactic change: Lexical diffusion in chinese syntax. *Journal of Chinese Linguistics*, pages 213–254.

M. Zhang. 2000. Syntactic change in southeastern mandarin: How does geographical distribution reveal a history of diffusion? In *In Memory of Professor Li Fang-Kuei: Essays of Linguistic Change and the Chinese Dialects*, pages 197–242. Academia Sinica and University of Washington, Taipei, Taiwan and Seattle, WA.

Y. S. Zhang. 2010. Cong cuopei dao tuoluo: Fu zhui "yu" de ling xing hua houguo yu xingrongci, dongci de ji wu hua. *Zhongguo Yuwen*, 2:135–145.

L. W. Zheng and R. L. Cheng. 1990. Cihui kuosan lilun zai jufa bianhua li de yingyong——jian tan taiwan guanhua "you" zi ju de jufa bianhua. *Yuyan Jiaoxue yu Yanjiu*, 1:66–73.

HisGraphRAG: GraphRAG for Vietnamese Historical question answering

Hoang-Thanh Nguyen[1,2], Tung Le[1,2], Huy Tien Nguyen[1,2,*]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
23C11048@student.hcmus.edu.vn, {lttung, ntienhuy}@fit.hcmus.edu.vn
[*]Corresponding author: Huy Tien Nguyen – ntienhuy@fit.hcmus.edu.vn

Abstract

Large Language Models (LLMs) often face issues with factual accuracy and hallucinations, especially in specific domains like history. While Retrieval-Augmented Generation (RAG) and Graph-based RAG (GraphRAG) improve reasoning by integrating external knowledge, applying GraphRAG to historical data presents three key challenges: (1) redundant entities in the knowledge graph; (2) a lack of temporal understanding for historical events; and (3) potential for LLM generative capacity to be diluted by excessive retrieved information. We propose HisGraphRAG, a novel framework designed for historical question answering. HisGraphRAG addresses these issues through entity alignment during graph construction, temporal reasoning and reranking for event contextualization, and filtering of irrelevant information to maintain LLM focus. Evaluated on a Vietnamese history university entrance exam dataset, HisGraphRAG significantly boosts reasoning performance across various LLMs. This work offers a more reliable and effective GraphRAG framework for historical inquiry, enhancing LLM applications in this crucial field.

1 Introduction

Multiple-choice question answering (MCQA) in the history domain poses significant challenges for large language models (LLMs), which often struggle with precise factual recall and chronological reasoning (Brown et al., 2020). This can lead to *hallucination*, the generation of factually incorrect information (Ji et al., 2023), a problem especially prevalent in less-documented historical contexts like Vietnamese history (Kumar and Pavlick, 2022).

Retrieval-Augmented Generation (RAG) mitigates this by supplementing LLMs with external documents, improving factual grounding (Lewis et al., 2020a). However, standard RAG struggles with the complex inter-entity relationships and temporal structures (Zhou et al., 2022) inherent in historical narratives (Lewis et al., 2020b). For instance, answering a question like, "During World War II, what was the impact of 1943 American aid to other countries?" requires reasoning about connections not explicitly present in the retrieved text. To address this, GraphRAG builds structured knowledge graphs into the RAG pipeline, modeling entities and their relationships (Microsoft Research, 2024). While this enhances context, it introduces new limitations in historical domains. First, despite overall improvements, GraphRAG can paradoxically perform worse than a naive LLM in certain cases, as shown in Figure 1. Second, it often creates *duplicated entities* (e.g., "Soviet Union" and "Soviet") during indexing, fragmenting the knowledge graph as detailed in Table 1. Third, it fails to preserve *temporal grounding*, which is crucial for answering time-sensitive questions.

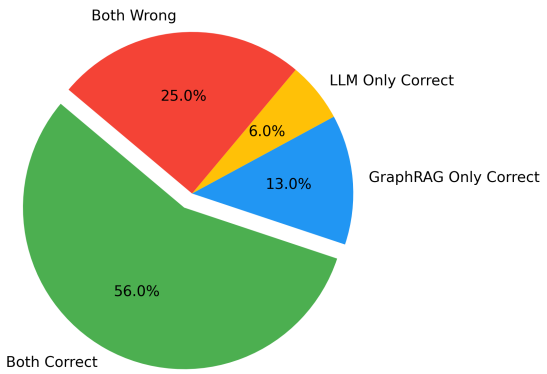


Figure 1: Accuracy Comparison: GraphRAG vs LLM-Only (Vietnamese University Entrance Exam)

To overcome these limitations, we propose HisGraphRAG, a novel framework with three targeted enhancements. First, we implement *entity alignment* during indexing to merge semantically equivalent nodes, creating a cleaner graph. Second, we

	Number of nodes	Number of relations
W/o Entity alignment	1334	681
Entity alignment	1186	510
% reduction	11.09%	25.11%

Table 1: Number of duplicated entity (Building on the summary version of grade 12 Vietnamese history book)

introduce *temporal retrieval*, using time expressions in the query to fetch chronologically relevant events. Finally, inspired by human test-taking strategies, we add an *answer candidate filtering mechanism* to eliminate implausible options before retrieval, sharpening the reasoning focus. Our main contributions are:

- Identify and address three critical limitations of GraphRAG in historical QA: duplicated entities, loss of temporal context, and distraction from irrelevant information.
- Introduce HisGraphRAG, a system that integrates entity alignment, temporal retrieval, and question candidates filtering to improve performance on Vietnamese historical QA. Then we demonstrate that our model's selective reasoning strategies effectively enhance accuracy on university entrance exam datasets.

2 Related work

Traditional RAG methods primarily rely on retrieving relevant text passages based on semantic similarity, but they struggle to capture complex relational and temporal knowledge inherent in domains like history (Procko and Ochoa, 2024). This limitation motivated the development of Graph Retrieval-Augmented Generation (GraphRAG) (Microsoft Research, 2024), which integrates structured knowledge graphs to provide richer, relational context for LLMs. GraphRAG retrieves graph elements such as nodes, triples, or subgraphs relevant to a query, enabling multi-hop reasoning over interconnected entities and relations.

GraphRAG is designed to minimize hallucinations and improve knowledge updates in LLM by integrating information from external graph knowledge (Procko and Ochoa, 2024; Han et al., 2024). G-Retriever (He et al., 2024) uses embedding cosine similarity to identify relevant nodes and edges for the query, then constructs the subgraph using the Prize-Collecting Steiner Tree (PCST) algorithm, which generates an accurate and compact subgraph for generation (He et al., 2024; Yiqian Huang, 2025). The RoG method (Jiang et al., 2024) proposes a "planning–retrieval–reasoning" model, in which reasoning paths are taken according to a pre-determined plan before performing inference. Meanwhile, GNN-RAG (Mavromatis and Karypis, 2024) utilizes Graph Neural Networks to handle the complex structure in the knowledge graph, combining retrieval augmentation techniques to increase the diversity of the input data. However, the performance of these methods depends deeply on the quality of the graph, and when the resulting subgraph contains a lot of noise or is irrelevant, the performance degrades significantly (He et al., 2024).

Despite these advances, GraphRAG faces several challenges. First, retrieving noisy and irrelevant information will downgrade LLM performence. It's researched and improved in Guo and al. (2025) research by adding two stages filtering in querying process. However, this improvement comes at the cost of increased response time in generating the final answer. Our entity alignment method will reduce duplicated or misaligned entities durign indexing step which reduces the knowledge graph size while remaining retrieval effectiveness. Second, temporal information and relationships between historical events are often overlooked, limiting the model's ability to reason about event sequences and causality. Third, excessive reliance on external graph knowledge can suppress the intrinsic reasoning capacity of the base LLM, then degrading overall performance.

3 Method

To address the key challenges identified in our preliminary studies—namely entity duplication, irrelevant retrievals, and lack of temporal reasoning—we propose a new framework **HisGraphRAG** designed specifically for historical question answering with GraphRAG. An overview of our framework is shown in Figure 2, containing three central principles:

- **Entity Alignment:** Redundant or variant entities (e.g., "President Ho Chi Minh" vs. "Ho Chi Minh") dilute knowledge and compromise reasoning. Aligning these into unified representations is essential to preserve semantic coherence and reduce graph fragment.

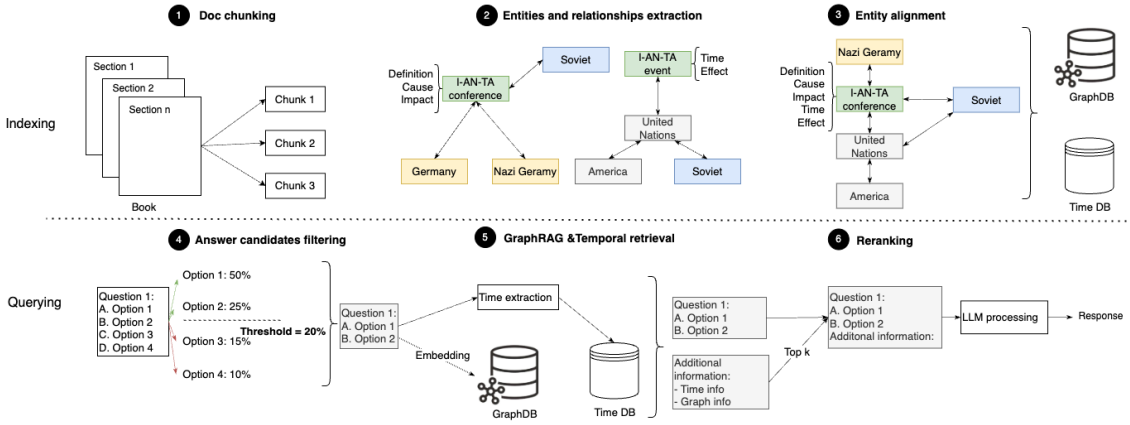


Figure 2: An overview of the HisGraphRAG process, from indexing through querying

- **Answer candidates Filtering:** GraphRAG's retrieval process often includes irrelevant or noisy entries. These not only add confusion but can override otherwise correct LLM responses. Selectively filtering and prioritizing useful information is thus critical.

- **Temporal Awareness:** Historical queries often hinge on timeline-sensitive information. Incorporating temporal cues into retrieval and reranking ensures the selected context aligns with the question's timeframe.

3.1 Doc Chunking

The input dataset for our system comprises history textbooks, which are inherently structured into well-defined chapters and sections. Traditional document chunking strategies—such as segmenting by a fixed number of tokens or characters—fail to preserve the semantic and topical coherence necessary for downstream tasks like retrieval and graph construction. To address this, we adopt a section-aware chunking strategy tailored to the structural granularity of historical texts.

Let a document be composed of a sequence of chapters $C = \{c_1, c_2, ..., c_n\}$, where each chapter c_i contains a sequence of sections $S_i = \{S_{i1}, S_{i2}, ..., S_{im}\}$. Each section S_{ij} is treated as a primary unit of chunking. To maintain context continuity across sections, we apply overlapping context windows by concatenating with a portion of the preceding section when available. Formally, each chunk x_{ij} is generated as:

$$x_{ij} = S_{ij-k} || S_{ij} \tag{1}$$

where k is the overlapping context parameter and $||$ denotes concatenation.

To comply with the LLM input limitations, If the resulting chunk x_{ij} exceeds a token threshold $T_{\max}$, we recursively subdivide it into smaller chunks $x_{ij}^1, x_{ij}^2, \ldots, x_{ij}^l$ such that:

$$\text{len}(x_{ij}^l) \leq T_{\max}, \quad \forall l \tag{2}$$

The history textbooks are provided in PDF format, which includes both text and non-textual elements (e.g., illustrations, charts, and exam questions). We preprocess these documents by converting each page into an image and then employ a visual-language model (VLM), or an LLM equipped with document vision capabilities, to extract textual descriptions from visual elements. This step ensures that all important content—especially figures relevant to historical narratives—is retained in natural language form. At the same time, we filter out irrelevant components, such as quiz questions, using a classification filter tuned to remove interrogative-style sentences. This preprocessing ensures that each document chunk is both semantically cohesive and enriched with textual equivalents of visual data, thus improving the quality of downstream retrieval and reasoning stages.

3.2 Entities and Relationships Extraction

Since our focus is exclusively on historical texts, we design the entity and relationship extraction process to target concepts that are particularly relevant to the domain. Specifically, we extract information related to historical events, key figures, locations, organizations, strategies, and their associated impacts. Each document chunk is processed using a LLM to identify named entities and infer relationships between them. For every pair of related entities, we assign a single relation type: `RELATE_TO`),

accompanied by a natural language description that explains the nature of the relationship. This design choice maintains simplicity while still preserving the contextual meaning between entities.

- A graph database that stores entities as nodes and their `RELATE_TO` relationships as edges, including descriptive annotations.
- Identities was categorized into: person, organization, event, place, action, strategy, impact.
- A SQL database that records temporal attributes of events and entities, enabling chronological filtering in later retrieval stages.

It is important to note that, unlike Microsoft's GraphRAG implementation, we do not include any community reports or crowdsourced annotations. Our method strictly constructs knowledge from the source documents themselves, maintaining a clean and verifiable information graph suitable for history domain.

3.3 Entity Alignment

As history textbooks are organized by chapters and sections, entities representing the same real-world concept (e.g., "Soviet", "Soviet Union") may appear multiple times across different chunks or chapters with slight variations in name or description. If not resolved, this entity duplication introduces noise in the retrieval process—returning redundant or conflicting results—and fragments relationships, weakening the structure of the knowledge graph. To address this, we perform entity alignment using a large language model (LLM) with extended context capabilities. We aggregate entities extracted across all sections and send them in batches to the LLM, which performs a semantic merge of duplicate or near-duplicate entities. For each group of similar entities, the model:

- Selects a canonical name and be validated by LLM then human
- Merges descriptions into a unified summary
- Merges associated relationships while eliminating redundancy

This process ensures that each unique historical figure, event, or organization is represented by a single unified node in the graph, improving both the retrieval precision and the integrity of relational links. By consolidating entity representations and summarizing associated content, the resulting graph becomes more compact, coherent, and interpretable—especially critical for query tasks that involve reasoning over historical timelines or cause-effect chains.

3.4 Answer candidates Filtering

To evaluate the effectiveness of our system, we consider a multiple-choice question-answering setting tailored for historical content. Each question is accompanied by four candidate options: A, B, C, D, with exactly one correct answer. The task is to select the correct option using the knowledge extracted from the document graph. A key challenge arises in the retrieval process: all answer options contribute equally to document retrieval regardless of correctness. This leads to the inclusion of distracting or irrelevant content from incorrect answers, which can degrade both retrieval precision and model confidence. In human examination settings, students naturally focus more on plausible answers while disregarding options that appear clearly incorrect. Inspired by this behavior, we introduce an Answer candidates filtering mechanism that prioritizes likely correct answers.

Given a question q and four answer choices $A = \{a_1, a_2, a_3, a_4\}$, we compute:

$$P(a_i \mid q) = \text{LLM\_score}(q, a_i) \tag{3}$$

Let θ be a confidence threshold. We retain only those options such that:

$$P(a_i \mid q) \geq \theta \tag{4}$$

This filtering strategy offers two major benefits. First, it aligns with the reasoning strategy of proficient human test-takers: before committing to an answer, they mentally discard distractors that are clearly implausible, thereby focusing attention on a narrower and more reliable candidate set. Second, from a technical perspective, the multiple-choice setup inherently introduces redundancy — both between the question and its answer candidates, and among the candidates themselves. Without filtering, this redundancy propagates into the retrieval process, increasing the risk of irrelevant or noisy evidence being incorporated. By retaining only those candidates whose likelihood surpasses the confidence threshold, our method reduces retrieval redundancy and sharpens the focus on semantically

meaningful information, ultimately improving both retrieval efficiency and answer accuracy.

3.5 Graph & Temporal Retrieval

In this step, we integrate semantic graph-based retrieval and temporal filtering to enhance the contextual information fed into the final prompt. These two retrieval processes leverage the structured outputs from Step 2: the graph database (containing entities and relationships) and the time-indexed event database.

Graph-based Retrieval:

Given a query q which contains question and answer candidates from Step 4, we embed them into a dense vector $\vec{q}$ using a pre-trained model. Each graph node e_j has an embedding $\vec{e_j}$. Compute similarity:

$$v_q = Embedded(q) \tag{5}$$

Simultaneously, each entity node e in the graph database is stored with its own semantic embedding v_e, computed from both the entity's name and its description. Retrieval is performed by computing similarity scores:

$$\text{sim}(q, e_j) = \cos(\vec{v_q}, \vec{v_e}) \tag{6}$$

Entities with similarity above a retrieval threshold δ are selected:

$$\varepsilon_{retrieved} = \{e \in \mathcal{E} \mid \text{sim}(q, e) \geq \delta\} \tag{7}$$

For each retrieved entity e, we also extract its associated relationships and descriptions to construct a subgraph relevant to the question context

Temporal Retrieval:

To complement graph-based retrieval, we perform time-aware filtering based on temporal cues in the query. If the query q contains a time reference (e.g., a year or period), we extract it and define a time window:

$$[t_{start}, t_{end}] = [t_q - \Delta t, t_q + \Delta t] \tag{8}$$

Finally, the union of retrieved graph-based and time-based information:

$$C_q = \varepsilon_{graph} \cup \varepsilon_{temporal} \cup \varepsilon_{textchunk} \tag{9}$$

3.6 Reranking

The retrieval stage often results in a large set of semantically and temporally relevant entities, relationships, and events. However, not all of these contribute equally to answering the given question. Including too many can overwhelm the prompt and reduce the effectiveness of the language model's reasoning. To ensure focus and precision, we introduce a reranking mechanism driven by the LLM itself.

Instead of relying on a traditional scoring function or trained reranker, we leverage the LLM's internal reasoning to identify the most relevant context items. Specifically, we prompt the LLM to act as a selector, choosing a subset of the retrieved content that is most likely to support answering the question.

Given a query q and retrieved context $C_q = \{c_1, c_2, ..., c_n\}$, we prompt the LLM to select the most relevant k items:

$$C_q^{\text{top-k}} = \text{Select}(q, C_q, k) \tag{10}$$

where k is typically set to a small number to maintain prompt compactness.

This LLM-guided reranking allows for context-aware prioritization, capturing subtle relevance signals that might be missed by static similarity metrics. The selected top-k entities and facts are then composed into the final input prompt for the answering model, significantly improving focus and factual consistency in the output.

4 Experiment

4.1 Dataset

To evaluate the effectiveness of our GraphRAG-based retrieval and reasoning framework in the domain of historical question answering, we construct both a retrieval corpus and a evaluation dataset.

Retrieval Corpus: We use the official Grade 12 Vietnamese History textbook, which serves as the knowledge base for information retrieval and graph construction. The textbook is representative of high school-level historical content and provides rich, structured information in the form of chapters and sections, including text, figures, and timelines.

Evaluation Dataset: For the downstream question answering task, we use real-world standardized multiple-choice questions from the Vietnamese University Entrance Exams in the years 2017 and 2018. For each year, we include three official exam

codes, resulting in a total of 120 multiple-choice questions (3 sets × 40 questions) called VUEE-2017 and VUEE-2018. Each question consists of one correct answer and three distractors, following the standard national testing format.

4.2 Experimental setting

We implement our GraphRAG-based system using a modular architecture combining large language models, graph databases, and vector search engines. The key components and hyperparameters are configured as follows:

- Main Language Model: We use ChatGPT-4o mini as the primary model for all reasoning tasks, including entity alignment, reranking, and final answer generation. Its ability to handle long-context inputs and follow prompt instructions makes it well-suited for this structured retrieval setting.
- Embedding Model: For dense retrieval and semantic similarity computation, we use the text-embedding-3-small model. All entities and question inputs are embedded using this model and stored for efficient nearest-neighbor search.
- Graph Database: We store structured entities and their relationships using Neo4j, which enables traversal and context construction via `RELATE_TO` edges.
- Vector Database: All embeddings are stored and queried through Milvus, which supports high-performance similarity search over dense vector representations.
  - Temporal Window: For time-based filtering, we extract time expressions from the query and apply a ±1 year buffer to retrieve relevant historical events.
  - Answer Filtering Threshold: For question option filtering (Step 4), we discard any answer option with a model-estimated confidence below 15%.
  - Reranking Top-k: After retrieval, we prompt the LLM to select the top 15 most relevant context items for final reasoning.
- Implementation Base: Our system builds upon and extends the open-source GraphRAG implementation from the public repository [1].

[1] https://github.com/ImmortalDemonGod/nanographrag

5 Result

5.1 Indexing result

To evaluate the efficiency cost of the proposed entity alignment step, we measured token usage and processing time during the indexing phase, both with and without alignment.

Cost	w/o entity alignment	with entity alignment
In token consumption	605k	669k in
Out token consumption	190k	219k out
Time consumption	322s	270s

Table 2: LLM consumption during indexing step

As shown in Table 2, applying entity alignment leads to a modest increase in input token consumption from 605k to 669k tokens, an approximate 10.6% increase. Similarly, output token usage increases from 190k to 219k tokens, reflecting the additional summarization and relational consolidation introduced during alignment. Despite this increase in token count, the overall time consumption decreases slightly, from 322 to 270 seconds, possibly due to the more compact and structured representation of merged entities reducing downstream processing complexity.

5.2 Querying results

We evaluate the effectiveness of our proposed system, HisGraphRAG, along with several baseline methods and incremental variations, on two standardized historical question-answering benchmarks: VUEE-2017 and VUEE-2018. Table 3 reports the accuracy, token consumption, and response time across all methods.

For the baseline comparison, the LLM-only setting—where the model answers without any external retrieval—achieves the lowest accuracy on both datasets (68.3% for VUEE-2017 and 50.8% for VUEE-2018), indicating the insufficiency of parametric knowledge alone for detailed historical reasoning. When retrieval is added in a naive form via NaiveRAG, which lacks any structured understanding of entities or relationships, performance improves modestly (72.5% and 54.2%), suggesting that even simple information augmentation can be beneficial. These baselines provide a reference point for the gains enabled by structured and filtered retrieval.

When moving to GraphRAG, which incorporates entity-level structure and relationship reasoning, the accuracy improves significantly to 77.5% for VUEE-2017 and 60.0% for VUEE-2018. Adding

Method	VUEE-2017			VUEE-2018		
	Acc	Token	Time	Acc	Token	Time
LLM-only	68.3%	22k	80s	50.8%	23k	80s
NaiveRAG	72.5%	102k	145s	54.2%	95k	150s
GraphRAG	77.5%	920k	210s	60.0%	907k	218s
GraphRAG + Entity alignment	76.7%	1.4M	377s	61.7%	1.4M	385s
GraphRAG + Temporal & Rerank	73.3%	1.7M	927s	56.7%	1.7M	773s
GraphRAG + Answer candidates Filter	80.8%	949k	384s	63.3%	983k	360s
HisGraphRAG (Ours)	**81.7%**	**2.2M**	**977s**	**70.8%**	**1.6M**	**922s**

Table 3: HisGraphRAG accuracy result on VUEE-2017 and VUEE-2018 dataset

entity alignment to GraphRAG yields mixed results: a slight decrease in 2017 (76.7%) but a notable improvement in 2018 (61.7%), indicating that alignment becomes more impactful when the dataset contains more ambiguity or entity duplication. Interestingly, adding temporal retrieval and reranking in isolation causes a decline in accuracy (73.3% and 56.7%), likely because this configuration pulls in excess temporal content not always directly relevant to the question—overloading the prompt and reducing model focus. However, when question filtering is applied alone, accuracy increases sharply (80.8% and 63.3%), highlighting the importance of removing low-confidence distractors before retrieval. Ultimately, the HisGraphRAG system, which integrates all components—including alignment, temporal retrieval, reranking, and filtering—achieves the highest accuracy on both datasets (81.7% and 70.8%).

While **HisGraphRAG** achieves the best results, it does so with increased computational cost. The number of tokens consumed rises to 2.2M and 1.6M respectively for VUEE-2017 and VUEE-2018, and average response time approaches 900 seconds. These costs reflect the additional reasoning steps introduced in the pipeline, but are justified in scenarios where accuracy and context richness are more important than latency. Notably, all systems perform worse on VUEE-2018 than on VUEE-2017, which is consistent with historical data showing that the 2018 exam included more difficult questions, as reflected in average student scores. Even under these conditions, HisGraphRAG maintains strong performance margins, reinforcing its robustness across varying levels of difficulty.

5.3 Error analysis

We analyze one illustrative example from our test data to demonstrate how each component of our method contributes to improved accuracy.

Example Question:

> At the *Second National Congress* (February 1951), the *Indochinese Communist Party* decided to establish in each country of Indochina:

A. **Marxist-Leninist Party (Correct)**

B. Coalition Government

C. United Front

D. Armed Force

We summarize the impact of each component in Table 4.

Our proposed method, **HisGraphRAG**, significantly improves the precision and relevance of retrieval by incorporating three key strategies:

- **Entity Alignment** resolves redundancies by merging semantically identical entities into canonical representations, thus eliminating unnecessary duplication and optimizing prompt usage.
- **Temporal Retrieval** leverages temporal context to filter information, constraining retrieved historical events within a precise chronological window relevant to the question.
- **Answer Candidate Filtering** systematically estimates the probability of correctness for each potential answer choice, proactively removing irrelevant distractors to avoid misleading the model.

Component	Without our Method (GraphRAG)	With our Method (HisGraphRAG)
Entity Alignment	Retrieved redundant entities *"Indochinese Communist Party"* and *"Communist Party of Indochina"*, wasting prompt space and causing duplication.	Merged duplicate entities into a single canonical form *"Indochinese Communist Party"*, reducing redundancy and increase useful relationships.
Temporal Retrieval	Retrieved irrelevant congresses (*First* and *Third*), confusing the model with unrelated decisions.	Extracted date (1951) and retrieved events within ±1 year only, filtering out irrelevant historical congresses.
Answer Candidate Filtering	Included distracting information about *Coalition Governments* formed in different historical contexts, misleading the model toward incorrect Option B.	Filtered out irrelevant candidates by estimating answer probabilities, ensuring only contextually relevant evidence is included.

Table 4: Comparison of outputs with and without key components of our proposed method (HisGraphRAG).

Together, these improvements collectively narrow evidence to highly relevant, concise, and contextually precise content, markedly enhancing the probability of correctly identifying the output—**Marxist-Leninist Party** instead of United Front (answer from GraphRAG). acknowledgement

6 Conclusion

In this work, we presented HisGraphRAG, a novel retrieval-augmented generation (RAG) framework specifically designed to address the challenges of historical multiple-choice question answering. Our method incorporates four key components—entity alignment, temporal-aware retrieval, contextual reranking, and candidate answer filtering—to deliver more accurate and contextually grounded responses. By aligning temporal and entity-level information through structured graph-based knowledge and targeted retrieval strategies, HisGraphRAG significantly enhances the reasoning capability of large language models in historical domains. Compared to standard RAG and GraphRAG baselines, our approach achieves notable gains in accuracy, demonstrating its ability to reduce confusion caused by duplicated entities, temporal inconsistencies, and irrelevant distractors. These findings underscore the critical importance of integrating structured semantic knowledge and temporal constraints to support complex historical understanding and decision-making.

While HisGraphRAG shows promising improvements, it also increases computational cost due to longer query times and higher token consumption. In future work, we plan to optimize the temporal retrieval process, explore more efficient indexing strategies, and develop scalable entity alignment techniques, such as hybrid embedding methods and rule-based clustering. These enhancements will help make the approach faster and more practical for large-scale and real-time applications.

7 Acknowledgement

This research is partially funded by the Vingroup Innovation Foundation (VINIF) under the grant number VINIF.2021.JM01.N2

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kai Guo and al. 2025. Empowering GraphRAG with Knowledge Filtering and Integration. *arXiv preprint arXiv:2503.13804v1*.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, and et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Poster #4607, East Exhibit Hall A-C.

Zhijing Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yujin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38.

Boran Jiang, Yuqi Wang, Yi Luo, Dawei He, Peng Cheng, and Liangcai Gao. 2024. Reasoning on efficient knowledge paths: Knowledge graph guides large language model for domain question answering. *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 142–149.

Shailza Kumar and Ellie Pavlick. 2022. Revisiting zero-shot question answering performance of multilingual language models on low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3669–3679.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Douwe Kiela, Anurag Batra, Tim Rocktäschel, and Sebastian Riedel. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Mandar Kulkarni, Mike Luo, Ian Wigham, Daniel Beck, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *ICLR 2025 Conference*.

Microsoft Research. 2024. GraphRAG: Graph-Augmented Retrieval for LLMs. `https://microsoft.github.io/graphrag/`.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169.

Xiaokui Xiao Yiqian Huang, Shiqi Zhang. 2025. A cost-efficient multi-granular indexing framework for graph-rag. *arXiv preprint arXiv:2502.09304*.

Yichong Zhou, Mo Yu, Tongtao Zhao, Huan Sun, and Claire Cardie. 2022. Temporal question answering with sequential temporal graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 740–753.

Time Tells: Temporal Event Ordering in Frontier LLMs — Performance, Limitations, and Human Comparison

Feifei Sun[1], Ziyi Tong[1,], Teeradaj Racharak[2], Minh Le Nguyen[1],

[1]Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]Advanced Institute of So-Go-Chi (Convergence Knowledge) Informatics, Tohoku University, Japan

Correspondence: racharak@tohoku.ac.jp, nguyenml@jaist.ac.jp

Abstract

Understanding temporal information is essential for natural language understanding, yet both humans and large language models (LLMs) face challenges when narratives mix absolute and relative time expressions. In this study, we systematically evaluate frontier LLMs on temporal event ordering tasks and compare their performance against human baselines under absolute-time (AT) and mixed-time (MT) conditions. Using a recently constructed temporal reasoning dataset [1], we analyze representative models including GPT-4, DeepSeek Reasoner, and QwQ-32B.

Our findings reveal three key insights: (1) Frontier LLMs can achieve near-human Kendall's τ value in AT settings, with GPT-4 and DeepSeek Reasoner performing competitively. (2) In MT scenarios, human performance drops more sharply than some LLMs, suggesting that frontier models can maintain stronger consistency in long-text temporal reasoning with mixed time expressions. (3) Targeted probing with time masking confirms that LLMs rely heavily on explicit temporal anchors, showing fragility when such cues are removed.

These results demonstrate that temporal reasoning remains a core challenge for both humans and LLMs, while also revealing conditions under which models can rival—or even approach—human performance. Our analysis provides actionable insights for improving the interpretability and robustness of LLMs in temporally grounded language tasks.

1 Introduction

Temporal reasoning is a fundamental aspect of language comprehension, enabling interlocutors to reconstruct event sequences and interpret narratives in context. In natural discourse, however, temporal information is rarely expressed in a fully explicit or linear fashion. Narratives across languages frequently weave together absolute time references (e.g., in 1945), relative expressions (e.g., two years later), and event-anchored cues (e.g., shortly after the war), producing timelines that are non-linear and partially implicit. This mixture poses challenges not only for computational models, but also for human readers—particularly when temporal cues are sparse, distributed, or dependent on discourse structure.

Recent advances in LLMs have shown promising results in temporal reasoning tasks. Yet, most evaluations have been conducted in settings dominated by explicit temporal anchors or simplified temporal relations. Such conditions do not fully reflect the complexity of real-world narratives, including those found in historical accounts, biographies, and cross-cultural storytelling, where temporal markers are often underspecified or require inferential bridging. However, existing benchmarks rarely capture this mixture of absolute and relative time, and few studies have systematically tested how LLMs behave under these more naturalistic, mixed-time conditions (Chu et al., 2023; Wang and Zhao, 2023; Tan et al., 2023). Moreover, while recent studies have identified anchoring biases in LLMs (Huang et al., 2025) and systematic weaknesses in handling relative temporal expressions (Chen et al., 2025), little is known about whether current evaluation methods can reveal the sources of model errors more comprehensively—specifically, whether they arise from difficulties in anchoring absolute references, integrating relative cues, or bridging across discourse gaps. From a computational linguistics perspective, understanding how models—and humans—navigate such mixed-time conditions is key to developing systems that are both robust and interpretable across languages.

Building on a recently introduced mixed-time temporal reasoning benchmark (Sun et al., 2025), this study makes three contributions:

[1]https://github.com/fantastic-Feifei/MTS-benchmark

- Human–model comparative evaluation: We establish a human baseline for temporal event ordering under absolute-time (AT) and mixed-time (MT) conditions, quantifying the inherent difficulty and ambiguity of hybrid-time narratives.
- Probing temporal robustness: We conduct targeted masking experiments that selectively remove explicit year expressions, revealing the extent to which LLMs rely on surface-level temporal anchors.

Our findings show that while frontier LLMs can match or even exceed human performance in explicit AT settings, both humans and models struggle when temporal anchors are obscured or replaced with relative expressions. Crucially, masking a single year often triggers a collapse in model ordering accuracy, underscoring the importance of context integration beyond explicit timestamps. These results have direct implications for the design of temporally aware NLP systems and contribute to the broader understanding of how temporal reasoning operates across varied linguistic and narrative structures.

2 Related Work

Temporal Reasoning in NLP Temporal reasoning has been a long-standing challenge in natural language processing, involving tasks such as temporal expression normalization (Verhagen et al., 2010), event ordering (Chambers and Jurafsky, 2008; Ning et al., 2019), and temporal question answering (Khot et al., 2020; Chen et al., 2021). Most existing benchmarks, including TimeBank, MATRES, and TORQUE, primarily focus on scenarios with explicit absolute time anchors or simplified temporal relations. However, real-world narratives often exhibit *mixed temporal structures* that combine absolute, relative, and event-anchored expressions, forming non-linear timelines. Deriving a coherent global event order from such narratives remains challenging, especially when temporal cues are implicit or distributed across long contexts.

Large Language Models for Temporal Understanding Recent research has explored the ability of LLMs to perform temporal reasoning in complex narratives. Although state-of-the-art LLMs such as GPT-4 and DeepSeek demonstrate strong capabilities in general reasoning tasks, their performance often degrades on event ordering when explicit date cues are removed or when relative temporal expressions dominate (Xiong et al., 2024; Ding and Wang, 2025; Yuan et al., 2024). LLMs typically rely on surface-level temporal signals to achieve partial ordering consistency, but they struggle with non-linear or mixed time settings, where reasoning requires integrating both explicit and implicit temporal cues. This gap motivates the need for systematic evaluations of LLM temporal robustness under both AT and MT conditions.

Probing Methods for Model Reasoning Probing techniques offer a principled approach for investigating the internal reasoning behavior of LLMs (Belinkov and Glass, 2019; Elazar et al., 2021). In the context of temporal reasoning, one widely-used strategy is to mask time expressions—whether absolute or relative—during inference or intermediate training, and then evaluate the model's ability to reconstruct event order purely from context.

For example, Cole et al. 2023 introduce Temporal Span Masking, where temporal expressions are selectively masked during intermediate training to enhance performance on downstream temporal tasks. Similarly, the TempoBERT model (Rosin et al., 2021) employs explicit time masking incorporated into the model's inputs, boosting accuracy in temporal prediction tasks (Rosin et al., 2022). More recently, Liu et al., 2025 propose the Time-R1 framework, which includes a "masked time entity completion" subtask—directly analogous to our design—in its multi-stage training to assess the model's reliance on narrative context rather than explicit temporal markers.

Our method similarly adopts a targeted masking strategy by selectively removing absolute or relative time expressions and evaluating whether models can still recover the correct event sequence. Unlike prior masking-based probing methods that primarily evaluate lexical recovery, our design masks *absolute* time anchors (e.g., "in 1995"), which are crucial chronological cues in narrative texts. This allows us to directly test whether models can maintain temporal coherence and reason based on commonsense or contextual inference when explicit anchors are missing. Such a probing setup enables fine-grained analysis of model reliance on temporal cues, highlights differences between human and model reasoning, and supports controlled experiments assessing temporal robustness in long-context event ordering tasks.

3 Dataset and Task Setup

3.1 Dataset Overview

Statistic	Value
Total passages	4,824
Avg. events per passage	7.99
Avg. relative per passage	4.53
Avg. absolute per passage	3.46
Relative time ratio (relative / all)	56.73%

Table 1: Full dataset statistics, showing event density and distribution of absolute vs. relative time expressions.

We conducted our experiments on the hybrid-time temporal reasoning dataset, which contains 4,824 Wikipedia-style biographical passages (Table 1) with time-stamped events and various temporal expressions (Table 2). The dataset can be accessed from our GitHub repository (linked in the abstract).

Terminology

Global event refers to an event's position in the complete chronological sequence of a passage (e.g., *1960* $\rightarrow$ *1980* $\rightarrow$ *1990* $\rightarrow$ *2000*).
Local context denotes the immediately surrounding sentences that help determine an event's position.
Temporal cues are explicit or implicit indicators of time, which in our dataset include: **Absolute** (e.g., "in 1945"), explicit chronological anchors; **Relative** (e.g., "two years later"), requiring contextual inference; **Event-anchored** (e.g., "the end of 47th Olympics"), dependent on prior events.

This *mixed-time* design reflects natural narratives where explicit dates are interwoven with relative and discourse-dependent cues, requiring models to combine surface-level anchors with contextual reasoning. The coexistence of multiple expression types enables controlled comparisons of reasoning strategies and fine-grained robustness analysis.

Each passage is annotated with normalized temporal values and aligned to a global event sequence, supporting two evaluation settings: **AT** and **MT**. This design probes model sensitivity to explicit vs. implicit temporal cues and offers reusability for cross-linguistic evaluation, temporal QA, timeline extraction, and discourse analysis.

3.2 Dataset Sampling

To facilitate human annotation and maintain clarity in subsequent analysis, we sampled 100 passages from the full 4,824-instance dataset (Table 3), each containing between 4 and 7 event sentences. This range strikes a balance between narrative richness and human annotator feasibility. Our sampling decision was also guided by evidence from cognitive psychology showing that human comprehension of long, complex texts is constrained by working memory and processing limitations (Sugawara et al., 2020; Kalyuga, 2011). Specifically, when discourse length increases, readers must concurrently integrate earlier information while processing new content, which strains limited memory resources. Therefore, to ensure high-quality human annotation in our probing tasks, we limit the sample size to 100 passages while retaining narrative richness. As later confirmed in Section 4.1, humans exhibit lower consistency than models when reasoning over longer passages.

The selected subset preserves a diverse mixture of time expressions (absolute, relative) and event ordering structures. All 100 passages were independently annotated by three trained annotators and subsequently used for both human-model comparison and fine-grained probing experiments.

All annotators are domain experts in Information Science, with research backgrounds in machine learning and strong proficiency in English. This expertise ensures both familiarity with the technical aspects of the task and the linguistic competence required to process the biographical passages, providing a reliable baseline for comparison with model predictions. In addition, one of the annotators is the co-author of this work.

3.3 Task Definition

We define a sentence ordering task that requires models to recover the global temporal order of events described in natural language. Each input consists of a set of four event sentences sampled from a temporally rich passage. The model is prompted to output the correct chronological order of sentence indices (e.g., "1,3,2,4").

We evaluate model performance under two settings:

- **Absolute-Time (AT)**: All original absolute time expressions (e.g., "in 1945") are preserved.

Table 2: Examples of temporal expressions in our dataset, categorized by expression type: Absolute, Relative, and Event-Anchored.

Type (Temporal Expression)	Example
Absolute	"in 1995", "on September, 1920"
Relative	"three years later", "shortly after the Expo"
Event-Anchored	"the end of 47th Olympics", "during the Great Depression"

Setting	AT	MT
#Samples	30	70
Avg Events	4.4	4.5
High Granularity (%)	53.3	42.9
Low Granularity (%)	46.7	57.1
Abs : Rel	53 : 47	43 : 57

Table 3: Statistics of the 100-sample subset for human annotation and probing experiments. This subset maintains temporal diversity while ensuring cognitive feasibility for human reasoning. *Avg Events* denotes the average number of annotated events per passage. *High Granularity* = expressions specifying *year+month+day* and *year+month*, and *Low Granularity* = expressions specifying only the *year*.

- **Mixed-Time (MT)**: Some absolute expressions are rewritten into relative forms (e.g., "eight years later") or event-anchored references (e.g., "August 2024" rewritten as "the end of 47th Olympics") to simulate hybrid time contexts.

To further probe model dependency on explicit time anchors, we introduce a masked-time variant of this task, in which one temporal expression is replaced with a `[MASK]` token (see Section 3.5). Models must still recover the correct sentence order, revealing their temporal inference capabilities under partial information.

3.4 Model Selection

To evaluate temporal reasoning capabilities across a diverse range of model families and architectural scales, we selected the following representative large language models for analysis:

GPT-4 (OpenAI) (Achiam et al., 2023): A frontier proprietary model known for its strong general reasoning and instruction-following abilities. It serves as a high-performance reference in our evaluations.

DeepSeek-Reasoner (DeepSeek) (Guo et al., 2025): A reasoning-optimized model designed for multi-hop and structured inference tasks. It represents a model explicitly trained with a focus on reasoning capabilities.

DeepSeek-v3 (DeepSeek) (Liu et al., 2024): A general-purpose instruction-tuned model from the same family, included to contrast with the reasoning-augmented variant.

QwQ-32B (Alibaba) (Team, 2025): A large-scale open-source model with competitive performance in general benchmarks.

Qwen2.5-7B (Alibaba) (Yang et al., 2024): A strong open-source base model with relatively smaller scale (7B), chosen to assess how compact models perform under temporal reasoning tasks.

These models span different training paradigms (instruction tuning, reasoning augmentation), sizes (7B–32B), and sources (open-source vs. proprietary), enabling a broad comparison of temporal reasoning performance across architectural and methodological dimensions.

3.5 Probing Design

Masked Time Prediction

To further examine the temporal reasoning ability of language models, we design a *masked time prediction* task that targets a model's capacity to infer missing temporal information from surrounding context (Table 4).

For each passage in both the AT and MT settings, we randomly select a single sentence containing an absolute time expression (e.g., "1945") and mask only the year component with a `[MASK]` token. The remainder of the sentence and surrounding context are preserved, enabling us to isolate the model's ability to recover or approximate the masked temporal anchor based on event order and discourse-level clues. In the MT setting, the masked sentence may contain a rewritten relative reference (e.g., "eight years later"), making the task more reliant on understanding inter-event relations rather than directly reading explicit anchors.

We deliberately mask only one temporal expression per passage for three reasons:

1. **Experimental control:** Masking a single anchor allows us to attribute any performance change directly to the removal of that cue, avoiding confounding effects from masking multiple references simultaneously.

2. **Linguistic realism:** In naturally occurring narratives, explicit temporal anchors are often sparse but not entirely absent; removing just one simulates this partial loss of explicit cues.

3. **Interpretability:** With only one masked expression, we can more clearly trace whether the missing anchor disrupts global ordering, making error patterns easier to analyze.

Rather than evaluating the exact lexical reconstruction of the masked expression—which may admit multiple valid rewrites—we assess the model's ability to recover the correct global event order when the masked sentence is included. This design enables us to measure whether the removal of a key temporal anchor significantly degrades downstream temporal reasoning, thereby revealing the extent to which models depend on explicit time cues versus contextual inference.

Prompt Construction

To assess whether language models can preserve temporal reasoning capabilities when explicit time anchors are partially removed, we formulate a masked time prediction task as a sentence reordering problem. Each instance consists of four event sentences, one of which has its year expression masked (e.g., in 1997" $\rightarrow$ in [MASK]").

To ensure consistent and well-structured outputs across models, we adopt a one-shot prompting strategy. Preliminary experiments in the zero-shot setting showed that only GPT-4 reliably followed the expected output format (i.e., returning a comma-separated list of sentence indices such as "1,2,3,4"). In contrast, open-source models such as DeepSeek and Qwen often produced verbose, unstructured, or incomplete responses. To mitigate this, we include a concrete in-context example at the beginning of each prompt.

Each prompt is composed of two parts: a worked example followed by a test instance. Both follow the same structure—numbered sentences and an instruction asking the model to reorder them chronologically using index notation.

This design allows us to probe the impact of masking a single temporal anchor on the model's ability to infer global event order. By applying this setup to both AT and MT settings, we can evaluate the degree to which models rely on explicit time expressions versus contextual temporal reasoning.

A detailed example of our prompt format is provided in Appendix A.

3.6 Evaluation Metrics

To comprehensively assess the temporal reasoning capability of language models, we employ the following evaluation metrics:

- **Exact Match (EM):** For event ordering tasks, EM is a strict metric that checks whether the predicted event sequence exactly matches the gold order. Unlike rank-based correlation measures (e.g., Kendall's τ), which provide partial credit for partially correct rankings, EM only assigns credit when the entire sequence is perfectly correct.

- **Kendall's τ:** For sentence reordering tasks, we compute Kendall's τ rank correlation coefficient between the predicted and gold-standard sentence orders. This metric captures the *pairwise consistency* of temporal relationships between events, providing a gradient of correctness even when the full order is not exact.

In addition to aggregate scores, we conduct fine-grained analyses along the following dimensions:

- **Time Expression Type:** We compare model performance on absolute (e.g., "in 1982") vs. relative (e.g., "eight years later") expressions, to evaluate their sensitivity to different temporal formats.

- **Time Granularity:** We define granularity as the level of specificity expressed in temporal anchors (year, year+month, year+month+day). This is distinct from the *type* of temporal expression (e.g., absolute vs. relative vs. event-anchored), which we treat as a separate factor.

- **Context Length:** We investigate whether the number of surrounding events in a passage influences the model's ability to infer temporal relations, shedding light on context sensitivity.

Together, these metrics provide a multi-faceted view of model behavior, balancing surface-form fidelity with structural reasoning competence.

Original Context (Before)	Masked Context
He graduated in **1998** and started working the following year.	He graduated in **[MASK]** and started working the following year.
He became president in **2009** after a decade in parliament.	He became president in **[MASK]** after a decade in parliament.
She left the ministry in **March 2003** and returned briefly in 2005.	She left the ministry in **[MASK]** and returned briefly in 2005.

Table 4: Examples of original and masked contexts used in the probing task.

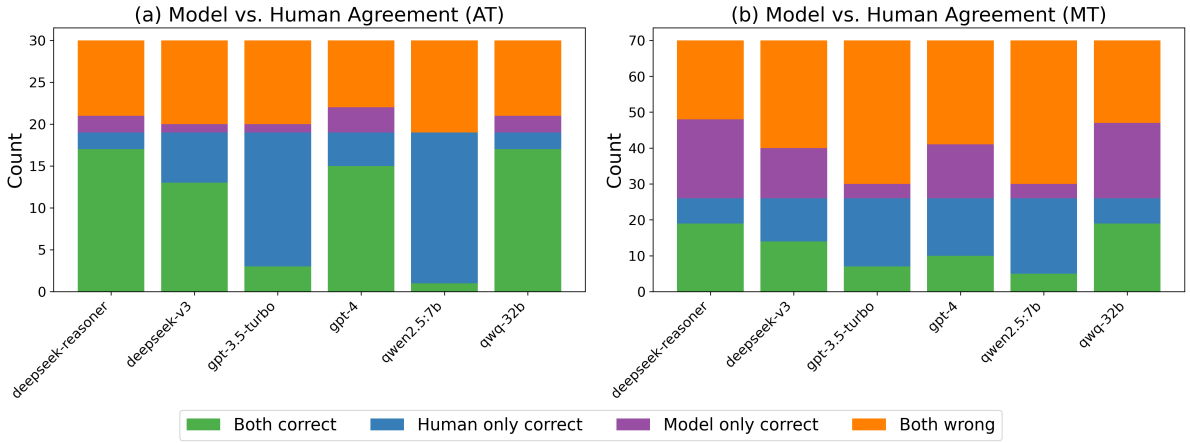


Figure 1: **Model vs. Human Agreement under AT and MT settings.** Stacked bar plots show the distribution of prediction outcomes for each model: *Both correct* (green), *Human only correct* (blue), *Model only correct* (purple), and *Both wrong* (orange). (a) **Absolute Time (AT):** Models such as DEEPSEEK-REASONER, GPT-4, and QWQ-32B achieve the highest human–model overlap, while DEEPSEEK-V3 and QWEN2.5-7B show more human-only correct cases, indicating weaker recovery of masked time expressions. (b) **Mixed Time (MT):** All models see a sharp drop in *Both correct* counts, with increased *Model only correct* and *Both wrong* cases. The gap between human and model judgments widens under ambiguous or relative time cues, especially for QWEN2.5-7B and DEEPSEEK-V3.

4 Results and Analysis

4.1 Overall Model Performance

To establish a performance baseline, we first evaluate six language models on the temporal ordering task under two settings: AT and MT, using two metrics: exact match (EM) accuracy and Kendall's τ rank correlation. Table 5 summarizes the results for four representative models and three human annotators.

Performance under AT Under the AT setting, all models achieve relatively high accuracy, with DEEPSEEK-REASONER and QWQ-32B both reaching an EM score of 0.63. GPT-4 also performs well with an EM of 0.60 and the highest Kendall's τ of 0.69, suggesting strong global ordering consistency. Compared to human annotators, most models perform on par or slightly better in EM, though ANNOTATOR 2 achieves the highest τ score of 0.73, indicating the strongest ordering alignment with gold labels.

Reliable Yet Nuanced: Human Baseline Under AT Interestingly, all three annotators achieved an identical exact match (EM) score of 0.56 under the AT setting, despite differing moderately in their Kendall's τ values. This suggests that while their full-sequence predictions aligned with the gold order in the same proportion of cases, the extent to which their orderings agreed with the gold ranking varied. Such consistency in EM across annotators reinforces the reliability of the human baseline under AT, and highlights that partial sequence disagreements (captured by τ) may still occur even when EM scores coincide.

Performance under MT The MT setting introduces a pronounced performance drop for both human annotators and models. While DEEPSEEK-REASONER and QWQ-32B retain relatively strong performance (EM = 0.60 and 0.59, respectively),

other models such as DEEPSEEK-V3 and GPT-4 show substantial degradation (EM = 0.41 and 0.42). Human performance declines even more noticeably: all three annotators exhibit reduced EM and Kendall's τ scores, with ANNOTATOR 1 dropping to an EM of just 0.30.

Humans Struggle, Models Endure under MT This gap between human and model performance may be attributed to the increased cognitive load imposed by long passages and temporally ambiguous references. Unlike the AT setting—where time expressions are explicit and reasoning is more straightforward—the MT condition requires interpreting implicit and relative temporal cues, often embedded within complex narratives. These challenges appear to hinder human consistency, whereas models like DEEPSEEK-REASONER and QWQ-32B demonstrate notable robustness, suggesting their superior ability to track and resolve temporal structure in long-context settings.

GPT-4 Mimics Human Temporal Reasoning Patterns Among all evaluated models, GPT-4 consistently demonstrated performance most closely aligned with that of human annotators. In the AT setting, its Kendall's τ of 0.69 is only marginally above the human range (0.62–0.73), and in the MT setting, its Kendall's τ of 0.46 remains within the variance of human annotators (0.32–0.50) (Table 5). This closeness in rank correlation suggests not only comparable accuracy but also similar ordering tendencies, indicating a reasoning style that aligns with human temporal judgments.

We hypothesize that GPT-4's relatively human-like behavior may stem from its training paradigm. Unlike open-source models such as DEEPSEEK-REASONER and QWQ-32B, which are often trained with strong emphasis on instruction tuning and retrieval-augmented generation, GPT-4 incorporates extensive reinforcement learning from human feedback (RLHF). This iterative alignment process likely encourages the model to mimic human preferences and inference styles, especially in ambiguous or underspecified contexts.

In contrast, DEEPSEEK-REASONER and QWQ-32B exhibit more decisive but less human-consistent behavior. Their superior performance under the MT condition suggests stronger capabilities in long-context tracking and structured reasoning, yet their predictions often diverge from human tendencies, possibly due to a more pattern-driven or memorization-based inference mechanism shaped by large-scale instruction-following pretraining.

Setting	Model	EM	Kendall's τ
AT	DEEPSEEK-REASONER	**0.63**	0.65
	DEEPSEEK-V3	0.46	0.47
	GPT-4	0.60	0.69
	GPT-3.5-TURBO	0.13	0.30
	QWEN2.5-7B	0.03	0.12
	QWQ-32B	**0.63**	0.66
	Annotator 1	0.56	0.62
	Annotator 2	0.56	**0.73**
	Annotator 3	0.56	0.63
MT	DEEPSEEK-REASONER	**0.60**	**0.65**
	DEEPSEEK-V3	0.41	0.49
	GPT-4	0.42	0.46
	GPT-3.5-TURBO	0.16	0.08
	QWEN2.5-7B	0.15	0.20
	QWQ-32B	0.59	**0.65**
	Annotator 1	0.30	0.32
	Annotator 2	0.34	0.43
	Annotator 3	0.42	0.50

Table 5: Overall model and human performance on event ordering under AT and MT conditions. EM = exact match accuracy; Kendall's τ measures rank correlation between predicted and gold orders.

Taken together, these findings suggest that GPT-4, while not always achieving the highest EM scores, may employ a reasoning process that is cognitively closer to human temporal understanding—a property valuable for downstream applications requiring interpretability or human-in-the-loop decision making.

Performance of Qwen2.5-7B and GPT-3.5-Turbo Both QWEN2.5-7B and GPT-3.5 underperform compared to larger frontier models. While QWEN2.5-7B yields near-random orderings (AT: EM=0.03, τ=0.12; MT: EM=0.15, τ=0.20), GPT-3.5 achieves slightly higher consistency (AT: EM=0.13, τ=0.30), yet still lags behind human annotators and fails to scale under MT (EM=0.16, τ=0.08). These results suggest that both models struggle with global event ordering, albeit for different reasons: QWEN2.5-7B appears particularly weak in leveraging absolute anchors, while GPT-3.5 shows instability in mixed-time reasoning.

Due to its consistently poor performance across both conditions, we exclude QWEN2.5-7B,GPT-3.5-TURBO, and DEEPSEEK-V3 from subsequent probing analyses.

4.2 Human-Model Agreement Patterns

To better understand the consistency between model predictions and human judgments, we analyze agreement patterns across six representative

models under both AT and MT conditions. Each instance in our dataset was independently annotated by three human annotators. To enable consistent comparison with model outputs, we adopted a majority vote strategy to determine the gold-standard human response for each instance.

Figure 1 presents stacked bar plots that categorize each prediction outcome into one of four types: *Both correct*, *Human only correct*, *Model only correct*, and *Both wrong*, based on comparison with the majority-vote human answer.

In the AT setting (Figure 1a), most models—including DEEPSEEK-REASONER, GPT-4, and QWQ-32B—achieve high agreement with human annotators, with a substantial proportion of cases falling into the *Both correct* category. However, DEEPSEEK-V3 and QWEN2.5-7B exhibit relatively higher numbers of *Human only correct* cases, suggesting challenges in precise recovery of masked time expressions.

In contrast, under the MT setting (Figure 1b), all models show performance degradation. The number of *Both correct* cases drops notably, while *Model only correct* and *Both wrong* instances increase. This trend highlights the difficulty of temporal reasoning in contexts with ambiguous or relative time expressions. Models such as QWEN2.5-7B and DEEPSEEK-V3 especially struggle, with many instances correctly identified by humans but missed by the models.

These findings underscore the value of probing model behavior across both structured and ambiguous temporal settings, as performance under absolute time does not necessarily generalize to more naturalistic, mixed-time scenarios.

Quantifying Human–Model Alignment To complement the categorical outcome analysis, we further quantify the degree of alignment between models and human annotators using Kendall's τ. In addition to correlations with gold orders (Section 4.1), we directly compute the correlation between each model prediction and each individual annotator's sequence, averaging across annotators. This provides a finer-grained measure of human–model agreement beyond majority-vote correctness.

Table 6 summarizes the results under both AT and MT conditions. Table 6a reports Kendall's τ for model–gold and human–gold comparisons, while Table 6b presents direct model–human correlations. Several patterns emerge: (1) frontier models such as DEEPSEEK-REASONER, GPT-4, and QWQ-32B show strong alignment with both gold orders and human judgments in the AT setting; (2) QWEN2.5-7B consistently lags behind, particularly in the MT condition, reflecting its difficulty with ambiguous or relative time references; and (3) across most models, direct model–human correlations are slightly lower than model–gold ones (e.g., GPT-4, DEEPSEEK-V3), highlighting that high accuracy with respect to gold standards does not always translate into close alignment with human reasoning.

System	AT τ	MT τ
DEEPSEEK-REASONER	**0.84**	**0.77**
QWQ-32B	0.82	0.76
GPT-4	0.75	0.65
DEEPSEEK-V3	0.68	0.68
GPT-3.5	0.51	0.45
QWEN2.5-7B	0.37	0.33
Annotator 1	0.80	0.55
Annotator 2	0.75	0.62
Annotator 3	0.83	0.67

(a) Model–Gold and Human–Gold Kendall's τ

System	AT τ	MT τ
DEEPSEEK-REASONER	**0.84**	**0.59**
QWQ-32B	0.83	0.58
GPT-4	0.72	0.46
DEEPSEEK-V3	0.71	0.53
GPT-3.5	0.52	0.39
QWEN2.5-7B	0.40	0.28

(b) Direct Model–Human Kendall's τ

Table 6: Human–model agreement measured by Kendall's τ across AT and MT settings.

4.3 Probing Model Temporal Inference via Time Masking

To investigate whether LLMs can maintain temporal reasoning ability when explicit time anchors are removed, we conduct a masked time prediction experiment in both AT and MT settings. In each instance, a single year expression is masked in one of the event sentences, and models are prompted to output the correct chronological order of the events (as described in Section 3.5). This setup isolates the model's reliance on explicit temporal cues and evaluates its ability to infer event order from context alone.

Results. Table 7 summarizes model performance under the masked time probing task. In the AT setting, GPT-4 achieves an Exact Match (EM) of only 0.033, with a negative average Kendall's τ

of -0.059, indicating that its predicted orders are slightly worse than random. QwQ-32B performs comparably, while DeepSeek-Reasoner completely fails (EM = 0). The MT setting is even more challenging: all models fail to recover any correct orders (EM = 0), and Kendall's τ approaches zero or negative, reflecting near-random or even inversely correlated rankings.

Setting	Model	EM	Avg Kendall τ
AT	GPT-4	0.033	-0.059
	QwQ-32B	0.033	-0.056
	DeepSeek-Reasoner	0.000	-0.006
MT	GPT-4	0.000	-0.026
	QwQ-32B	0.000	-0.050
	DeepSeek-Reasoner	0.000	0.015

Table 7: Performance of LLMs under the masked time probing task. EM denotes the fraction of exact matches to the gold event order; Kendall's τ measures correlation between predicted and gold orders.

Analysis and Insights.

Our probing results reveal three key insights:

1. **Strong reliance on explicit temporal anchors.** Masking just a single year drastically degrades performance, indicating that models primarily leverage surface-level time expressions rather than robust event reasoning.

2. **Failure to generalize in mixed-time narratives.** In MT settings, where relative or vague temporal expressions dominate, masking any remaining anchor causes complete failure, with EM = 0 for all models.

3. **Temporal reasoning collapses without explicit cues.** Negative or near-zero Kendall's τ scores suggest that model predictions are effectively random, and sometimes even inversely correlated with the true order.

These findings demonstrate that current LLMs exhibit *shallow temporal reasoning*, heavily dependent on explicit timestamps. Our probing methodology thus exposes a critical weakness in LLM temporal inference, highlighting the necessity for models that can robustly infer event order in partially observed or naturally vague timelines.

5 Conclusion

This study examined the temporal reasoning abilities of frontier LLMs and human annotators under both AT and MT conditions, using a hybrid-time dataset that integrates absolute, relative, and event-anchored expressions. By combining controlled evaluation with a masked-time probing task, we identified clear performance gaps and reasoning patterns: while frontier models can approach human-level ordering accuracy in AT settings, both humans and models struggle when explicit temporal anchors are reduced or removed. The masking experiments further revealed a strong dependence on explicit cues, with ordering accuracy collapsing when even a single anchor is missing.

From a broader perspective, these findings highlight that temporal reasoning in naturally occurring narratives is shaped by more than surface-form date recognition: it requires resolving vague relative expressions, reconciling narrative sequencing with chronological order, and integrating long-distance discourse anchors. The hybrid-time design adopted here provides a reusable, language-agnostic framework for diagnosing such challenges, and can be readily adapted to other languages and narrative genres.

Future work will expand the dataset to multilingual settings, enabling cross-linguistic comparisons of temporal reasoning strategies. We also aim to explore model architectures and training objectives that promote robust integration of explicit and implicit temporal cues, moving towards temporally aware systems capable of handling the complexities of real-world discourse.

Acknowledgements

Feifei Sun acknowledges the support of the China Scholarship Council (CSC) under scholarship number 202408050023.

This work was also supported partly by JSPS KAKENHI Grant Number JP25H00459. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the author(s)' organization, JSPS or MEXT.

We thank all reviewers for their insightful and constructive feedback, which has helped us improve the clarity, precision, and completeness of this work.

Limitations

Our study has several limitations. First, the probing experiments are conducted on a small 100-sample subset to ensure human annotation feasibility and interpretability. While this allows for detailed human-model comparison, it limits the

statistical generalizability of our findings to larger-scale temporal reasoning tasks.

Second, our current probing approach masks only one time expression per passage and evaluates its impact on event ordering. This simplified design highlights model reliance on explicit temporal cues but does not fully capture the complexity of real-world narratives with multiple missing or ambiguous temporal anchors.

Third, our evaluation focuses on surface-level ordering accuracy (EM and Kendall's τ) and does not analyze intermediate reasoning steps or latent temporal representations. As a result, our conclusions about model reasoning are inferential and may not fully reveal the internal mechanisms driving model predictions.

Fourth, although we include multiple model families (GPT-4, DeepSeek, QwQ), our selection omits smaller or domain-specialized models, and we exclude QWEN2.5-7B, GPT-3.5-TURBO and DEEPSEEK-V3 from probing due to its low baseline performance.

Fifth, a potential concern is that our passages originate from Wikipedia biographies, raising the possibility that models may have partially memorized specific dates or event sequences during pretraining. To mitigate this risk, during dataset construction, we performed data cleaning and randomization procedures to reduce direct overlap with seen text (details described in our companion work (Sun et al., 2025)). In particular, events were extracted and re-ordered into new narrative contexts, such that models could not rely on surface recall of document-level sequences.

Nevertheless, we acknowledge that memorization cannot be entirely excluded, as noted in our Limitations section. Importantly, our evaluation requires models to reconstruct global event orders across passages: even if individual dates were known, successful performance hinges on reasoning over relative and hybrid time references rather than verbatim recall.

Finally, our design of the Masked-Time probing task primarily aims to stress-test model robustness when crucial temporal anchors are missing. As the masked expression may admit multiple plausible human interpretations, constructing a unique human "gold" reference would be problematic. Therefore, we did not collect human annotations for this setting. We acknowledge this as a limitation, since direct human-model comparison could further illuminate the gap in reasoning strategies.

Future work should expand the model coverage and explore more fine-grained temporal reasoning diagnostics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu. 2025. Perceive the passage of time: A systematic evaluation of large language model in temporal relativity. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8304–8313.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Jeremy R Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. *arXiv preprint arXiv:2303.12860*.

Xi Ding and Lei Wang. 2025. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1855–1868.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. 2025. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. *arXiv preprint arXiv:2505.15392*.

Slava Kalyuga. 2011. Cognitive load aspects of text processing. *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, pages 114–132.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv preprint arXiv:2505.13508*.

Qiang Ning, Zhili Feng, and Dan Roth. 2019. A structured learning approach to temporal relation extraction. *arXiv preprint arXiv:1906.04943*.

Guy D Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the fifteenth ACM international conference on Web search and data mining*, pages 833–841.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2020. Benchmarking machine reading comprehension: A psychological perspective. *arXiv preprint arXiv:2004.01912*.

Feifei Sun, Ziyi Tong, Houjing Wei, Cheng Peng, Teeradaj Racharak, and Le-Minh Nguyen. 2025. Benchmarking temporal reasoning: Can large language models navigate time when stories refuse to follow a straight line? In *First Workshop on Foundations of Reasoning in Language Models (FoRLM@NeurIPS)*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1963–1974.

A Prompt Example for Masked Time Prediction

Below is a representative one-shot prompt used in our masked time prediction (sentence reordering) experiment. The example demonstrates the format provided to language models during inference:

> `Here is an example: Input Sentences:`
>
> 1. He was born in 1960.
>
> 2. He graduated in 1980.
>
> 3. He joined IBM in 1990.
>
> 4. He became a manager in 2000.
>
> `Answer:` 1,2,3,4
>
> `Now reorder the following sentences in chronological order. Respond only with the sentence indices in the correct order, separated by commas.`
>
> `Input Sentences:`
>
> A campaign, called Save the Ampelmännchen, was launched by the public and Ampelmännchen enthusiasts, resulting in the preservation of Peglau's Ampelmännchen in [MASK].
>
> Karl Peglau died in Berlin, Germany, on 29 November 2009, at the age of 82.
>
> Karl Peglau was a German traffic psychologist who invented the iconic Ampelmännchen traffic symbols used in the former East Germany in 1961.
>
> Peglau designed the glass human figures for the stop (red) and go (green) lights on the traffic signal in 1961, which became known as the Ampelmännchen.
>
> `Answer: ___`

In this example, the first sentence contains a masked year expression. The model must infer its correct position within the *global event timeline*—the ordered sequence of events in the passage, whether represented as explicit dates (e.g., *1960* $\rightarrow$ *1980* $\rightarrow$ *1990* $\rightarrow$ *2000*) or as major life milestones (e.g., *birth* $\rightarrow$ *graduation* $\rightarrow$ *career start* $\rightarrow$ *promotion*)—by leveraging both local context and surrounding temporal cues.

Semantic Meaning or Script Shape? A Comparative Study of Cross-Lingual Transfer in mBERT and PIXEL

Zhenming Li
Kyushu Institute of Technology
Fukuoka, Japan
li.zhenming714@mail.kyutech.jp

Kazutaka Shimada
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

Abstract

Multilingual BERT (mBERT) has been extensively investigated for cross-lingual transfer learning (CLTL), achieving strong performance owing to its rich semantic representations. Recent studies have demonstrated that visually grounded models, such as PIXEL, can also be applied to CLTL by leveraging character-level glyph information. In this work, we present a comparative study of mBERT and PIXEL in a zero-shot cross-lingual transfer setting across 5 languages. Our results show that mBERT consistently outperforms PIXEL in overall accuracy, underscoring the effectiveness of semantic representations for CLTL. Nevertheless, we find that PIXEL exhibits competitive performance for visually similar language pairs and maintains robustness when semantic information is limited, suggesting the usefulness of visual information in cross-lingual transfer scenarios.

1 Introduction

Cross-lingual transfer learning (CLTL) aims to exploit knowledge acquired in a source language to improve performance in a target language. multilingual BERT (mBERT) (Devlin et al., 2019), has been widely explored in CLTL tasks and shows great performance. mBERT adopts a multilingual subword tokenization strategy and learns shared semantic representations from large-scale multilingual corpora, thereby enabling robust cross-lingual generalization. It is well established that semantic-based models, such as mBERT, can serve as effective backbones for CLTL tasks.

Complementary to these token-based approaches, recent work in Visual NLP explores the potential of processing written language through its visual form, rather than as sequences of discrete tokens. A notable example is PIXEL (Rust et al., 2022), a tokenizer-free language model that processes text as images. PIXEL renders text into fixed-size grayscale images, partitions them into non-overlapping patches, and processes these patches with a Vision Transformer (ViT) backbone trained using a masked patch prediction objective. By bypassing language-specific tokenization, PIXEL inherently supports script-agnostic processing and facilitates cross-lingual transfer without relying on a shared subword vocabulary. Empirical results across multiple multilingual benchmarks (Rust et al., 2022) show that PIXEL achieves competitive performance, demonstrating robustness to diverse scripts, orthographic variation, and visual distortions. These findings suggest that visual-based models such as PIXEL can also serve as effective backbones for CLTL tasks.

Given that both semantic-based models and vision-based models are capable of supporting transfer learning, in this work, we ask a simple question: Is cross-lingual transfer learning more effective when models focus on semantic meaning, like mBERT, or when they focus on the visual form of text, such as the shapes of scripts and glyphs, like PIXEL? To answer this question, we compare semantic-based and shape-based transfer in a CLTL setting. In the setting, the source task is sentiment classification in five languages: Chinese, Japanese, English, French, and Spanish, while the target task is Chinese offensive language classification. We conduct experiments with both mBERT and PIXEL, followed by a comparative analysis of their performance across model types and writing scripts.

Our key findings are summarized as follows:

1. mBERT outperforms PIXEL in overall accuracy, indicating the superior effectiveness of semantic information in transfer learning.

2. PIXEL transfers effectively for visually similar language pairs and performs robustly un-

der limited semantic information by exploiting glyph cues.

2 Related Work

2.1 Cross-Lingual Transfer learning

Transfer learning (Pan and Yang, 2010) is a widely adopted machine-learning methodology in which a model trained on a source task or domain is reused and adapted to improve performance on a related target task or domain. Cross-lingual transfer learning (CLTL) constitutes a transfer learning paradigm that emphasizes the transfer of knowledge between two distinct languages. This approach has been recognized as an effective framework for tasks such as offensive language detection. Ranasinghe and Zampieri (2020) and Montariol et al. (2022) employed zero-shot cross-lingual transfer in their respective studies, demonstrating its capability to facilitate knowledge transfer to previously unseen languages. However, Nozza (2021) highlights that such zero-shot transfer may incur performance degradations and be susceptible to cultural and lexical pitfalls, particularly when transferring from English to other languages without access to labeled data in the target language. In contrast, De la Peña Sarracén et al. (2023), Röttger et al. (2022), and Caselli and Plaza-del Arco (2025) investigated the few-shot setting, wherein the model is provided with a limited number of labeled examples in the target language to enhance adaptation and mitigate transfer-related deficiencies.

The relationship between the source and target languages has been extensively examined in the context of CLTL. Blaschke et al. (2025) analyzed 263 languages across three NLP tasks and conclude that choosing the similarity measure based on the task is important, with lexical similarity being most predictive for lexicon-heavy tasks. Lim et al. (2024) examined multiple source–target setups and found that multi-source transfer performs best when at least one typologically close language is combined with several diverse sources, while random selection can be suboptimal. Lin et al. (2024) introduced a model-embedding-based similarity metric and showed it predicts cross-lingual transfer performance better than typology-based metrics, especially in low-resource settings. Prior findings indicate that higher similarity between the source and target languages can enhance the effectiveness of cross-lingual transfer.

2.2 Vision Transformer

As we introduced in section 1, PIXEL (Rust et al., 2022) is one of the representatives of Vision Transformer (ViT)-based models. Original PIXEL only pretrained on English corpora, PIXEL-M4 (Kesen et al., 2025) extended the approach through multilingual pretraining on four visually and linguistically diverse languages: English, Hindi, Ukrainian, and Simplified Chinese, yielding substantially improved performance and cross-script transfer in non-Latin scripts compared to its monolingual counterparts. The Vision Transformer (ViT), originally proposed by Dosovitskiy et al. (2020), has inspired a wide range of subsequent works, including ViLT (Kim et al., 2021), ALBEF (Li et al., 2021), and PaLI (Chen et al., 2022).

Vision Transformer (ViT) concepts have been increasingly applied in multimodal NLP tasks. Kim et al. (2021) proposed ViLT, a convolution-free vision-and-language model that achieves competitive performance on Visual Question Answering, image–text retrieval, and visual reasoning, while being substantially more efficient than prior visual models. Ganz et al. (2024) presented QA-ViT, a Question-Aware Vision Transformer that embeds question-specific awareness directly within the vision encoder, allowing dynamic visual feature adaptation to queries and achieving consistent improvements across various multimodal reasoning tasks. Chochlakis et al. (2022) introduced VAuLT, extending ViLT with BERT to enhance semantic representations in multimodal sentiment analysis, improving sentiment prediction accuracy.

3 Zero-shot CLTL from Sentiment to Offensiveness

In this study, we explain the details of zero-shot cross-lingual transfer learning from sentiment analysis to offensive language detection across multiple languages, as illustrated in Figure 1.

First of all, our experimental design follows the inductive transfer learning paradigm defined by Pan and Yang (2010), wherein the source and target tasks differ. We focus on the transfer from sentiment classification (source task) to offensive language detection (target task). In sentiment classification, each sentence is labeled as either “positive” or “negative”. In offensive language detection, sentences are labeled as either “non-offensive” or “offensive”. We adopt a label alignment assumption, mapping negative sentiment to the offensive

Dataset	language	offensive	non offensive	total
COLD	Chinese	18041	19439	37480
Dataset	**language**	**negative**	**positive**	**total**
WeiboSenti	Chinese	10000	10000	20000
WRIME	Japanese	11604	10834	22438
Sentiment140	English	9980	9968	19948
French_multi	French	10000	10000	20000
Spanish_multi	Spanish	9994	9976	19970

Table 1: Data statistics of datasets in the experiment

class and positive sentiment to the non-offensive class. This mapping is consistent with prior work leveraging sentiment features for offensive or sarcasm language detection (Islam, 2024; Husain and Uzuner, 2021).

Under this alignment, we fine-tune multilingual models on sentiment datasets and evaluate them directly on offensive language detection without exposure to offensive data during training, constituting a zero-shot transfer setting. We employ sentiment datasets in 5 source languages: Chinese, Japanese, English, French, and Spanish. Evaluation is performed exclusively on a Chinese offensive language test set.

For each experiment, we fine-tune a model using a single source language, without combining datasets across languages. In addition to zero-shot transfer, we establish a supervised baseline by fine-tuning the models on Chinese offensive language training data and evaluating them on the same Chinese offensive language test set.

We experiment with two multilingual pre-trained models: mBERT and multilingual PIXEL (PIXEL-M4). Both models are pre-trained on corpora including Chinese and English; however, PIXEL-M4 does not include Japanese, French, or Spanish in its pretraining, whereas mBERT does. Hyperparameter configurations for fine-tuning both models are reported in Table 5 and Table 6 in the Appendix A. Fine-tuning parameters are kept consistent across all source languages and the baseline.

4 Data Collection

We now describe the datasets used in our experiments. Our setup requires both offensive language data for target evaluation and multilingual sentiment datasets for source languages.

For the Chinese offensive language dataset, we utilize the COLD dataset (Deng et al., 2022), a publicly available corpus comprising 37,480 social media comments annotated with binary offensive labels. Following the original data partitioning scheme proposed by the authors, we divide the corpus into training, development, and test subsets. For the baseline configuration, the mBERT and PIXEL models were trained on the training set, with hyperparameters optimized using the development set of the offensive language dataset, while the test set is reserved exclusively for final evaluation. The test set comprises 5,323 instances from the offensive language dataset and is consistently employed across both the baseline and all transfer learning settings.

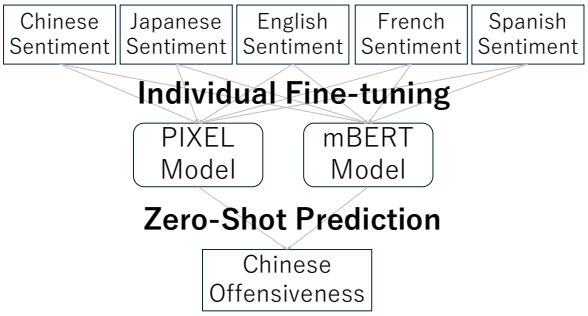


Figure 1: Zero-shot CLTL from Sentiment to Offensiveness. We fine-tune the PIXEL and mBERT with sentiment data in different languages individually and conduct zero-shot prediction on Chinese offensiveness.

Then we introduce the sentiment classification datasets for five languages. The Chinese sentiment data is drawn from the work of (Wan et al., 2020), named WeiboSenti. For Japanese, we adopt the WRIME (Kajiwara et al., 2021) dataset, which provides labeled sentiment annotations for Japanese texts. English sentiment data is obtained from the widely used Sentiment140 (Go et al., 2009) dataset, developed by a research team at Stanford University. For French and Spanish, we aggregate samples from multiple publicly available sources, guided by documentation provided by

Language	PIXEL-M4	mBERT
Baseline	0.7367	0.7950
Chinese	0.4885(▼0.2482)	0.5784(▼0.2166)
English	0.4708(▼0.2659)	0.5533(▼0.2417)
Japanese	0.4582(▼0.2785)	0.5525(▼0.2425)
French	0.3900(▼0.3467)	0.6084(▼0.1866)
Spanish	0.3472(▼0.3895)	0.4880(▼0.3070)

Table 2: F1-Score (Macro Average for two classes) for PIXEL-M4 and mBERT on zero-shot sentiment transfer. The symbol ▼ denotes a performance decrease relative to the Baseline.The baseline is trained with the target data set, namely the offensive language data. In other words, this essentially corresponds to the upper bound for this target task against zero-shot approaches from sentiment analysis.

Brand24/mms (Łukasz Augustyniak et al., 2023) and named them French_multi and Spanish_multi. The French data is drawn from the datasets of (Narr et al., 2012; Keung et al., 2020). The Spanish data is sourced from datasets of (Cruz et al., 2008; Keung et al., 2020; Keith Norambuena et al., 2019; Patwa et al., 2020; Mozetič et al., 2016).

To ensure fair cross-lingual comparison, we control the size of each sentiment dataset to approximately 20,000 examples per language. For Chinese and English, where the sentiment data originates from a single source, we randomly sample subsets from the original datasets. For French and Spanish, where data is drawn from multiple sources, we allocate samples proportionally to achieve a total of 20,000 instances per language. For Japanese, we use the entire original dataset, as its size is already close to the target. During data collection, we balance the label distribution within each dataset to achieve an approximately 1:1 ratio between positive and negative sentiment classes, thereby mitigating potential bias from class imbalance in training. After data collection, we apply a simple preprocessing pipeline to remove hashtags, emojis, and URLs. Some instances are discarded during this process, and the final dataset statistics used in our experiments are reported in Table 1.

We follow the original data splits provided by the authors for COLD, using their train, evaluation, and test sets. For the sentiment analysis datasets (WeiboSenti, WRIME, Sentiment140, French_multi, and Spanish_multi), we partition the data into training (80%) and evaluation (20%) subsets, as no test data of sentiment is required.

5 Experiment Results and Analysis

The experimental results are presented in Table 2. We conduct 3 runs under the same experimental setting and report the average performance across 3 runs as the final result. We report only the macro-averaged F1 score over the two classes (offensive and non-offensive) in Table 2 for analysis. Additional details, including training statistics and other classification metrics, are provided in Table 7 in Appendix B and Table 8 in Appendix C. The Baseline setting corresponds to fine-tuning the models on the COLD training set and evaluating them on the COLD test set. Names of the languages correspond to the source sentiment datasets in Table 1.

Across all transfer scenarios, we observe a decrease in performance relative to the baseline. This performance degradation can be attributed to task mismatch or language differences between the source and target. For example, a conceptual and label mismatch leads to inevitable information loss: 0.2482 for PIXEL-M4 and 0.2166 for mBERT, even in the same language transfer setting from Chinese sentiment to Chinese offensiveness.

Comparing the two models, we first observe that the baseline performance of mBERT surpasses that of PIXEL. Furthermore, mBERT consistently exhibits smaller performance decrease than PIXEL-M4. For instance, in the English case, the decrease is 0.2417 for mBERT compared to 0.2659 for PIXEL-M4, and this pattern holds for all languages. This observation suggests that semantic information may play a more robust and critical role in cross-lingual transfer than visual information.

Although PIXEL-M4 underperformed in terms of F1-score, this outcome does not necessarily imply that visual information is useless in transfer learning. From a cross-lingual perspective, languages may share substantial visual similarity in their orthographic forms. For example, Chinese and Japanese share a large inventory of Chinese characters, many of which are identical or visu-

Dataset	original	replaced with radical	label
COLD	真恶心啊那个男的 (That guy is disgusting)	具亚心口二丨力勺	offensive
WeiboSenti	人口就是多国庆出门 (Many people go out on National Day)	人口京日夕口大凵门	positive
WRIME	友達コロナ疑惑とか焦るー (My friend is worried about coronavirus)	又辶コロナ疋心とか灬るー	negative

Table 3: Examples of replacement with radicals in Chinese and Japanese datasets. Inside the () is the meaning of the sentences.

Dataset	PIXEL-M4	mBERT
COLD	0.7367	0.7950
COLD_radicals	0.5360 (▼ 0.2007)	0.4789 (▼ 0.3161)
WeiboSenti	0.4885	0.5784
WeiboSenti_radicals	0.4670 (▼ 0.0215)	0.4368 (▼ 0.1416)
WRIME	0.4582	0.5525
WRIME_radicals	0.5002 (▲ 0.0420)	0.5287 (▼ 0.0238)

Table 4: F1-Score (Macro Average for two classes) for Ablation study: replacement with radicals. The symbol ▼ denotes a performance decrease relative to the no replacement. The symbol ▲ denotes performance increase.

ally similar. Likewise, English, French, and Spanish employ closely related Latin scripts with only minor orthographic variations. Focusing on the PIXEL-M4 model alone, we find evidence that visually similar languages can still yield better results. PIXEL-M4 was pretrained on Chinese and English; accordingly, we compare transfer performance from these two languages and observe that transfer from Chinese sentiment achieves higher accuracy than from English (0.4885 vs. 0.4708). For languages not seen in PIXEL-M4 pretraining: Japanese, French, and Spanish, Japanese yields the highest performance (0.4582), followed by French (0.3900) and Spanish (0.3472). These results indicate that, for PIXEL-M4, transferring from visually similar languages (e.g., Chinese, Japanese) to Chinese offensiveness classification tends to be more effective.

In summary, while mBERT appears more robust overall likely due to its stronger semantic representation, results of PIXEL-M4 demonstrate that visual similarity remains a contributive factor in cross-lingual transfer. We further conduct a paired t-test and a Wilcoxon signed-rank test to assess the significance of the performance difference between the two models. For the paired t-test, the results are $t = -4.8359$ and $p = 0.0047$, while for the Wilcoxon signed-rank test, the results are $W = 0.0$ and $p = 0.0312$. Both tests indicate a statistically significant difference at the 0.05 level, thereby demonstrating the robustness and reliability of the observed results.

6 Ablation Study on Visual Similarity

As discussed in the previous section, visual similarity can be a contributing factor in cross-lingual transfer, though it is not decisive as semantic representation. We now investigate an extreme scenario in which the semantic content of sentences is removed, to assess whether visual information alone can still facilitate transfer learning.

To this end, we conduct an ablation study by replacing all Chinese characters with their corresponding radicals. A radical in Chinese is a subcharacter component that often contributes to the meaning or pronunciation of the full character, but is not sufficient on its own to convey the original word’ s semantic meaning. Importantly, radicals preserve substantial visual information, such as stroke patterns, visual structure, and positional arrangement, thereby maintaining a high degree of script-level similarity to the original characters even when semantic content is removed.

We apply this replacement procedure to all 3 datasets containing Chinese characters, namely COLD, WeiboSenti, and WRIME. It is worth noting that a single Chinese character (kanji in Japanese) may contain multiple components that can be interpreted as radicals. In our experiments, we employ the RadicalFinder module from the

cjkradlib [1] Python library and select the first candidate radical for a Chinese character suggested by the tool. For COLD, we replace only the training set while keeping the original test set unchanged. For WeiboSenti and WRIME, we replace the entire datasets, as no test data is required from these datasets. We replace only the texts, while preserving their original sentiment and offensiveness labels. Some examples of the replacement are illustrated in Table 3. The modified sentences become semantically uninterpretable while preserving visual similarity to the original languages. These radical scripts can be regarded as a pseudo-language devoid of semantic content, and thus treated as the source languages in the zero-shot CLTL framework described in Section 3. For COLD, We fine-tune the models on the radical-based training data and evaluate them on the original test set. For WeiboSenti and WRIME, we apply the same procedure as in Section 3, fine-tuning on radical-based scripts and evaluating on the original COLD test set.

The experimental results are presented in Table 4. Performance is measured using the macro-averaged F1-score over the two classes (offensive and non-offensive), consistent with Table 2. The modified pseudo-scripts data are named _radicals after their original datasets. We observe that replacing characters with radicals generally leads to performance degradation across all three datasets for both models. For instance, on COLD, PIXEL-M4 experiences a decrease of 0.2007, while mBERT suffers a greater decrease of 0.3161. All settings show performance decreases, except for WRIME with PIXEL-M4, where we observe a slight improvement (+0.0420). These findings reinforce the conclusion that semantic meaning is critical for both PIXEL-M4 and mBERT. Although PIXEL-M4 is effective in capturing script-level features, it remains susceptible to semantic information loss, which in turn degrades performance.

Interestingly, when comparing models within each dataset, we find the opposite trend from the previous section: PIXEL-M4 consistently exhibits less performance decrease than mBERT in the case of radical replacement. For example, in WeiboSenti_raidcals, the decrease of PIXEL-M4 is only 0.0215, whereas the decrease of mBERT is 0.1416. In the WRIME_radicals, the performance of PIXEL-M4 even improves with the radical replacement, which warrants further investigation. These results suggest that when semantic meaning is absent, visual similarity can still be leveraged for transfer learning, and visual-based models such as PIXEL are more robust in such cases. Consequently, visual similarity between source and target languages may serve as an important criterion when selecting language pairs for cross-lingual transfer, particularly in scenarios where semantic information is scarce or unavailable.

[1] https://pypi.org/project/cjkradlib/

7 Conclusion

In this paper, we examine the impact of transfer learning on two multilingual models: the semantic-oriented mBERT and the visual-oriented PIXEL. We adopt a zero-shot cross-lingual transfer setting across five languages to compare their performance. Experimental results indicate that semantic information constitutes a reliable and effective basis for transfer learning, outperforming purely visual cues. However, a closer analysis of PIXEL reveals that it facilitates transfer particularly well between visually similar source and target language pairs, suggesting that visual information remains a non-negligible factor in cross-lingual transfer. Furthermore, our ablation study shows that even in the absence of semantic information, PIXEL can achieve robust transfer performance. This finding highlights the potential of visual information as a viable alternative for transfer learning in scenarios where semantic information is limited. Our current work focuses exclusively on standard quantitative classification metrics such as the F1 score. For future work, we aim to conduct a detailed error analysis to identify qualitative linguistic patterns (e.g, failure cases and script-specific phenomena), thereby providing deeper insights into the role and impact of visual information.

Limitations

This study has several limitations. Most notably, the sentiment-to-offensiveness transfer setting relies on the simplifying assumption that offensive language is conceptually equivalent to negative sentiment. While there is empirical overlap between the two, this assumption may not capture the full complexity of offensive language, which can include sarcastic, provocative, or contextually ambiguous expressions that are not necessarily associated with negative sentiment. Future work should aim to develop stronger transfer settings, poten-

tially involving intermediate tasks, richer label taxonomies, or adversarial adaptation strategies, to more accurately bridge the gap between sentiment and offensiveness.

Moreover, in the current study, we focus exclusively on the case of sentiment-to-offensiveness transfer across five languages. To enable a more comprehensive analysis, future work should extend the experiments beyond this setting to encompass additional transfer scenarios and a broader range of languages.

Acknowledgments

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2133.

References

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Tommaso Caselli and Flor Miriam Plaza-del Arco. 2025. Learning from disagreement: Entropy-guided few-shot selection for toxic language detection. In *Proceedings of The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 53–66, Vienna, Austria. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, and Shrikanth Narayanan. 2022. Vault: Augmenting the vision-and-language transformer for sentiment classification on social media. *arXiv preprint arXiv:2208.09021*.

Fermin L Cruz, Jose A Troyano, Fernando Enriquez, and Javier Ortega. 2008. Experiments in sentiment classification of movie reviews in spanish. *Procesamiento del Lenguaje Natural*, 41:73–80.

Gretel De la Peña Sarracén, Paolo Rosso, Robert Litschko, Goran Glavaš, and Simone Ponzetto. 2023. Vicinal risk minimization for few-shot cross-lingual transfer in abusive language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4069–4085, Singapore. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13861–13871.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Fatemah Husain and Ozlem Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 364–369.

Khondoker Ittehadul Islam. 2024. Leveraging sentiment for offensive text classification. *arXiv preprint arXiv:2412.17825*.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Brian Keith Norambuena, Exequiel Lettura, and Claudio Villegas. 2019. Sentiment analysis and opinion mining applied to scientific paper reviews. *Intelligent Data Analysis*, 23:191–214.

Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. *ArXiv*, abs/2505.21265.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. *Preprint*, arXiv:2402.13562.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, Andre Martins, and Hinrich Schuetze. 2024. mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, St. Julian's, Malta. Association for Computational Linguistics.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.

Sascha Narr, Michael Hülfenhaus, and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. In *Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012)*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2022. Language modelling with pixels. *ArXiv*, abs/2207.06991.

Shuo Wan, Bohan Li, Anman Zhang, Wenhuan Wang, and Donghai Guan. 2020. S2ap: Sequential sentiweibo analysis platform. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part III*, page 745–749, Berlin, Heidelberg. Springer-Verlag.

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. *Preprint*, arXiv:2306.07902.

A Fine-tuning Parameters

Parameters	Values
Rendering backend	PyGame
Classification head pooling	Mean
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	0
Learning rate	3e-5
Learning rate warmup steps	100
Learning rate schedule	Linear decay
Max sequence length	529
Batch size	64
Max steps	15000
Early Stopping	Active
Eval interval	100 steps
Dropout probability	0.1

Table 5: Fine-tuning parameters for PIXEL models.

Parameters	Values
Classification head pooling	CLS embedding
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	0
Learning rate	3e-5
Learning rate warmup steps	100
Learning rate schedule	Linear decay
Max sequence length	256
Batch size	64
Max steps	15000
Early Stopping	Active
Eval interval	100 steps
Dropout probability	0.1

Table 6: Fine-tuning parameters for mBERT models.

B Best Evaluation F1 during Training

Datasets	mBERT	PIXEL-M4
WeiboSenti	0.8845	0.7851
WRIME	0.8527	0.7304
Sentiment140	0.8039	0.6657
French_multi	0.9231	0.7960
Spanish_multi	0.7782	0.6878

Table 7: Best F1 score recorded when evaluating the training effect using sentiment evaluation data. The checkpoint corresponding to the best F1 score is saved at the end of training.

C Full Metrics for Experiment Results

Dataset	Model	Precision (Mean)	Precision (Std)	Recall (Mean)	Recall (Std)	F1-score (Mean)	F1-score (Std)
COLD	mbert	0.7961	0.0030	0.8090	0.0034	0.7950	0.0046
COLD	pixel	0.7395	0.0051	0.7501	0.0055	0.7367	0.0062
WeiboSenti	mbert	0.6651	0.0045	0.6375	0.0137	0.5784	0.0259
WeiboSenti	pixel	0.5972	0.0067	0.5683	0.0082	0.4885	0.0181
Sentiment140	mbert	0.5621	0.0112	0.5589	0.0177	0.5533	0.0157
Sentiment140	pixel	0.5109	0.0119	0.5105	0.0108	0.4708	0.0698
WRIME	mbert	0.6591	0.0046	0.6221	0.0053	0.5525	0.0109
WRIME	pixel	0.4986	0.0154	0.5002	0.0065	0.4582	0.0357
French_multi	mbert	0.6214	0.0099	0.6241	0.0095	0.6084	0.0032
French_multi	pixel	0.4551	0.0306	0.4966	0.0010	0.3900	0.0097
Spanish_multi	mbert	0.6841	0.0035	0.5972	0.0249	0.4880	0.0497
Spanish_multi	pixel	0.5909	0.0306	0.5178	0.0145	0.3472	0.0557

Table 8: Macro Avg Metrics: Mean and Standard Deviation(Std) of Precision, Recall, and F1 score over 3 Runs of the CLTL experiment.

Reference Points in LLM Sentiment Analysis: The Role of Structured Context

Junichiro Niimi[1,2]
Faculty of Business Management, Meijo University,
1-501, Tempaku-ku, Nagoya, Aichi 4688502, Japan
RIKEN AIP, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
jniimi@meijo-u.ac.jp

Abstract

Large language models (LLMs) are now widely used across many fields, including marketing research. Sentiment analysis, in particular, helps firms understand consumer preferences. While most NLP studies classify sentiment from review text alone, marketing theories, such as prospect theory and expectation–disconfirmation theory, point out that customer evaluations are shaped not only by the actual experience but also by additional reference points. This study therefore investigates how the content and format of such supplementary information affect sentiment analysis using LLMs. We compare natural language (NL) and JSON-formatted prompts using a lightweight 3B parameter model suitable for practical marketing applications. Experiments on two Yelp categories (Restaurant and Nightlife) show that the JSON prompt with additional information outperforms all baselines without fine-tuning: Macro-F1 rises by 1.6% and 4% while RMSE falls by 16% and 9.1%, respectively, making it deployable in resource-constrained edge devices. Furthermore, a follow-up analysis confirms that performance gains stem from genuine contextual reasoning rather than label proxying. This work demonstrates that structured prompting can enable smaller models to achieve competitive performance, offering a practical alternative to large-scale model deployment.

1 Introduction

1.1 Background

In recent years, with the rapid advancements in large language models (LLMs; Brown et al., 2020), both industrial and academic areas in wide range of domains have utilized LLMs for data analytics, automation, and decision support. In particular, due to their high applicability in textual data, many studies have implemented sentiment analysis using LLMs (cf. Krugmann and Hartmann, 2024).

LLMs indeed demonstrate remarkable capabilities in understanding textual context; however, the actual 'context' is referred to as a relationship between the tokens, which is captured through Transformer (Vaswani et al., 2017) and attention mechanisms (Bahdanau et al., 2015) and most existing approaches limit their analysis to the linguistic context within review texts alone. Regarding real-world marketing applications, the actual context of consumer evaluation contains the factors which extend far beyond the written review, such as past purchasing patterns, prior experiences with the business, comparative evaluations against competitors, and opinions from social media.

This gap is particularly relevant in customer relationship management (CRM; Oliver, 1999; Reinartz et al., 2004), where understanding customer sentiment accurately drives business decisions. Marketing research has long established through prospect theory (Kahneman and Tversky, 2013) and expectation-disconfirmation theory (EDT; Oliver, 1980) that consumers evaluate experiences relative to these broader reference points. This insight remains largely unexplored in LLM-based sentiment analysis.

Furthermore, practical deployment, particularly for real-time recommendation, has two critical challenges. First, regarding computational efficiency, many business cannot deploy large-scale models (over 70 billion parameters) due to latency and infrastructure constraints. Second, despite having rich contextual data (e.g., user contents, browsing history), current methods cannot efficiently incorporate this information into LLM-based sentiment analysis.

1.2 Research Gap

Despite LLMs' ability to process diverse input formats, sentiment analysis studies predominantly focus on review text alone (e.g., You et al., 2015; Flanagin and Metzger, 2013; Sparks and Brown-

ing, 2011). However, real-world platforms possess rich contextual information. From these gaps, we sequentially derive four research questions (RQs 1–4):

RQ1 Reference-point utilization: Does supplying user- and business-average ratings actually help an LLM classify sentiment more accurately?

RQ2 Prompt format: If the same information is presented in a machine-readable structure or plain texts, does the prompt format affect the model performance?

RQ3 Proxy effect: If supplying the reference points improve accuracy, is it due to their implicit encoding of the ground-truth labels?

RQ4 Reference interactions: How do interactions between multiple reference points affect prediction accuracy?

We address these RQs by three experimental studies. In Study 1, we set up two approaches for the prompts: natural-language (NL) and machine-readability (JSON), and several combinations of contextual factors: user average (U), business average (B), and other attributes (O). We compare the model performance across these models. In Study 2, we test whether such information reflect the label; average ratings may act as a proxy of ground truth. Finally, in Study 3, we further examine how accuracy changes according to the interactions of those two reference points. We progressively explore not only whether reference points improve performance, but also how they function within the model's inference process.

The remainder of this study is constituted as follows: Section 2 reviews related studies, Section 3 outlines the model construction, and Section 4 presents empirical analyses. We discuss the key findings and implications in Section 5. Finally, we list our research limitations in Section 6.

2 Related Study

2.1 Sentiment Analysis

Sentiment analysis has been conducted with various methodologies, including lexicon-based models (Hutto and Gilbert, 2014), machine learning approaches with embeddings (Mikolov et al., 2013; Bojanowski et al., 2017), and deep neural networks such as BERT and RoBERTa (Devlin et al., 2018; Liu et al., 2019).

Recently, LLM-based approaches have gained attention for sentiment analysis (Krugmann and Hartmann, 2024). Models like GPT (Brown et al., 2020; OpenAI, 2023a) and Llama (Touvron et al., 2023a,b; Grattafiori et al., 2024) demonstrate superior performance compared to fully-supervised models and even fine-tuned RoBERTa (Wang et al., 2024; Krugmann and Hartmann, 2024). Key advantages include broad applicability due to pre-training and ability to process raw text without extensive preprocessing, achieving high accuracy without fine-tuning.

Other approaches to sentiment analysis using LLMs include aspect-based sentiment analysis (AbSA; Do et al., 2019; Nazir et al., 2022), which simultaneously predicts multiple aspects in reviews such as price and service quality. Ensemble approaches combine the decisions of multiple LLMs to create the robust model (Xing, 2025; Huang et al., 2024; Niimi, 2025; Chen et al., 2025).

However, existing approaches, including AbSA and ensemble methods, predominantly focus on review text alone. While this concentration highlights interesting challenges from the viewpoint of NLP, it does not necessarily guarantee practical utility for marketing or business decision-making, where contextual information beyond the textual modality plays a crucial role. Although some multimodal sentiment analysis utilizes the other modalities such as images and audio (Das and Singh, 2023; Gandhi et al., 2023), they rarely address psychological reference points, such as prior expectations expressed through numeric values.

2.2 A Role of Reference Point in Service Evaluation

In the field of marketing, extensive research has examined how consumers evaluate the product and service quality. Notable frameworks include prospect theory and EDT. Prospect theory posits that consumers evaluate services by comparing their actual experience to a pre-established reference point (Kahneman and Tversky, 2013). If their experience falls short of this reference point, they tend to feel dissatisfied; conversely, if it exceeds the reference point, satisfaction is more likely. Furthermore, EDT explains the evaluation process from two perspectives. Absolute evaluation involves assessing whether the perceived quality meets a fixed standard, while relative evaluation is based on com-

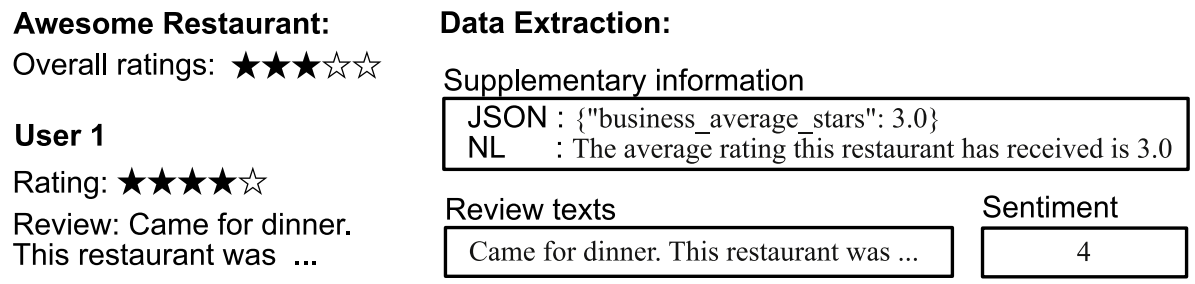


Figure 1: Sample extraction from the dataset.

paring prior expectations with the perceived quality (Oliver, 1980). In both cases, prior expectations play a significant role in determining overall customer satisfaction.

Prior studies on a wide range of products and services have examined the factors that affect or shape expectations. Some of the key factors include consumers' past experiences (Oliver, 1980; Kopalle and Lehmann, 2001; Bolton, 1998; Cooil et al., 2007), reputations provided by other customers (Keiningham et al., 2015; Ryu et al., 2008; Babin et al., 2005), and perceived value (Ryu et al., 2008; Qin and Prybutok, 2008). In particular, reputations—such as an average rating and the helpful reviews from other consumers—serve as important reference points when comparing prior expectations with actual experiences.

However, as noted above, because sentiment analysis is predominantly based on actual review texts, few existing studies have taken these supplementary factors into account. To more accurately capture customers' preferences, it is crucial to incorporate such information into sentiment analysis. We therefore propose a framework that effectively leverages this additional information within LLMs.

3 Proposed Model

3.1 Pre-trained Model

To implement LLM-based sentiment analysis, we adopt Llama 3.2 with 3 billion (3B) parameters and is instruction-tuned (Llama-3.2-3B-Instruct[1]). Llama family has been widely adopted in sentiment analysis (Mai et al., 2024; Roumeliotis et al., 2024; Gautam et al., 2025). In particular, the 3B architecture is a relatively lightweight, compared to mainstream 70–100B models, which is suitable for resource-constrained environments, such as on-device processing, with maintaining privacy.

For the immediate deployment, we do not apply fine-tuning. The hyper-parameters for the model are set in temperature=1.0. The model will stop inference after generating one token (equivalent to underbar in the prompt shown in Fig. 2).

3.2 Basic Prompt

To obtain sentiment values, we use text-completion function which the model completes the continuous texts of the given prompt.

As shown in (Zhang et al., 2024), first, one-shot model is significantly outperforms the zero-shot model. Additionally, few-shot prompt clearly outperforms the one-shot; however, In this study, using multiple examples introduces additional variables (e.g., the reference points of example reviews, their alignment patterns, example selection biases), making it difficult to interpret core findings regarding structured contextual information. Therefore, we use one-shot prompt (Fig. 2) to maintain experimental clarity.

```
### Instruction
You are a helpful assistant evaluating the review texts about
the restaurant. Please evaluate the review text and assign an
integer score ranging from 1 for the most negative comment
to 5 for the most positive comment. The output should be a
single integer from 1 to 5.

### Example
User review: {example_review}
Output: {example_label}

### Task
User review: {user_review}
Output: _
```

Figure 2: Basic Prompt

3.3 Displaying Supplementary Information

To display additional information, the method of presentation needs to be discussed. Learning struc-

[1]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

tural information is highly dependent on in-context learning, which means that locating the explanations before the tabular data improves the understanding of the structure (Sui et al., 2024). Therefore, we set up and compare two display methods, text and JSON format. We include the supplementary information, such as the average user rating and the average venue rating after the review section of the prompt.

Display Method (NL) First, we adopt text method, which displays the average evaluations with the explanations. With this method, although the prompt gets longer, any forms of information, including texts and numeric values, can be input into the model as long as the information can be explained in natural language. Thus, the necessary amount of computing resources becomes larger. Therefore, we set up the explanations and the input as Fig. 3.

When evaluating the review, consider both the textual sentiment and the supplementary information. Use user's average score which user has given in their past reviews to understand the user's typical rating behavior and restaurant's average score which the restaurant has received across all users to compare this restaurant's performance relative to others. Additionally, use restaurant's name, open hours (the total number of hours the restaurant is open in a week), and open days (the total number of days the restaurant is open in a week) to contextualize the review.

Example
User review: This restaurant was...
Supplementary Information: The average of this user's past ratings is 2.6. The average rating this restaurant has received is 3.0
Output: 5

Figure 3: NL Prompt for Supplementary Information

Display Method (JSON) Second, we adopt JSON method, which displays supplementary information with JSON format. This method can provide information in the structured form. One study (Sui et al., 2024) which investigates the impact of using table data, such as CSV, JSON, and HTML, on the understanding of the information indicates that LLMs can comprehend the contents of the table unless the data structure is not too complex and the ability is improved when explanations are added before the structured input. Therefore, we set up the explanations and the structured input as Fig. 4.

When evaluating the review, consider both the textual sentiment and the supplementary information in JSON format. Use 'user_average' (the average score this user has given in their past reviews) to understand the user's typical rating behavior and 'restaurant_average' (the average score this restaurant has received across all users) to compare this restaurant's performance relative to others. Additionally, use 'restaurant_name', 'open_hours' (the total number of hours the restaurant is open in a week), and 'open_days' (the total number of days the restaurant is open in a week) to contextualize the review.

Example
User review: This restaurant was...
Supplementary Information: {"user_average_stars": 2.6, "business_average_stars": 3.0}
Output: 5

Figure 4: JSON Prompt for Supplementary Information

3.4 Dataset

We adopt Yelp Open Dataset (Yelp, 2022) which contains the evaluations and reviews for a wide range of establishments. Compared with the popular benchmarks, such as IMDb movie reviews (Maas et al., 2011) and Amazon Reviews (Jianmo Ni, 2019), Yelp covers diverse user backgrounds and business conditions, resulting in higher heterogeneity across both users and businesses. This heterogeneity makes predictions more challenging since these latent differences are not fully captured within review texts alone.

For the comprehensive analysis, we set up the two different groups for the analysis: Restaurant and Nightlife, which are extracted using the category tags given for the establishments (Table 1). For each group, to prevent data leakage between train and test sets, we ensured that both user IDs and store IDs are mutually exclusive between the training and test sets of the entire Yelp dataset. Under this constraint, we use at most 500 unique user–business pairs for our evaluation set. Reviews are written in English.

For preprocessing the review texts, we at least remove the line break codes of the texts to maintain the format of the prompt. Table 2 shows the summary statistics of the dataset. A number of tokens is counted with tiktoken (OpenAI, 2023b) which is adopted in Llama 3. As shown in the statistics, some samples have significantly long texts.

Datasets	Selected	Excluded
Restaurant	Restaurant	Fast Food, Food Truck, Bar, Nightlife
Nightlife	Bar, Nightlife	Fast Food, Food Truck

Table 1: Selected and excluded category tags for each dataset. We avoid duplicate samples between datasets and select business with fixed addresses.

	Mean	Std	Min	Max
Restaurant				
Stars	3.724	1.515	1	5
Chars	431.726	368.464	42	2552
Tokens	98.744	85.544	9	606
Nightlife				
Stars	3.544	1.605	1	5
Chars	511.082	484.695	65	4998
Tokens	117.104	112.465	15	1118

Table 2: Summary statistics of each category in the dataset

4 Analyses and Results

4.1 Study 1: Impact of Reference Points and Display Methods

First, to address RQ1 (reference-point utilization) and RQ2 (prompt format), we implement sentiment analysis in restaurant and bar evaluations. We compare evaluation metrics across different models and assess the effectiveness of incorporating additional information in two categories. The supplementary information consists of following three elements. U: the user's average rating, indicating the mean rating of the past evaluations given by the user on Yelp, B: the business' average rating, indicating the mean rating which the restaurant has received from all users, and O: other contextual factors, indicating additional information both of textual and numerical attributes, such as the restaurant name, operating hours and the number of days the restaurant is open per week. Both U and B are expected to serve as reference points that affect user's prior expectations.

The LLM-based approach is evaluated with multiple variations, considering the type of supplementary information used and its machine-readability. Accordingly, the LLM-based models are categorized as follows: JSON-UBO / NL-UBO: Utilizing all supplementary information, presented in JSON format or natural language, respectively; JSON-UB / NL-UB: Incorporating only the average ratings; JSON-O / NL-O: Incorporating only contextual factors; and LLM (None): A baseline model without any supplementary information. We also establish four well-established baselines: BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al.). These pretrained models are fine-tuned for 5 epochs with additionally extracted 1000 training samples and the test performances are computed when the validation losses become the lowest. Since the proposed model employs 3B model which focused on lightweight and fast inferences, reference models are also base-sized (e.g., RoBERTa-Base-Uncased).

Sentiment analysis has not only the classification aspect but also the regression due to that the sentiment label is on the ordinal scale, which means that the magnitude of prediction error is important in addition to the concordance. Therefore, we adopt both Macro-F1 score and root mean square error (RMSE) for the evaluation metrics. For the baseline models we extracted an additional training set and fine-tuned each model.

Table 3 and 4 report the results for the Restaurant and Nightlife datasets, respectively. First, among the two datasets, JSON-UBO achieves the highest score in both datasets and improves significantly over LLM (None), which receives no supplementary information. Among the reference models, the strongest baseline differs by domain: LLM (None) ranks first in Restaurant, whereas RoBERTa-Base leads in Nightlife. In general, RoBERTa and DeBERTa outperform BERT, followed by DistilBERT.

Comparing the models within each display format, first, JSON prompts consistently improve performance as more information is added. These results indicate that, supplying a reference point in a machine-friendly format helps the model capture the complex relationships among factors, enabling effective inference. Increasing information from JSON-UB to JSON-UBO leads particularly

$n = 500$		UBO	UB	O
Macro-F1	LLM (JSON)	**0.612**†	**0.598**	**0.588**
	LLM (NL)	**0.593**†	**0.599**	0.524
	LLM (None)	0.587		
	DeBERTa	0.538		
	RoBERTa	0.533		
	BERT	0.474		
	DistilBERT	0.465		
RMSE	LLM (JSON)	**0.564**	**0.616**	**0.647**
	LLM (NL)	**0.620**	**0.624**	0.686
	LLM (None)	0.675		
	DeBERTa	0.703		
	RoBERTa	0.742		
	BERT	0.758		
	DistilBERT	0.804		

Table 3: Study 1 Results (Dataset 1: Restaurant). † indicates the statistically significant difference ($p < .05$) by two-sided McNemar test against LLM (None). Bold number indicates that the model surpasses all the reference models; shaded cells indicate the overall best value.

$n = 500$		UBO	UB	O
Macro-F1	LLM (JSON)	**0.635**†	0.622	0.592
	LLM (NL)	0.602	**0.628**	0.580
	LLM (None)	0.526		
	DeBERTa	0.523		
	RoBERTa	0.625		
	BERT	0.574		
	DistilBERT	0.481		
RMSE	LLM (JSON)	**0.597**	**0.613**	0.665
	LLM (NL)	0.672	**0.647**	0.666
	LLM (None)	0.709		
	DeBERTa	0.688		
	RoBERTa	0.657		
	BERT	0.668		
	DistilBERT	0.746		

Table 4: Study 1 Results (Dataset 2: Nightlife). † indicates the statistically significant difference ($p < .05$) by two-sided McNemar test against LLM (None). Bold number indicates that the model surpasses all the reference models; shaded cells indicate the overall best value.

to reduce RMSE in both datasets. As a result, prediction accuracy rises; for Restaurant, Macro-F1 rises by 4.3% (from 0.587 to 0.612) and RMSE reduces by 16.44% (from 0.675 to 0.564) relative to both LLM (None). For Nightlife, the improvements are even larger relative to LLM (None), +20.7% (from 0.526 to 0.635) and -15.8% (from 0.709 to 0.597), respectively, and still effective compared with RoBERTa (+1.6% / -9.1%). This result aligns with prospect theory and expectation-disconfirmation theory, providing the clear answer to RQ1. The user- and business-level average ratings act as reference points for the LLM and improve the prediction accuracy.

By contrast, the results of NL prompts do not follow this pattern; increasing information from NL-UB to NL-UBO does not contribute on the performance, and in Nightlife the accuracy even decreases below the best baseline. This suggests that, although LLMs can process natural language using the large context window, the models still struggle to capture their complex relationship particularly when large quantities of contextual factors are embedded as plain text.

These empirical results also provide the response to RQ2. Supplying additional information in a machine-readable structure allows the LLMs to effectively utilize those reference points and contextual factors for the prediction.

4.2 Study 2: Relationship with the Expectation

From the results of Study 1, a remaining concern is that these reference points may have worked as proxies for the labels. Therefore, to address RQ3 (proxy effect), Study 2 investigates whether model performance decreases as the review score diverges from these reference points according to the extent of the gaps. We use the JSON-UBO results for the examination.

Since prior expectations are formed based on the user's past behavior and the reputations, we treat such average score as indicators of prior expectations. We define expectation–evaluation gap for both user (U) and business (B) average from user i to store j as follows:

$$gap_{i,j}^{(U)} = rating_{i,j} - user\_average_i \quad (1)$$

$$gap_{i,j}^{(B)} = rating_{i,j} - business\_average_j \quad (2)$$

The data set is divided into five bins according to the extent of gaps, yielding groups that range

from "far below expectations" to "far above expectations." For each bin, we measure the Micro-F1 and RMSE. If the averages were merely proxies for the labels, performance would peak in the middle bin (where the gap is smallest) and decline sharply as the gap widens.

The results are shown in Table 5 (Restaurant) and Table 6 (Nightlife). The leftmost group includes cases where the actual rating falls below expectations, while the rightmost group consists of cases where the actual rating exceeds expectations.

	Expectation				
	below ←		met	→ beyond	
User Average					
$gap_{i,j}^{(U)}$	-1.788	-0.185	0.224	0.748	1.653
Micro-F1	**0.690**	0.636	0.667	**0.890**	**0.870**
RMSE	0.700	0.621	0.619	**0.374**	**0.436**
Business Average					
$gap_{i,j}^{(B)}$	-2.130	-0.640	0.415	0.945	1.565
Micro-F1	0.760	0.450	0.760	**0.860**	**0.910**
RMSE	0.624	0.819	0.490	**0.447**	**0.300**

Table 5: Study 2 Results (Dataset 1: Restaurant). Bold number indicates that the group surpasses the expectation-met group. Shaded cells indicate the overall best value.

First, in the Restaurant category, prediction performances increase in the upper two quantile groups for both the user's and store's average, compared to the middle group where the actual rating was close to the reference point. This suggests that the most accurate predictions were made when the actual experience exceeded prior expectations. In particular, when the experience was beyond the expectation, the performance improved by +25.1% for Micro-F1 and -39.6% for RMSE in user-average compared with the middle group while +13.2% for Micro-F1 and -8.8% for RMSE in business-average.

Next, in the Nightlife category, as in Study 1, merely meeting expectations did not consistently lead to better predictions. However, unlike the Restaurant category, prediction accuracy improved not only when experiences exceeded expectations, but also when they fell significantly short. Notably, the highest accuracy was observed in the group where the actual rating was far below the prior

	Expectation				
	below ←		met	→ beyond	
User Average					
$gap_{i,j}^{(U)}$	-2.083	-0.312	0.133	0.523	1.375
Micro-F1	**0.743**	**0.709**	0.700	**0.752**	**0.798**
RMSE	**0.507**	0.660	0.548	0.554	0.611
Business Average					
$gap_{i,j}^{(B)}$	-2.465	-0.995	0.350	0.975	1.520
Micro-F1	**0.830**	0.520	0.710	**0.850**	**0.800**
RMSE	**0.412**	0.911	0.566	**0.510**	**0.447**

Table 6: Study 2 Results (Dataset 2: Nightlife). Bold number indicates that the group surpasses the expectation-met group. Shaded cells indicate the overall best value.

expectation (the leftmost group).

Although there are substantial differences in business nature and customer behavior between restaurants and nightlife venues, at least, we do not confirm that the performance metrics increase in the group with the closest reference points to actual labels. In both cases, reference points do not simply function as proxies for the correct labels, but rather as relative evaluation values in inference, meaning that they function literally as reference points, which is a strong answer to RQ3.

4.3 Study 3: Error Analysis

To further understand how the model interacts with the reference points, we finally combine user and business average scores to create a 5×5 matrix where each cell represents the Micro-F1 score.

Restaurant	BA				
UA	1	2	3	4	5
1	-	1.000	0.812	0.750	-
2	1.000	0.833	0.771	0.643	0.000
3	1.000	0.643	0.688	0.716	0.000
4	-	0.250	0.679	0.816	0.750
5	-	1.000	0.909	1.000	1.000

Table 7: Error analysis by user average (UA) and business average (BA) for restaurant dataset

Table 7 (Restaurant) and 8 (Nightlife) show the results. First, in both categories, the model achieves highest performance (100% in most cases) when UA is 5. Second, two different reference points

Nightlife	BA				
UA	1	2	3	4	5
1	-	1.000	0.920	1.000	-
2	-	0.714	0.680	0.545	-
3	-	0.750	0.623	0.756	-
4	-	0.556	0.682	0.746	1.000
5	-	1.000	1.000	1.000	1.000

Table 8: Error analysis by user average (UA) and business average (BA) for nightlife dataset

show a clear interaction for the prediction. The accuracy tends to improve when two reference points align (along the diagonal), indicating that, when user's past evaluation is close to other consumers' average ratings, the actual rating becomes easier to predict. Notably, in some combinations, the accuracy results in 0%. This indicates cases where conflicting references make prediction challenging. However, as shown in Study 1, this accuracy even outperforms other models. Therefore, our approach can identify unreliable or difficult samples to predict based on reference point conflicts. These results clearly answer RQ4.

These interaction patterns enable practical deployment strategies. Companies can employ adaptive inference where samples with aligned reference points (UA≈BA) are processed on-device environment, while conflicting cases are routed to larger cloud-based models. Additionally, low-confidence predictions can be systematically collected as training data for fine-tuning domain-specific models, enabling continuous performance improvement.

5 Conclusion

5.1 Key Findings

In this study, we enhance LLM-based sentiment analysis by incorporating supplementary information as reference points and other contextual factors, based on prospect theory and EDT.

Study 1 compared two prompting strategies (NL and JSON) with multiple combinations of contextual information. The JSON-UBO model significantly outperformed both NL prompts and four strong baselines. Notably, while JSON prompts showed consistent gains with increasing information, NL prompts failed to leverage the same contexts despite the same context window.

Study 2 addressed the potential concern about reference points serving as label proxies. Accuracy improved more for reviews whose ratings deviated from the average than for those close to the average, indicating that the model was not simply copying the reference points for JSON-UBO model.

Study 3 revealed that the interactions of two different reference points affect the model performance. Accuracy improved when those ratings align while conflicting references indicate inherently challenging cases.

These findings comprehensively answer our research questions. **RQ1 (Effect of reference points):** We demonstrated that the proposed model with U/B/O information significantly improves performance with 4.3–20.7% gains in Macro-F1 and 15.8–16.4% reductions in RMSE over baselines without fine-tuning. **RQ2 (Effect of machine readability):** NL prompts fail to leverage complex information, highlighting the importance of prompt design even for models with large context windows. **RQ3 (Effect of proxy labels):** Follow-up analysis confirms that model performance improves when ratings deviate from expectations, indicating that reference points assist contextual inference rather than serving as mere label proxies. **RQ4 (Effect of reference interactions):** We revealed that aligned reference points improve prediction accuracy, while conflicting reference points indicate inherently challenging prediction cases.

5.2 Implications

This study has both academic and practical implications.

Academic. First, by incorporating the theoretical approach into the LLM-based sentiment analysis, model performance significantly improved, indicating that LLM's rich capability to handle complex context contributes to the predictions. By using JSON format, it is possible to input various information into LLMs, and combining more abundant information may further improve prediction accuracy. Second, even if the amount of information is same across the several prompts, the results of the inferences, including the prediction and performance, vary depending on the display methods, despite the large context window of modern LLMs. Second, simple scalar values, such as 1–5 star averages, can be used directly in the JSON prompt; no discretization or embedding tricks are required. These suggest that we can flexibly employ various factors, including textual and numeric information,

into sentiment analysis. Furthermore, although we employ sentiment analysis for the model verification, the proposed approach using JSON-based contextual information is transferable in wide domain of document classification task, including marketing analysis.

Practical. Since the proposed method only relies on prompt construction, companies can feed existing database contents to LLMs with JSON prompt and immediately construct their own extended models. As shown in the results, our approach achieves RMSE of 0.564 (restaurant) and 0.597 (nightlife), meaning the average prediction error is less than 1-star on a 5-point scale. This level of accuracy is sufficient for practical applications such as the simple recommendation systems, where distinguishing between adjacent rating categories (e.g., 4 vs 5 stars) is often less critical than identifying overall sentiment polarity. Achieving this performance with a 3B parameter scale without fine-tuning indicates that the company can immediately deploy the recommendation agent on edge application environments combined with the rich customer database. Furthermore, our error analysis revealed that samples with aligned reference points can be accurately predicted while cases with conflicting references could be routed to larger models or LLM-based ensemble strategy (Xing, 2025; Huang et al., 2024; Niimi, 2025). This enables the energy efficient processing where computational resources are allocated based on prediction difficulty.

6 Limitations

This study has several limitations. First, while our approach is grounded in prospect theory and EDT, we did not empirically test whether the psychological mechanisms underlying these theories actually explain the model's improved performance. Therefore, we need to further validate those relationships by the psychological experiments to support our findings.

Second, our experiments were conducted using only a single model architecture (Llama-3.2-3B-Instruct). The effectiveness of structured prompting may vary across different model families and scales, limiting the generalizability of our findings. Particularly, we cannot conclude whether the benefits of JSON formatting extend to larger models or different architectures.

Third, our evaluation is restricted to English reviews from two categories within the Yelp Open Dataset (Yelp, 2022). Additional verifications for other domains, languages, benchmarks are also required to support the effectiveness.

In addition, we did not compare our method with simple post-hoc calibration baselines (e.g., shifting predictions according to user or business averages). Such heuristics may adjust main effects but cannot capture interaction effects between users and venues, which our structured prompting approach is designed to address. Future research may provide a systematic comparison between these alternatives, incorporating effect-size measures for more robust evaluation.

Finally, while we argue that our approach is computationally efficient due to the absence of fine-tuning, we did not provide quantitative measurements of inference time or memory usage across different prompt formats for the actual inferences.

Acknowledgment

We are grateful to the two anonymous reviewers for their insightful comments. Their feedbacks have greatly improved our study.

Both the dataset and model were managed and used in an appropriate environments that comply with the terms of use. We do not collect additional information that could lead to the identification of individuals.

This study is supported by JSPS KAKENHI (Grant Number: 24K16472).

References

Barry J Babin, Yong-Ki Lee, Eun-Ju Kim, and Mitch Griffin. 2005. Modeling consumer satisfaction and word-of-mouth: restaurant patronage in korea. *Journal of Services Marketing*, 19(3):133–139.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Ruth N Bolton. 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, 17(1):45–65.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S Yu. 2025. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*.

Bruce Cooil, Timothy L Keiningham, Lerzan Aksoy, and Michael Hsu. 2007. A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of marketing*, 71(1):67–83.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, 118:272–299.

Andrew J Flanagin and Miriam J Metzger. 2013. Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, 29(4):1626–1634.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Himanshu Gautam, Abhishek Gaur, and Dharmendra Kumar Yadav. 2025. A survey on the impact of pre-trained language models in sentiment classification task. *International Journal of Data Science and Analytics*, pages 1–39.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Ting Liu, and Bing Qin. 2024. Ensemble learning for heterogeneous large language models with deep parallel collaboration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Julian McAuley Jianmo Ni, Jiacheng Li. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Timothy Lee Keiningham, Bruce Cooil, Edward C Malthouse, Bart Lariviere, Alexander Buoye, Lerzan Aksoy, and Arne De Keyser. 2015. Perceptions are relative: an examination of the relationship between relative satisfaction metrics and share of wallet. *Journal of Service Management*, 26(1):2–43.

Praveen K Kopalle and Donald R Lehmann. 2001. Strategic management of expectations: The role of disconfirmation sensitivity and perfectionism. *Journal of Marketing Research*, 38(3):386–394.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv preprint arXiv:1907.11692.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Zhelu Mai, Jinran Zhang, Zhuoer Xu, and Zhaomin Xiao. 2024. Financial sentiment analysis meets llama 3: A comprehensive analysis. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, MLMI '24, pages 171–175, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. ArXiv preprint arXiv:1301.3781.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Junichiro Niimi. 2025. A simple ensemble strategy for llm inference: Towards more stable text classification. In *Proceedings of the 30th International Conference on Natural Language & Information Systems (NLDB 2025)*, Lecture Notes in Computer Science. Springer.

Richard L Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, 17(4):460–469.

Richard L Oliver. 1999. Whence consumer loyalty? *Journal of marketing*, 63(4):33–44.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. *tiktoken: a fast BPE tokeniser for use with OpenAI's models*.

G Qin and Victor R Prybutok. 2008. Determinants of customer-perceived service quality in fast-food restaurants and their relationship to customer satisfaction and behavioral intentions. *Quality Management Journal*, 15(2):35–50.

Werner Reinartz, Manfred Krafft, and Wayne D Hoyer. 2004. The customer relationship management process: Its measurement and impact on performance. *Journal of marketing research*, 41(3):293–305.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Kisang Ryu, Heesup Han, and Tae-Hee Kim. 2008. The relationships among overall quick-casual restaurant image, perceived value, customer satisfaction, and behavioral intentions. *International journal of hospitality management*, 27(3):459–469.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Beverley A Sparks and Victoria Browning. 2011. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism management*, 32(6):1310–1323.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. ArXiv preprint arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. ArXiv preprint arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is chatgpt a good sentiment analyzer? In *First Conference on Language Modeling (COLM 2024)*.

Frank Xing. 2025. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Yelp. 2022. *Yelp Open Dataset, An all-purpose dataset for learning*. Yelp.

Ya You, Gautham G Vadakkepatt, and Amit M Joshi. 2015. A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, 79(2):19–39.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

RoSRL: Adaptive Rule-of-Sum Reinforcement Learning for Efficient and Reliable Summarization

Thu Phuong Tran Thi[1,2], Vinh Nguyen Van[2], Thai Nguyen Phuong[2], Quang Vu Ngoc[3], Khoa Nguyen Dang[4]

[1]Hanoi Metropolitan University, Hanoi, Vietnam

[2]VNU University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

[3]FPT IS Company Limited [4]University of Rochester

tttphuong2@daihocthudo.edu.vn vinhnv@vnu.edu.vn thainp@vnu.edu.vn

quang.vn@outlook.com knguy42@u.rochester.edu

Abstract

Hallucination remains a critical challenge for summarization, especially in domains such as law and large-scale news where factual accuracy is paramount. While reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) can reduce hallucination, they require costly preference annotations, limiting applicability in low-resource preference-label settings. Extractive methods inherently avoid unsupported content but often rely on static heuristics, constraining adaptability and coherence. We present RoSRL (Rule-of-Sum Reinforcement Learning), an efficient and reliable label-free extractive framework that combines interpretable, feature-based sentence scoring with reinforcement learning. RoSRL incorporates Lean-Proof Lite, a lightweight validator ensuring numerical and entity-level consistency, and adapts feature weights via multi-dimensional rewards—faithfulness, coverage, coherence, brevity, and section balance—optimized with Proximal Policy Optimization (PPO). Experiments on Vietnamese and English news and U.S. legal texts show consistent gains under ROUGE/BERTScore and other model-based summary evaluation methods, corroborated by human judgments. These results highlight RoSRL's efficiency, scalability, and reliability as a resource-conscious alternative to heuristic baselines and preference-optimized models, and underscore its potential as a foundation for future abstractive extensions.

1 Introduction

Automatic text summarization condenses large volumes of text into concise summaries that preserve essential information, enabling efficient information access in domains such as news, scientific publications, and legal documents. Existing approaches are typically categorized as extractive, which assemble summaries by selecting salient source sentences, and abstractive, which generate paraphrased content. Abstractive methods, powered by large neural language models, produce highly fluent summaries but are prone to factual errors. Extractive methods maintain stronger grounding in the source text, thus reducing the risk of hallucination, but often sacrifice coherence.

Hallucination remains a critical challenge for real-world deployment, especially in domains where factual reliability is paramount, such as law and high-volume news reporting. Neural abstractive summarizers frequently generate unsupported statements, making them unsuitable in contexts where factual errors carry legal or societal consequences (Maynez et al., 2020; Ji et al., 2023). Recent approaches such as RLHF (Stiennon et al., 2020; Ouyang et al., 2022) and direct preference optimization (DPO) (Rafailov et al., 2023) have shown promise in reducing hallucination, but both require large-scale annotated preference datasets, which are expensive to produce and scarce in low-resource languages like Vietnamese.

Extractive summarization provides a lightweight alternative by directly composing summaries from source sentences, inherently lowering the hallucination risk. However, sentence misplacement or incoherent selection can still distort meaning if taken out of context. Classic extractive methods such as Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), LexRank (Erkan and Radev, 2004), and TextRank (Mihalcea and Tarau, 2004) rely on manually tuned heuristics (e.g., sentence centrality, redundancy reduction) that limit adaptability to domain-specific needs for factual consistency and coherence. Neural extractive methods such as BERTSum (Liu and Lapata, 2019) leverage pretrained language models to improve performance but remain supervised and require labeled summaries, limiting scalability and cross-domain applicability. More recent reference-free approaches like OTExtSum (Tang et al., 2022) optimize semantic coverage but do not explicitly

ensure factual reliability, structural balance, or multilingual generalization (Min et al., 2025; Adams et al., 2023).

Reinforcement learning (RL) offers a promising pathway to overcome these limitations. Models such as the Reinforced Neural Extractive Summarizer (RNES) (Wu and Hu, 2018) and BanditSum (Dong et al., 2018) demonstrate that optimizing informativeness and coherence via bandit or policy-gradient methods can improve extractive summarization without gold-standard supervision. However, existing RL-based extractive methods do not explicitly address multi-dimensional quality objectives—such as factual consistency, coverage, and section balance—and show limited generalization to multilingual and low-resource contexts.

To address these gaps, we propose RoSRL: Adaptive Rule-of-Sum Reinforcement Learning for Efficient and Reliable Summarization, a label-free extractive framework that unifies interpretable feature-based sentence scoring with policy optimization. RoSRL is designed to be efficient in computation and reliable in factuality. It integrates two components: (*i*) Lean-Proof Lite, a lightweight factuality validator that enforces numerical, entity, and semantic consistency; and (*ii*) ADAPT_RULE, a PPO-based adaptive mechanism that dynamically refines feature weights under multi-dimensional rewards covering faithfulness, coverage, coherence, brevity, and section balance. By removing the need for labeled summaries and optimizing directly for complementary quality dimensions, RoSRL attains competitive accuracy with modest resources across languages and domains.

Our contributions are as follows:

- We introduce RoSRL, an adaptive rule-of-sum reinforcement learning framework for extractive summarization that combines interpretable feature scoring with dynamic weight refinement and Lean-Proof Lite for numerical and entity-level consistency.
- We design multi-dimensional, reference-free reward signals that drive PPO-based adaptation, enabling balanced optimization across factuality, coverage, coherence, brevity, and section representation without reliance on costly preference annotations.
- Through extensive experiments on Vietnamese news (VNExpress), CNN/DailyMail, and U.S. legal texts (BillSum), we show that the proposed reward design and adaptive weighting yield consistent improvements over strong heuristic and unsupervised baselines in automatic metrics, model-based summary evaluation, and human judgments.
- We demonstrate that RoSRL operates efficiently with modest GPU resources, underscoring its practicality for multilingual and low-resource-compute scenarios.

By bridging the gap between static-rule extractive methods and reinforcement learning–based adaptability, RoSRL provides a reliable, factuality-aware, and resource-conscious foundation for extractive summarization, and establishes a clear pathway toward future abstractive extensions.

2 Related Work

2.1 Heuristic and centrality-based extractive summarization

Early extractive summarization systems relied heavily on manually designed heuristics. MMR (Carbonell and Goldstein, 1998) balanced relevance and novelty using static weights, while graph-based algorithms such as LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) computed sentence salience via similarity graphs and damping factors. Submodular optimization (Lin and Bilmes, 2011) later formalized the trade-off between coverage and diversity, achieving strong results on DUC benchmarks. Extensions such as PacSum (Zheng and Lapata, 2019) incorporated section bias through position-aware edge weights, emphasizing leading sentences while preserving discourse centrality.

2.2 Learning-based methods: reinforcement learning, unsupervised, and optimal transport

The advent of deep learning enabled learning-based approaches to replace handcrafted features with representations learned from data. Reinforcement learning emerged as a label-free alternative: BanditSum (Dong et al., 2018) formulates summarization as a contextual bandit problem, directly optimizing ROUGE with rapid convergence; RNES (Wu and Hu, 2018) uses policy gradients to jointly optimize informativeness and coherence, yielding strong performance on CNN/DailyMail. Beyond RL, unsupervised neural methods leverage pretrained encoders. Xu et al. (Xu et al., 2020)

pre-trained a hierarchical transformer on unlabeled corpora and ranked sentences via self-attention, achieving competitive results among unsupervised systems, though primarily optimizing semantic coverage without explicit factuality checks. The reference-free OTExtSum (Tang et al., 2022) employs optimal transport to match semantic distributions between a document and its extract, achieving state-of-the-art results among unsupervised systems; however, it focuses on coverage and salience, lacking explicit numeric/entity faithfulness and sentence-level coherence, and has not been evaluated in multilingual contexts (Min et al., 2025; Adams et al., 2023).

2.3 Faithfulness and preference-optimization approaches

In abstractive summarization, reducing hallucination has been a major research focus. RLHF (Stiennon et al., 2020; Ouyang et al., 2022) and DPO (Rafailov et al., 2023) have achieved notable gains in English by aligning models with human preferences. However, these methods require large-scale preference-annotated datasets, which are costly and scarce in low-resource languages. In Vietnamese summarization, VSum-HB (Tran Thi et al., 2025) introduced a human feedback dataset of 5,000 samples from the Vietnews corpus to explore RLHF in a low-resource setting—marking an important first step toward preference-optimization for Vietnamese. Yet, the dependence on substantial annotated data and high computational cost motivates the search for lighter, label-free methods that maintain factual accuracy.

2.4 Vietnamese extractive summarization

For Vietnamese, Lam et al. (Lam et al., 2022) compared multiple extractive architectures (RNN, GRU-RNN, LSTM, BiLSTM, BERT) on a news dataset. BERT achieved the highest ROUGE-1 score (0.449) but the lowest ROUGE-2 score (0.186), suggesting a bias toward sentences with overlapping keywords and weaker performance on longer n-gram coherence. The method is fully supervised, relying on sentence-level labels, which may entail limited generalization to different domains or low-resource contexts. To date, no Vietnamese extractive summarization framework has combined reinforcement learning with unsupervised feature-based scoring to jointly ensure coverage, robustness, and factual consistency.

2.5 Diffusion-based and structure-aware extractive models

Diffusion-based approaches such as DiffuSum (Zhang et al., 2023) generate summary sentence embeddings via diffusion processes, then select sentences by alignment, achieving state-of-the-art ROUGE-2 scores and strong cross-domain generalization. TermDiffuSum (Dong et al., 2025) incorporates legal-term-aware diffusion scheduling for legal text summarization, improving relevance in the legal domain. Structure-aware approaches such as SumHiS (Pavel et al., 2024) exploit latent document clustering to guide sentence selection, yielding substantial ROUGE-2 gains. However, the two-stage architecture involving sentence ranking and hidden-structure discovery introduces additional computational overhead, and the method does not incorporate explicit mechanisms for factuality or coherence verification.

RoSRL is conceptually closest to heuristic systems such as MMR, LexRank, and PacSum, as it starts from feature-weighted, sentence-scoring baselines. It extends these by incorporating factuality-oriented features—including NLI-based inference, SBERT-based alignment, numeric/date indicators, and section-aware priors—and by introducing ADAPT_RULE, a PPO-based adaptive mechanism that dynamically adjusts weights under multi-dimensional, reference-free rewards. In contrast to supervised extractive models or resource-intensive diffusion-based approaches, RoSRL maintains computational efficiency on modest GPU resources while explicitly optimizing for faithfulness, coverage, coherence, brevity, and structural balance in multilingual and low-resource settings. This positions RoSRL as a bridge between interpretable heuristic methods and adaptive reinforcement learning frameworks.

3 Proposed Method

3.1 Method Overview

RoSRL is a label-free extractive summarization framework designed to combine the interpretability of feature-based scoring with the adaptability of reinforcement learning. The pipeline (Figure 1) begins with document segmentation and feature extraction, producing a representation for each candidate sentence. These features are scored using a linear combination of interpretable metrics, forming both a rule-based baseline summary and a policy-generated summary. Rewards are computed for

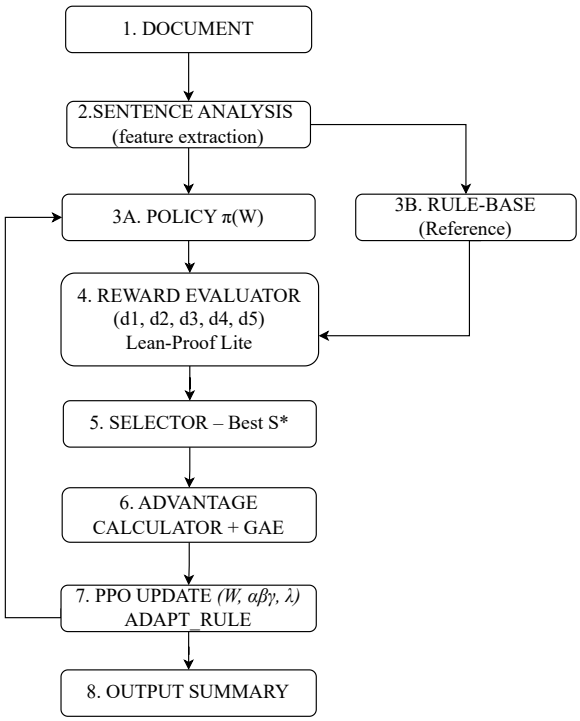


Figure 1: **RoSRL pipeline.** Reward dimensions: d_1 = faithfulness, d_2 = coverage, d_3 = coherence, d_4 = brevity, d_5 = section balance. The overall reward combines these dimensions (see Eq. 3).

candidates that pass a lightweight factuality validator, Lean-Proof Lite; a selector then identifies the best candidate summary and its return/advantage is estimated using generalized advantage estimation (GAE) before updating the policy parameters with PPO (Schulman et al., 2017). In addition, the framework includes an ADAPT_RULE mechanism that schedules updates to feature weights and combination coefficients in two phases to enhance training stability. This combination ensures RoSRL retains the factual reliability inherent in extractive methods while adaptively improving faithfulness, coverage, coherence, brevity, and section balance.

3.2 Feature-based Extractive Policy

Given a document D_i, we segment it into sentences $\{s_1, \ldots, s_m\}$, each represented by a 7-dimensional interpretable feature vector: TF–IDF centrality, sentence position, section bias, length normalization, novelty, named-entity density, and keyword overlap. We adopt a minimal, interpretable, domain-agnostic basis that spans four complementary priors: (i) topical centrality (TF–IDF); (ii) document structure (position, section bias); (iii) redundancy/length control (novelty, length normalization); and (iv) salience cues (named-entity density, keyword overlap). This 7-D design is computationally light and directly controllable via W and the mixing weights $\{\alpha, \beta, \gamma\}$, which is crucial for ADAPT_RULE and PPO stability. *Feature–reward alignment:* features $\{1, 6, 7\}$ primarily support coverage/faithfulness; feature 5 improves coherence by reducing redundancy; feature 4 enforces brevity; and features $\{2, 3\}$ promote section balance.

A rule-based score is computed as

$$\text{score}_{\text{rule}}(s_j) = \sum_{d=1}^{7} w_d \, f_{j,d}, \tag{1}$$

with human-initialized weights w_d (Appendix A) ensuring transparency.

The policy π_W models a Bernoulli select/skip decision for each sentence, subject to budget and novelty constraints. To enable domain adaptation, a mixing stage refines the base score:

$$\begin{aligned}\text{score}_{\text{mix}}(s_j) = \alpha\, \text{score}_{\text{rule}} + \beta\, \text{centrality} \\ + \gamma\, \text{section\_bias}.\end{aligned} \tag{2}$$

Here, (α, β, γ) are trainable re-weighting parameters, initialized neutrally and optimized via PPO together with reward weights $\boldsymbol{\lambda}$, enabling adaptive re-balancing of heuristic features while retaining interpretability. $\text{score}_{\text{mix}}$ degenerates to the rule baseline when α=1 and β=γ=0, i.e., sentences are selected purely by $\text{score}_{\text{rule}}$ in Eq. (1). Under ADAPT_RULE, we only summarize the schedule here and defer details to Section 3.5.

The rule baseline S_{rule} is produced *greedily* from Eq. (1) using the default weights (Default_W, Table 6) with mixing disabled (α=1, β=γ=0), under the same budget and novelty constraints as the policy. For each document we form two candidates, S_{rule} and S_{pol} from π_W; both pass the Lean-Proof Lite gate and are evaluated along five reward dimensions (Eq. 3). The better candidate $S^\star$ supplies the rollout return and the advantage (GAE) used in Algorithm 1.

3.3 Multi-dimensional Reward and Lean-Proof Lite

To avoid reliance on costly preference labels, RoSRL employs multi-dimensional, reference-free reward signals. For a generated summary S, we compute normalized scores along five dimensions: faithfulness (entity and number consistency with the source), coverage (content overlap with salient

source segments), coherence (logical and structural flow), brevity (conciseness relative to budget), and section balance (distribution of coverage across document sections). Each score $r_k \in [0, 1]$ reflects the degree to which S satisfies dimension k.

Before aggregation, candidate summaries are filtered through the Lean-Proof Lite validator, which acts as a lightweight factuality gate with three checks: (i) numeric consistency—every numerical mention in the summary must appear in the source; (ii) entity consistency—named entities in the summary must be supported by the source (string-aligned NER); and (iii) lexical overlap—the Levenshtein similarity (normalized edit distance) between the summary and the source must be at least $\delta = 0.35$. Only summaries satisfying all checks are passed forward for reward computation. The threshold δ was selected via development experiments; prior work emphasizes that similarity thresholds should be tuned to the dataset and the chosen metric, and practical duplicate-detection systems routinely employ fixed thresholds tuned to their feature design and similarity functions (van Bezu et al., 2015). This lightweight validation is especially important in legal and news domains, where hallucinated numbers or entities can have severe consequences.

Validated dimension scores $\{r_k\}_{k=1}^{5}$ are then combined via a learned convex weighting:

$$R(S) = \sum_{k=1}^{5} \lambda_k r_k, \quad \lambda_k \geq 0, \quad \sum_{k=1}^{5} \lambda_k = 1 \quad (3)$$

where λ_k are learned parameters updated during training to adaptively prioritize dimensions that most improve downstream summarization performance. This design ensures both interpretability and adaptive reward shaping for PPO-based policy optimization.

3.4 PPO-based Adaptive Optimization

RoSRL employs PPO (Schulman et al., 2017) for stable policy optimization, leveraging its clipped surrogate objective to prevent destructive updates. Given old and new policy probabilities

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad (4)$$

Here, s_t denotes the sentence-level state and a_t the binary select/skip action.

$$L^{\text{clip}}(\theta) = \hat{\mathbb{E}}_t\big[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\big] \quad (5)$$

where ϵ follows the commonly used clipping range in PPO (Schulman et al., 2017). Advantages $\hat{A}_t$ are estimated using Generalized Advantage Estimation (GAE) (Schulman et al., 2016), following the setup in (Stiennon et al., 2020). In our sentence-level setting, rewards are computed at the document level and propagated back to individual actions, preserving cross-sentence dependencies.

Integrated Training Loop. The overall training procedure for each pass p over the dataset follows the steps in Algorithm 1. For each document D_i, we extract features, generate a rule-based summary S^{rule} and a policy summary S^{pol}, and compute their rewards using the multi-dimensional function gated by Lean-Proof Lite. The better-performing summary is selected as S^*, with its log-probabilities and GAE-based advantages $\hat{A}$ stored in a buffer B. After processing all documents, we run K PPO epochs over minibatches from B, computing L^{clip} and auxiliary losses. Parameter updates, including those under the ADAPT_RULE scheme, are described in Section 3.5. Early stopping is applied if validation rewards stagnate.

Training logs show that the share of policy-selected summaries increases consistently across rollout passes on all datasets, with the largest gains by the third pass (see Fig. 2 in the appendix A).

3.5 ADAPT_RULE Scheduling

Following the main PPO update loop in Algorithm 1, RoSRL applies a two-phase parameter update scheme to improve training stability. In Phase 1, only the λ weights for reward dimension combination are updated, allowing the reward signal to stabilize without perturbing the sentence scoring policy. In Phase 2, ADAPT_RULE is activated, enabling joint updates to W, α, β, γ, and λ. This staged approach mitigates reward instability and expedites convergence, consistent with curriculum-style optimization strategies in deep and multi-objective reinforcement learning (Portelas et al., 2020; Kang et al., 2023). Empirically, the ADAPT_RULE schedule yields a steady reward trajectory and a monotonic rise in policy-selected rollouts across passes (see Fig. 3 and Fig. 2).

Algorithm 1: RoSRL training with Lean-Proof Lite gate and ADAPT_RULE scheduling.

Input: $\{D_i\}$; init W, (α, β, γ), λ;
Output: W^*, $(\alpha, \beta, \gamma)^*$, λ^*, policy π^*
for $p \leftarrow 1$ **to** P **do**
  $B \leftarrow \emptyset$
  foreach D_i **do**
    Segment & extract features
    S^{rule} (rule), $S^{pol} \leftarrow \pi_W(D_i)$
    Rewards R^{rule}, R^{pol} **with Lean-Proof Lite gate**
    Select S^*, attach $\hat{A}$ via GAE, store $(\log \pi, R, \hat{A})$ in B
  end
  for *epoch* $\leftarrow 1$ **to** *PPO_EPOCH* **do**
    Sample minibatch; compute r_t & clipped L^{PPO} + aux losses
    Update λ **always;** $W, (\alpha, \beta, \gamma)$ **only if ADAPT_RULE**
  end
  Eval valid reward, update best, early stop if no improve
end

We interpret ADAPT_RULE as an implicit ablation: Phase 1 ($\boldsymbol{\lambda}$-only) disables the mixing pathway by freezing W and $\{\alpha, \beta, \gamma\}$—equivalently α=1, β=γ=0—thus isolating the effect of $\boldsymbol{\lambda}$ in Eq. (3). Phase 2 unfreezes W and $\{\alpha, \beta, \gamma\}$ and jointly updates them with $\boldsymbol{\lambda}$, revealing the incremental contribution of `score_mix` and the feature weights. Empirically, the moving-average reward increases smoothly from ≈ 1.0 to ≈ 1.8 without early oscillations (Fig. 3), while policy-win counts grow from R1 to R3 (Fig. 2).

Learned parameters at the best validation checkpoint. At the checkpoint with the highest validation reward (see Table 7), the learned mixing coefficients in Eq. (2) highlight the contribution of the mixing pathway, once Phase 2 is enabled, *section_bias* contributes more prominently than *centrality*. The learned reward weights in Eq. (3) suggest a clear prioritization of factuality and coverage (including section balance), while the penalty on length is kept moderate to avoid over-compression. The learned feature weights W allocate more mass to salience cues (named-entity density, keyword overlap), followed by topical centrality and structural signals (TF–IDF, position), with length normalization playing a comparatively smaller role. These patterns are consistent with the two-phase ADAPT_RULE schedule: Phase 1 stabilizes λ, Phase 2 jointly refines W and (α, β, γ), yielding smoother reward trajectories and higher policy-win rates during training.

4 Experiment setup

4.1 Datasets

We evaluate RoSRL on three datasets spanning two languages (Vietnamese, English) and two domains (news, legal), enabling cross-lingual and cross-domain testing.The Vietnamese set-VnExpress [1] is a curated VnExpress news corpus (2022–2024) comprising 13,468 samples, covering 15 categories and 80+ subtopics. The English datasets are CNN/DailyMail[2] for news and BillSum[3] for legislative texts. These datasets differ in language, style, and complexity, providing a robust evaluation setting.

4.2 Baselines and Initialization

RoSRL is initialized with interpretable feature weights W, (α, β, γ), and λ, which control sentence scoring and reward aggregation. These values, provided in Table 6, are chosen based on heuristic principles from prior extractive summarization studies—such as emphasizing sentence lead position, penalizing redundancy, and balancing coverage with brevity. We experiment with two encoder backbones for sentence representation: (i) the multilingual `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019) for cross-lingual applicability, and (ii) the Vietnamese domain-adapted `VoVanPhuc/sup-SimCSE-VietNamese-phobert-base` (Gao et al., 2021) for stronger performance in Vietnamese news.

4.3 Hyperparameters for RL

All experiments run on a single NVIDIA T4 GPU (15 GB). We use the full VnExpress training set, and randomly sample 10,400 documents from

[1] https://huggingface.co/datasets/quancute/VnExpress_news_summarization_sft_dataset
[2] https://huggingface.co/datasets/abisee/cnn_dailymail
[3] https://huggingface.co/datasets/FiscalNote/billsum

CNN/DailyMail and 14,000 from BillSum for training, ensuring diverse yet computationally manageable rollouts. A training pass consists of a complete sweep of rollouts over the sampled training set, followed by PPO updates and validation reward checks.

4.4 PPO settings

Adam optimizer (learning rate 3×10^{-4}), clipping parameter $\epsilon = 0.2$ (Schulman et al., 2017), GAE parameter is 0.95 (Schulman et al., 2016), batch size 4, and 4 PPO epochs per rollout batch. Rollouts are batched at 64 documents for stability.

4.5 Evaluation Metrics

We employ a combination of automatic, LLM-based, and human evaluation metrics to comprehensively assess summary quality. ROUGE-1/2/L (Lin, 2004) measures lexical overlap between generated and reference summaries, with higher scores indicating greater n-gram matching, while BERTScore (Zhang et al., 2020) evaluates semantic similarity via token-level cosine similarity in contextual embedding space. Additionally, UniEval (Zhong et al., 2022) evaluates generated summaries along four dimensions—coherence (quality of logical flow), consistency (alignment with the source content), fluency (linguistic quality), and relevance (coverage of essential information). For LLM-based evaluation, we adopt GPT-4.0 due to its strong alignment with human judgments and demonstrated state-of-the-art performance in text evaluation tasks (Liu et al., 2023), using tailored prompts to assess contextual consistency, relevance, and coherence on a 5-point Likert scale (Likert, 1932). For human evaluation, three research volunteers independently assessed 200 randomly selected samples from the BillSum test set using prompts created with GPT-4.0, carefully reviewing each sentence before assigning scores from 1 to 5 for each criterion, with the final score for each summary calculated as the arithmetic mean of the three annotators' scores. Detailed definitions and guidelines for these evaluation criteria are provided in Appendix B.

5 Evaluation

5.1 ROUGE Results

Table 1 compares RoSRL with strong extractive baselines. On CNN/DailyMail, SumHiS achieves the highest ROUGE score, while RoSRL attains competitive ROUGE-1 (0.3133) despite using no gold summaries and focusing on multi-dimensional quality rather than lexical overlap alone. On BillSum, RoSRL outperforms all methods in ROUGE-1 (0.4119) with comparable ROUGE-2 and ROUGE-L. For the Vietnamese news domain, RoSRL surpasses the supervised baseline reported by Lam et al. (2022) across all ROUGE variants—even though their experiment uses the CTUNLPSum corpus (95,579 articles) and ours uses a smaller, curated VnExpress dataset (13,468 articles, 15 categories, 80+ subtopics)—demonstrating RoSRL's strong adaptability to low-resource Vietnamese settings.

5.2 BERTScore Results

Table 2 compares semantic similarity scores (F1) between the baseline (fixed initialization weights) and the optimized (PPO-adapted) RoSRL.

Dataset	Baseline	Optimized
VnExpress	0.892	0.923
CNN/DailyMail	0.844	0.855
BillSum	0.783	0.784

Table 2: BERTScore (F1) comparison for RoSRL baseline vs. optimized.

BERTScore results show consistent improvements after PPO optimization, with the largest gain (+0.031) on VnExpress, reflecting RoSRL's ability to enhance semantic similarity without supervised fine-tuning. Gains on CNN/DailyMail are smaller but consistent, while BillSum sees marginal change due to its inherently formulaic legislative style.

5.3 UniEval Results

Table 3 reports RoSRL's UniEval quality scores. On CNN/DailyMail, coherence improves from 0.629 to 0.722, and relevance from 0.514 to 0.618, while consistency and fluency remain high (>0.81). On BillSum, relevance gains +0.095, with coherence improving from 0.689 to 0.728.

5.4 LLM and Human Evaluation Results

Table 4 reports the average scores (1–5) for contextual consistency, relevance, and coherence on CNNDailyMail, VnExpress, and BillSum. For BillSum, LLM-based scoring was complemented by human verification on 200 random test samples to ensure reliability in the legal domain. Overall, optimized RoSRL achieves consistently higher

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L
CNN/DailyMail	TextRank (2004)	0.3126	0.1118	0.1940
	LexRank (2004)	0.3245	0.1146	0.2013
	LSA (2001)	0.2933	0.0909	0.1816
	OT_ExtSum-BS (BERT)	*0.3450*	*0.1280*	*0.2780*
	SumHiS (w/ filtering)	**0.4348**	**0.3252**	**0.4244**
	RoSRL (Baseline)	0.3088	0.1157	0.1919
	RoSRL (Optimized)	0.3133	0.1195	0.1951
BillSum	LexRank (2004)	0.3845	0.1888	0.2460
	TextRank (2004)	0.3638	0.1735	0.2168
	LSA (2001)	0.3480	0.1406	0.2115
	OT_ExtSum-BS (Word2Vec)	*0.4010*	*0.1940*	**0.3430**
	OT_ExtSum-BS (BERT)	0.3750	**0.1970**	*0.3260*
	RoSRL (Baseline)	0.4005	0.1773	0.2393
	RoSRL (Optimized)	**0.4119**	0.1874	0.2563
CTUNLPSum	BERT(Lam et al. (2022))	0.4490	0.1860	0.0325
VnExpress	RoSRL (Baseline)	*0.6581*	*0.3822*	*0.4075*
	RoSRL (Optimized)	**0.6646**	**0.3902**	**0.4125**

Table 1: ROUGE scores across CNN/DailyMail, BillSum, and Vietnamese news. Best results are in **bold**, second-best are *italicized*.

Dataset / Setting	Coherence	Consistency	Fluency	Relevance
CNN/DailyMail (Baseline)	0.629	0.902	0.814	0.514
CNN/DailyMail (Optimized)	0.722	0.909	0.816	0.618
BillSum (Baseline)	0.689	0.857	0.649	0.546
BillSum (Optimized)	0.728	0.899	0.857	0.641

Table 3: UniEval scores for RoSRL before and after PPO optimization.

scores than the baseline across all datasets, with notable gains in relevance and coherence, while maintaining high contextual consistency. These results indicate that the optimized system produces more informative and better-structured summaries without sacrificing faithfulness to the source. Some examples of a generated summary in Appendix C.

6 Discussion

6.1 Adaptive: Cross-Domain Robustness via Multi-Dimensional Rewards

RoSRL is designed as a label-free extractive framework that adapts to diverse domains without requiring gold summaries or costly preference annotations. Its reinforcement learning component dynamically adjusts interpretable feature weights—including faithfulness, coverage, coherence, brevity, and section balance—using multi-dimensional rewards optimized with PPO.

Dataset / Method	Cons.	Rel.	Coh.
CNN/DailyMail Baseline	4.5	3.6	3.5
CNN/DailyMail Optimized	4.7	4.0	3.8
VnExpress Baseline	4.4	3.6	4.2
VnExpress Optimized	4.6	3.9	4.4
BillSum Baseline	4.2	3.2	3.3
BillSum Optimized	4.4	3.5	3.5

Table 4: LLM-based scores (1–5) for contextual consistency (Cons.), relevance (Rel.), and coherence (Coh.). BillSum results verified by human annotators on a 200-sample subset.

This adaptability is evident in its competitive ROUGE, BERTScore and UniEval gains across CNN/DailyMail, BillSum, and Vietnamese news. Such results demonstrate that RoSRL effectively generalizes across both high-resource and low-

Dataset	Avg. src length	Avg. RoSRL
VnExpress	460	138
CNN/DailyMail	696	124
BillSum	1361	237

Table 5: Average word length of source documents and RoSRL summaries on the test sets.

resource settings, surpassing static-heuristic extractive baselines while maintaining domain sensitivity.

6.2 Efficient: Resource-Conscious Reinforcement Learning

RoSRL's design emphasizes computational efficiency, enabling scalable training and deployment in resource-constrained environments. Leveraging mixed-precision training and lightweight PPO updates, RoSRL completes full training in under 7 GPU-hours per dataset on a single NVIDIA T4 (15GB), with peak memory usage below 6GB. Length statistics in Table 5 show that RoSRL produces summaries between 124 and 237 words, achieving aggressive compression for short-form news and moderate compression for lengthy legislative texts. This ability to tailor summary length to document complexity ensures optimal trade-offs between brevity and information preservation, further supporting cross-domain usability.

6.3 Reliable: Factual Consistency and Human-Aligned Quality

Reliability is reinforced through Lean-Proof Lite, a lightweight validation module that ensures numerical and entity-level consistency during extraction. Improvements in BERTScore, UniEval and LLM-based evaluations confirm that RoSRL delivers summaries with stronger semantic fidelity and structural coherence. On BillSum, human verification of 200 random test samples corroborates GPT-4.0 evaluations, indicating that the observed automatic metric gains translate into higher human-perceived quality. This alignment between model-based and human assessments underscores RoSRL's robustness as a dependable summarization framework.

7 Conclusion

This study introduced RoSRL, a reinforcement learning framework for extractive summarization that combines interpretable feature-based scoring with the Lean-Proof Lite factual consistency checker and the ADAPT_RULE mechanism for dynamic weight adaptation using PPO. This approach ensures factual reliability and enables adaptation to multiple domains without supervised fine-tuning.

Experiments on multiple news and legal datasets show that RoSRL generates concise summaries that preserve the core content and achieve reasonably high quality in automatic evaluation, LLM-based assessment, and human evaluation. UniEval results indicate consistent improvements across all criteria, particularly in consistency, coherence, and relevance, while maintaining high fluency.

However, as an extractive summarizer, RoSRL is limited in its ability to naturally rephrase, restructure sentences, and integrate information—capabilities often seen in abstractive methods. The system also relies on predefined feature functions and manually designed reward components. Future work will focus on integrating a lightweight abstractive summarization module, developing domain-adaptive feature learning to reduce manual tuning, and testing improvements on low-resource and rare-language domains, with the ultimate goal of creating a summarization system that produces fluent, domain-adaptive outputs and consistently excels across factuality, coherence, and relevance.

Acknowledgments

We extend our sincere gratitude to the open-source communities whose datasets and tools have provided essential resources for this research. We are deeply thankful to the volunteer evaluators for their valuable contributions in assessing the quality of the generated summaries. This research has been done under the research project QG.23.73 of Vietnam National University, Hanoi.

References

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. *arXiv preprint arXiv:2303.03948*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. Association for Computing Machinery.

Xiangyun Dong, Wei Li, Yuquan Le, Zhangyue Jiang, Junxi Zhong, and Zhong Wang. 2025. TermDif-

fuSum: A term-guided diffusion model for extractive summarization of legal documents. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3222–3235. Association for Computational Linguistics.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yanfei Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiachen Kang and 1 others. 2023. Learning multi-objective curricula for robotic policy learning. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *PMLR*.

Khang Nhut Lam, Tuong Thanh Do, Nguyet-Hue Thi Pham, and Jugal Kalita. 2022. Vietnamese text summarization based on neural network models. In *Artificial Intelligence in Data and Big Data Processing*, pages 85–96, Cham. Springer International Publishing.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 404–411. Association for Computational Linguistics.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of llm summarization across domains and languages. *arXiv preprint arXiv:2506.00549*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2022)*, pages 27730–27744. Curran Associates, Inc.

Tikhonov Pavel, Anastasiya Ianina, and Valentin Malykh. 2024. Sumhis: Extractive summarization exploiting hidden structure. *Preprint*, arXiv:2406.08215.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. 2020. Automatic curriculum learning for deep rl: A short survey. In *Proceedings of IJCAI*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2023)*, pages 53728–53741. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. 2016. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2020)*, pages 3008–3021. Curran Associates, Inc.

Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. 2022. OTExtSum: Extractive Text Summarisation with Optimal Transport. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1128–1141, Seattle, United States. Association for Computational Linguistics.

Thu Phuong Tran Thi, Vinh Nguyen Van, Thai Nguyen Phuong, Quy Nguyen Minh, and Anh Quan Nguyen Duc. 2025. Vsum-hb: A vietnamese text summarization dataset for reinforcement learning from human feedback. In *Information and Communication Technology*, pages 148–161, Singapore. Springer Nature Singapore.

Ronald van Bezu, Sjoerd Borst, Rick Rijkse, Jim Verhagen, Damir Vandic, and Flavius Frasincar. 2015. Multi-component similarity method for web product duplicate detection. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, page 761–768, New York, NY, USA. Association for Computing Machinery.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5602–5609. AAAI Press.

Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. 2020. Unsupervised extractive summarization by pre-training hierarchical transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1784–1795, Online. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. DiffuSum: Generation enhanced extractive summarization with diffusion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13089–13100. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Parameter	Value
DEFAULT_W	[0.4, 0.4, 0.2, 0.6, 0.3, 0.7, 0.5]
DEFAULT_MIX	$\alpha = 0.5, \beta = 0.3, \gamma = 0.2$
DEFAULT_LAMBDA	$\lambda_1 = 0.45$ (faithfulness), $\lambda_2 = 0.20$ (coherence), $\lambda_3 = 0.10$ (length-control), $\lambda_4 = 0.25$ (coverage/relevance), $\lambda_5 = 0.10$ (section-coverage)

Table 6: Default initialization parameters used in RoSRL.

Parameter	Value
LEARNED_MIX	$\alpha = 0.558$, $\beta = 0.169$, $\gamma = 0.272$
LEARNED_LAMBDA	$\lambda_1 = 0.269$ (faithfulness), $\lambda_2 = 0.192$ (coherence), $\lambda_3 = 0.102$ (length control), $\lambda_4 = 0.226$ (coverage/relevance), $\lambda_5 = 0.212$ (section balance)
LEARNED_W	[0.832, 0.818, 0.603, 0.546, 0.734, 1.121, 0.979]

Table 7: **Learned parameters at the best validation step.** Values correspond to the checkpoint with the highest validation reward used for reporting. The order of W matches the 7 features in §3.2

A Initialization Settings

Table 6 lists the default initialization parameters for RoSRL. These values were manually tuned, drawing on principles from prior empirical work in extractive summarization (Carbonell and Goldstein, 1998; Erkan and Radev, 2004; Tang et al., 2022) regarding sentence centrality, redundancy reduction, and structural balance. The weighting scheme prioritizes faithfulness-related metrics, aligned with recommendations from prior studies on minimizing factual errors in high-stakes domains (Maynez et al., 2020; Ji et al., 2023).

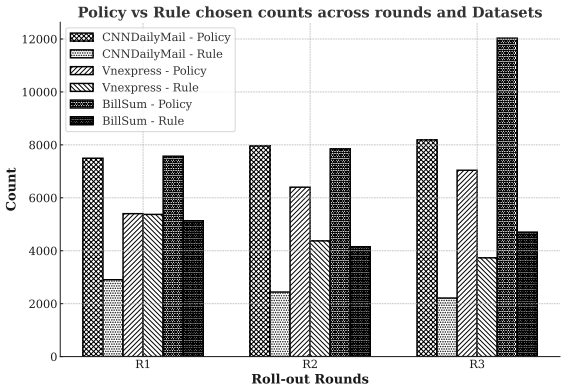


Figure 2: **Policy vs. Rule selection across rollout passes.** Each bar shows the number of documents for which the policy-generated summary was chosen over the rule-based reference (and vice versa) in rounds R1–R3, for VnExpress, CNN/DailyMail, and BillSum. Training logs indicate a consistent shift from rule dominance to policy dominance as PPO adapts the weights (cf. Algorithm 1).

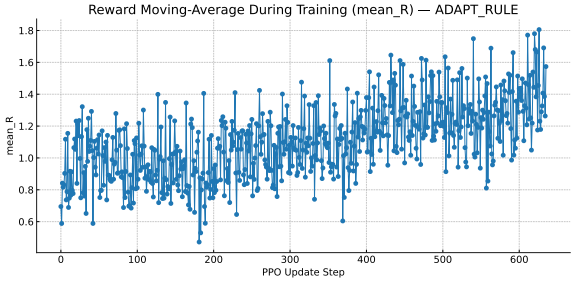


Figure 3: **Reward moving average across PPO updates (ADAPT_RULE).** The document-level reward steadily increases (about 1.0 to 1.8) without early-stage oscillations, indicating that stabilizing λ before updating $\{W, \alpha, \beta, \gamma\}$ prevents reward drift. Together with Fig. 2, this supports the stability–then–adapt dynamics of ADAPT_RULE.

Description of the ChatGPT evaluation
You are a summary quality evaluation expert. Based on the original source text, evaluate the extractive summary according to the following three criteria. Each criterion is scored on a scale from 1 to 5 (1 = very poor, 5 = excellent). **1. Contextual Consistency** — The degree to which the summary preserves the meaning and context of the source text without distortion or factual errors. 1: Many major factual/context errors 2: Some important details misrepresented or out of context 3: Mostly correct, with minor misleading points 4: Almost entirely accurate, negligible issues 5: Fully accurate and preserves context **2. Relevance** — The extent to which the summary covers the key content points. Step 1: Identify and assign weights to each key content point (total = 100%). Step 2: Mark "Yes"/"No" for each point depending on its presence in the summary. Step 3: Score based on the total weight covered. **3. Coherence** — The degree of logical connection between sentences, clarity, and readability of the summary. 1: Disjointed, hard to read 2: Weak connections, frequent abrupt transitions 3: Moderate coherence, some breaks in flow 4: Good flow, clear and easy to follow 5: Excellent flow, natural and fully coherent narrative

Table 8: Description of the ChatGPT evaluation prompt.

B Prompt Evaluation

C Generation Samples

Source document:
"SECTION 1. SHORT TITLE. This Act may be cited as the ``Special Agent Scott K. Carey Public
Safety Officer Benefits Enhancement Act". TITLE I--EDUCATIONAL ASSISTANCE TO OFFICERS DISABLED IN THE LINE OF DUTY. SEC. 101. BASIC ELIGIBILITY. Section 1212(a)(1) of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796d-1(a)(1)) is amended-- (1) by striking ``a dependent" and inserting ``an eligible dependent"; and (2) by striking ``education" and all that follows through the period at the end and inserting ``education.".
SEC. 102. APPLICATIONS; APPROVAL. Section 1213 of the Omnibus Crime Control and Safe Streets Act of
1968 (42 U.S.C. 3796d-2) is amended-- (1) in subsection (b)-- (A) by striking ``the dependent" each place it apears and inserting ``the applicant"; and (B) by striking ``the dependent's" each place it appears and inserting ``the applicant's"; and (2) in subsection (c), by striking ``a dependent" and inserting ``an applicant". SEC. 103. RETROACTIVE BENEFITS.Section 1216(a) of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796d-5(a)) is amended to read as follows: ``(a) Retroactive Eligibility.--Notwithstanding any other provision of law, but subject to the limitations of this subpart, an eligible dependent of a public safety officer shall be eligible for assistance under this subpart if such an officer-- ``(1) dies in the line of duty on or after January 1, 1978; or ``(2) becomes permanently and totally disabled as the direct result of a catastrophic injury sustained in the line of duty on or after January 1, 1978.". SEC. 104. DEFINITIONS. Section 1217 of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796d-6) is amended by adding at the end the following new paragraphs. (…) TITLE II--SURVIVOR PENSIONSSEC. 201. SURVIVOR PENSIONS. Part L of the Omnibus Crime Control and Safe Streets Act of 1968 is further amended by adding after section 1218 (42 U.S.C. 3796d-7) the following new subpart:
(…)``SEC. 1222. PAYMENTS TO BENEFICIARIES. ``(a) Beneficiaries Determined.--An annual pension under this subpart shall be paid to one or more survivors of the deceased public safety officer as follows:
``(1) If there is a surviving spouse of such officer, a pension equal to 80 percent of the applicable amount under section 1223(a), paid to the surviving spouse. (…)(4) If none of the above, a pension equal to 20 percent of the applicable amount under section 1223(a), paid-- ``(A) in the case of a claim made on or after the date that is 90 days after the date of the enactment of this subparagraph, to the individual designated by such officer as beneficiary under this subpart in the officer's most recently executed designation of beneficiary on file at the time of death with such officer's public safety agency, organization, or unit, provided that such individual survived such officer; or ``(B) if there is no individual qualifying under subparagraph (A), to the individual designated by such officer as beneficiary under such officer's most recently executed life insurance policy on file at the time of death with such officer's public safety agency, organization, or unit, provided that such individual survived such officer. (…) TITLE III--PUBLIC SAFETY OFFICER SCHOLARSHIPS SEC. 301. PUBLIC SAFETY OFFICER SCHOLARSHIPS. (a) In General.-- (1) Scholarship awards.--The Secretary of Education is authorized to award a Public Safety Officer scholarship, in accordance with this title, to-- (A) any eligible applicant who is attending, or who has been accepted for attendance at, any eligible institution providing instruction for one or more grades of kindergarten, elementary school, or secondary school; and (B) any eligible applicant who is enrolled, or has been accepted for enrollment, as a full-time or part- time postsecondary student in any eligible institution providing a degree-granting program for one or more postsecondary degrees.(2) Application.—(…) (2) Postsecondary awards.--For any academic year, the maximum amount of a scholarship award under this section for a postsecondary student shall not exceed the lesser of the following: (A) The average cost of attendance (as defined in section 472 of the Higher Education Act of 1965 (20 U.S.C. 1087kk)), (…) SEC. 302. ADDITIONAL AWARD REQUIREMENTS.SEC. 303. AGREEMENTS WITH ELIGIBLE INSTITUTIONS. For the purposes of this title, the Secretary is authorized to enter into agreements with eligible institutions in which any student receiving a scholarship award under this title has enrolled or has been accepted for enrollment.
SEC. 304. TREATMENT OF SCHOLARSHIPS FOR PURPOSES OF FINANCIAL AID. (…) SEC. 305. DEFINITIONS.
In this title: (1) Deceased or disabled officer.--The term ``deceased or disabled officer" means a public safety officer with respect to whom the Bureau of Justice Assistance has determined, in accordance with section 1201 of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796) (…)SEC. 401. COMPENSATION IN CASE OF DEATH. Section 8133(b)(1) of title 5, United States Code, is amended by striking ``or remarries before reaching age 55".SEC. 402. BENEFITS DEFINITION CONFORMING AMENDMENT. Section 1204 of the Omnibus Crime Control and Safe Streets Act of 1968 (42 U.S.C. 3796b) is amended by striking ``As used in this part--" and inserting ``Except as otherwise expressly provided, as used in this part--""

Figure 4: **BillSum comparison (long sample).** The source document contains **2,741 words**; In the qualitative comparison, the **baseline** selects underlined sentences from the source, while the **optimized** version highlights yellow sentences. Both **RoSRL (baseline)** and **RoSRL (Optimized)** compress it to roughly ≈1/5 length. *Factual consistency:* both preserve statutory provisions without adding unsupported claims. *Relevance:* the Optimized version covers a broader set of core sections (adds Scholarships and Survivor Pensions details in addition to Title I), while the baseline focuses on Title I and parts of Title II. *Coherence:* the Optimized version groups related provisions more clearly, whereas the baseline reads like disjoint excerpts.

Reference summary. *Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act – Amends the Omnibus Crime Control and Safe Streets Act of 1968 to extend: (1) educational benefits to public safety officers who become permanently and totally disabled in the line of duty and to their spouses and children; (2) allow payment of retroactive benefits to dependents of such disabled officers; and (3) establish a program of pension payments for certain survivors of deceased public safety officers. Authorizes the Secretary of Education to: (1) award a Public Safety Officer scholarship to disabled public safety officers, their spouses, and their children; and (2) enter into agreements with educational institutions to carry out such scholarship program. Amends federal personnel law to allow widows or widowers of federal employees killed on the job to continue to receive monthly compensation even if they remarry before reaching age 55.*

Source document:
Thousands of people flocked to San Francisco's Golden Gate Bridge on Sunday for a spectacular celebration of the famous landmark's 75th birthday. But a less cheery presence at the festivities was this display of 1,558 shoes representing those who have killed themselves by jumping off the bridge into the San Francisco Bay. (..). 'We're still losing 30 to 35 a people a year off the bridge.' Poignant: These 1,558 pairs of shoes represent all the people who have committed suicide by throwing themselves off the Golden Gate Bridge . Moving: The shoes were installed by the Bridge Rail Foundation, which pushes to cut down on the number of suicides at the bridge . The day-long party attracted pleasure boats, tug boats and other vessels to the waterfront ahead of a magnificent evening fireworks display. Crowds gathered for the exciting events taking place along the shoreline from Fort Point, south of the bridge, to Pier 39 along The Embarcadero. Many walked and biked across the 1.7-mile-long bridge bridge before rounding off the day by watching the nighttime show over the city's enduring symbol. Memorable sight: Fireworks explode over the Golden Gate Bridge as of its 75th anniversary celebrations draw to a close .(…). Since it opened in 1937, more than 2billion vehicles have crossed the mammoth structure. The imposing tourist attraction was named after the Golden Gate Strait, the entrance of water to San Francisco Bay from the Pacific Ocean, which was championed by engineer Joseph Strauss in the 1920s. Dazzling: The steel suspension bridge was bathed in colour and pleasure boats and tugs encircled the water beneath (...). Workers represented by the Inlandboatmen's Union walked off the job on a day strike, forcing the cancellation of ferries operated by Golden Gate between Larkspur, Sausalito and San Francisco. (…) . Shimmering delight: The water was bathed in gold and scarlet as the 18-minute extravaganza brought a day-long party to an end . California Governor Jerry Brown issued a statement on Saturday evening, saying that he was appointing a board to investigate the strike, which he said had disrupted public service. Ms Secchitano disputed the governor's claim, questioning the motivation to call for an investigation after a one-day strike and saying: '[This is] an action to try to silence us. 'They're counting on this process to back our membership off the issue.' Service resumed Sunday when workers returned to work. (…) Eleven men died during construction from 1933 to 1937 - ten of them when scaffolding fell through a safety net that had been set up to protect workers. The conditions were difficult, cold, foggy and windy, and workers who helped construct supports for the south tower had to contend with dangerous tides. The current retrofit project is expected to extend the bridge's lifespan by another 150 years.

RoSRL (baseline Summary):
Thousands of people flocked to San Francisco's Golden Gate Bridge on Sunday for a spectacular celebration of the famous landmark's 75th birthday. Poignant: These 1,558 pairs of shoes represent all the people who have committed suicide by throwing themselves off the Golden Gate Bridge . The day-long party attracted pleasure boats, tug boats and other vessels to the waterfront ahead of a magnificent evening fireworks display. Crowds gathered for the exciting events taking place along the shoreline from Fort Point, south of the bridge, to Pier 39 along The Embarcadero.Workers represented by the Inlandboatmen's Union walked off the job on a day strike, forcing the cancellation of ferries operated by Golden Gate between Larkspur, Sausalito and San Francisco.

RoSRL (optimized Summary): Thousands of people flocked to San Francisco's Golden Gate Bridge on Sunday for a spectacular celebration of the famous landmark's 75th birthday. But a less cheery presence at the festivities was this display of 1,558 shoes representing those who have killed themselves by jumping off the bridge into the San Francisco Bay. 'We're still losing 30 to 35 a people a year off the bridge.' Memorable sight: Fireworks explode over the Golden Gate Bridge as of its 75th anniversary celebrations draw to a close . Since it opened in 1937, more than 2billion vehicles have crossed the mammoth structure. Shimmering delight: The water was bathed in gold and scarlet as the 18-minute extravaganza brought a day-long party to an end .

Reference sumary:"Bridge Rail Foundation erected a moving display of 1,558 pairs of shoes to represent those who have jumped from the bridge to their death .Landmark was heralded as engineering marvel when it opened in 1937 as the Great Depression came to an end ."

Figure 5: **CNN/DailyMail.** Source document: 1,004 words; baseline: 118 words; optimized: 120 words (≈1/8.5 each). Compared to baseline, optimized preserves key facts (75th anniversary, 30–35 suicides/year, over 2 billion vehicles, 18-minute extravaganza..), improves relevance by omitting strike details, and enhances coherence by logically sequencing celebration, memorial, and historical context.

Source document: Bộ Lao động Thương binh và Xã hộiđềxuất nghỉ Tết Âm lịchtừ 26 tháng chạp đến hết mùng 5 tháng giêng (25/1-2/2/2025), gồm 5 ngày nghỉ chính thức, 4 ngày nghỉ cuối tuần. Bộ đã gửi văn bản xin ý kiến 16 cơ quan, bộ ngành về phương án nghỉ Tết Âm lịch 2025 trước khi trình Thủ tướng quyết định. Tôi thấy năm nào vào tầm này cũng có tin đề xuất phương án nghỉ Tết. Rồi sau đó một thời gian mới có tin về lịch nghỉ chính thức. Một bài viết vào năm 2022 giải thích vì saolịch nghỉ Tết khó cố địnhlà do ngày nghỉ chính thức có thể trùng cuối tuần, xen kẽ với ngày làm việc và cần áp dụng nghỉ bù hoặc hoán đổi. Nhiều người nói không nên tiêu hao nhiều thời gian và năng lượng, chỉ để bàn bạc tới lui mỗi lịch nghỉ Tết. >>'Đã đến lúc người Việt được thong thả nghỉ Tết chín ngày' Nhưng theo tôi, ắt hẳn Bộ Lao Động cũng chẳng muốn năm nào cũng xin ý kiến 16 cơ quan, bộ ngành. Theo Luật Lao động 2019, mỗi năm, người lao động được nghỉ 5 ngày Tết âm lịch. Tuy nhiên, trên thực tế, số ngày nghỉ Tết hàng năm đều nhiều hơn 5 ngày này chính nhờ hoán đổi, nghỉ bù. Và luật cũng chỉ quy định số ngày, mà không nói rõ là nghỉ bắt đầu từ ngày nào, nên phải xin ý kiến. Tôi cho rằng, nếu như 5 ngày nghỉ Tết là ít (vì trên thực tế rất nhiều người lao động ở xa đã mất hai ngày đi lại, số ngày ở bên gia đình không nhiều, thì nên đề xuất tăng thêm số ngày nghỉ Tết. Lấy ví dụ như đề xuất số ngày nghỉ Tết Ất Tỵ 2025 là 9 ngày, thì cố định số ngày nghỉ là 9. Sau đó cố định luôn mốc thời gian nghỉ, ví dụ từ 26 tháng Chạp năm cũ đến mùng 5 Tết. Sau Tết, nếu rơi vào ngày cuối tuần thì vẫn đi làm bình thường. Điều này Trung Quốc đã áp dụng. Ví dụ: Lễ Quốc tế Lao động lịch tiêu chuẩn họ được nghỉ một ngày. Tùy thuộc từng năm sẽ có thay đổi, năm 2024, nghỉ từ ngày 1 đến ngày 5/5, tổng cộng có 5 ngày nghỉ. Làm việc bù vào ngày 28/4 (Chủ Nhật) và ngày 11/5 (Thứ Bảy). *Quan điểm của bạn thế nào? Chia sẻ bài viếttại đây. Vĩnh Phan
RoSRL (Baseline Summary): Bộ Lao động Thương binh và Xã hộiđềxuất nghỉ Tết Âm lịchtừ 26 tháng chạp đến hết mùng 5 tháng giêng (25/1-2/2/2025), gồm 5 ngày nghỉ chính thức, 4 ngày nghỉ cuối tuần. Bộ đã gửi văn bản xin ý kiến 16 cơ quan, bộ ngành về phương án nghỉ Tết Âm lịch 2025 trước khi trình Thủ tướng quyết định. Tôi thấy năm nào vào tầm này cũng có tin đề xuất phương án nghỉ Tết. Rồi sau đó một thời gian mới có tin về lịch nghỉ chính thức. Một bài viết vào năm 2022 giải thích vì saolịch nghỉ Tết khó cố địnhlà do ngày nghỉ chính thức có thể trùng cuối tuần, xen kẽ với ngày làm việc và cần áp dụng nghỉ bù hoặc hoán đổi.
RoSRL (Optimized Summary): Bộ Lao động Thương binh và Xã hộiđềxuất nghỉ Tết Âm lịchtừ 26 tháng chạp đến hết mùng 5 tháng giêng (25/1-2/2/2025), gồm 5 ngày nghỉ chính thức, 4 ngày nghỉ cuối tuần. Bộ đã gửi văn bản xin ý kiến 16 cơ quan, bộ ngành về phương án nghỉ Tết Âm lịch 2025 trước khi trình Thủ tướng quyết định. Tôi thấy năm nào vào tầm này cũng có tin đề xuất phương án nghỉ Tết. Một bài viết vào năm 2022 giải thích vì saolịch nghỉ Tết khó cố địnhlà do ngày nghỉ chính thức có thể trùng cuối tuần, xen kẽ với ngày làm việc và cần áp dụng nghỉ bù hoặc hoán đổi. Tôi cho rằng, nếu như 5 ngày nghỉ Tết là ít (vì trên thực tế rất nhiều người lao động ở xa đã mất hai ngày đi lại, số ngày ở bên gia đình không nhiều, thì nên đề xuất tăng thêm số ngày nghỉ Tết.
Reference Summary: Bộ Lao động đề xuất nghỉ Tết Âm lịch 2025 từ 26 tháng Chạp đến mùng 5 tháng Giêng (9 ngày, gồm 5 ngày nghỉ chính thức và 4 ngày cuối tuần), đang chờ ý kiến các bộ ngành trước khi trình Thủ tướng. Tác giả cho rằng thay vì hàng năm xin ý kiến về lịch nghỉ Tết, nên xem xét tăng số ngày nghỉ Tết lên 9 ngày cố định và cố định thời gian nghỉ, tương tự như cách Trung Quốc thực hiện, để người lao động có thời gian nghỉ ngơi trọn vẹn.

Figure 6: **VNExpress.** Both summaries preserve factual accuracy on key points: the holiday period (lunar Dec 26–Jan 5; Jan 25–Feb 2, 2025; 5 official + 4 weekend days), the consultation with 16 agencies before submission to the Prime Minister, and the reason the schedule is not fixed (overlap with weekends requiring adjustments). The optimized version adds the author's viewpoint that 5 days are insufficient due to travel time, suggesting an extension, thus adding information and improving narrative coherence compared to the baseline.

Online Information Extraction System (EAOS) for Social Media Comments

Khanh-Hung Huynh[1,2], Phuc Bui Hoang Gia[1,2], Huyen Dinh Doan My[1,2], Phi-Long Nguyen[1,2], Thien-Vu Nguyen-Ho[1,2], Trong-Hop Do[1,2],

[1]University of Information Technology,
[2]Vietnam National University Ho Chi Minh city
Corresponding author: hopdt@uit.edu.vn

Abstract

An EAOS (Entity–Aspect–Opinion–Sentiment) information extraction system is proposed to analyze Vietnamese user comments on the YouTube platform. This work is the first to address a unified multi-task framework for Vietnamese social media content, jointly extracting entities, aspect categories, opinion spans, and sentiment—contrasting with prior studies that focused only on single tasks such as sentiment analysis or aspect-based sentiment analysis. The data is collected from highly interactive videos containing various opinions about products, services, and social topics. The system includes an offline module for data preprocessing, model construction, and training using deep learning techniques, and a prediction module for applying the trained model to real-world data. In the offline phase, the pretrained PhoBERT model is used to encode the contextual semantics of text, followed by classification layers that simultaneously predict sentiment, aspect category, and the start–end positions of both entity and opinion spans within each sentence. The dataset is normalized and manually annotated in the EAOS format, then divided into training, development, and testing sets. Experimental results show that PhoBERT outperforms traditional baseline methods, especially in accurately identifying aspect-related and opinion-bearing expressions. To ensure practical applicability, the trained models are deployed on an Apache Spark–based streaming architecture, enabling the system to process large-scale and continuous social media data in real time. The system is designed to be flexible and scalable, enabling adaptation to other social media platforms such as Facebook or TikTok. These findings emphasize both the pioneering nature of this research in Vietnamese multi-task information extraction and the effectiveness of transformer-based models in extracting complex information from Vietnamese social media content.

1 Introduction

In the current digital age, social media serves as a significant medium for expressing public opinions, emotions, and evaluations regarding products, services, and public figures. YouTube, as one of the most popular platforms, generates a continuous stream of user comments, providing a rich and real-time data source for opinion mining and sentiment analysis. The task addressed in this study involves the extraction of EAOS (Entity – Aspect – Opinion – Sentiment) structures from Vietnamese YouTube comments, with the goal of identifying the mentioned subject, the specific aspect being discussed, the expressed opinion, and the corresponding sentiment. To the best of our knowledge, this is the first work to address a unified multi-task framework for Vietnamese social media content, jointly extracting entities, aspect categories, opinion spans, and sentiment, rather than focusing on a single task as in prior studies such as sentiment analysis or aspect-based sentiment analysis.

The EAOS framework includes four core components: Entity, which refers to the object being mentioned (such as an artist, program, or product); Aspect, which denotes the aspect or topic associated with the entity (such as appearance, personality, or expertise); Opinion, which is the textual phrase reflecting the user's judgment; and Sentiment, which captures the polarity of the opinion (positive, negative, or neutral). The task requires models to detect and correctly associate these elements within each comment, considering that a single sentence may contain multiple EAOS tuples or express components in implicit ways.

Unlike conventional sentiment analysis approaches that focus on classifying the sentiment of entire sentences, the EAOS task demands a more fine-grained understanding of sentence semantics and structure. Additionally, the real-time and large-scale nature of YouTube comments intro-

duces the need for a system capable of processing data instantly as it is posted. To meet both the accuracy and scalability requirements, our proposed EAOS extraction system integrates deep learning techniques with Apache Spark Streaming, enabling large-scale, real-time processing of social media data. This architecture allows the trained models to detect entities, aspects, opinions, and sentiments in newly submitted comments without delay.

Designed for applications in social media monitoring and decision support, the system not only captures general public sentiment but also provides detailed insights into specific aspects of interest. By combining a pioneering multi-task approach, a manually annotated EAOS dataset for Vietnamese, and a scalable streaming-based deployment, this work lays the foundation for practical, real-time EAOS extraction from Vietnamese-language social media and contributes to advancing semantic understanding technologies in this domain.

2 Related Work

2.1 EAOS Extraction Systems

Over the past decade, the field of Aspect-Based Sentiment Analysis (ABSA) has attracted significant attention from the research community, with various approaches proposed to jointly extract and associate semantic components such as entities, aspects, opinions, and sentiments. Key subtasks have evolved from basic ones like Aspect Extraction (AE), Aspect-Based Sentiment Classification (ABSC), and Aspect-Based Opinion Extraction (ABOE), to more complex joint extraction tasks involving triplets and quadruples, such as ACSTE (Aspect–Category–Sentiment Triplet Extraction), AOSTE (Aspect–Opinion–Sentiment Triplet Extraction), and ACOSQE (Aspect–Category–Opinion–Sentiment Quadruple Extraction). Notable contributions include the work of Yunsen Xian et al., who introduced a detailed sentiment extraction framework for Entity–Aspect– Opinion–Sentiment quadruples, and Hongjie Cai et al., who proposed leveraging implicit aspect/opinion components to improve the system's completeness and accuracy. Recent approaches have also explored semi-supervised learning with incomplete annotations, such as iACOS, and context-aware conditional tagging strategies, as in the work of Wang et al., enhancing the practical applicability of EAOS extraction in real-world scenarios.

In the Vietnamese context, although research in this domain remains limited, notable progress has been made. The UIT-ViSD4SA dataset, developed by the University of Information Technology – VNU-HCM, comprises over 11,000 user reviews annotated with ten distinct aspect categories. Additionally, datasets from the VLSP shared tasks in 2016 and 2018 have significantly contributed to the development of named entity recognition, text classification, and sentiment analysis systems for Vietnamese. On the modeling side, PhoBERT—a Vietnamese-specific variant of BERT—has demonstrated strong performance across various natural language processing tasks and is employed as the core encoder in the proposed EAOS architecture to fully exploit semantic information in Vietnamese texts. Real-Time Prediction Systems In recent years, a growing number of studies have explored the integration of machine learning techniques with big data processing platforms to build real-time prediction systems for social media data. Most of these systems focus on sentiment analysis, disease prediction, or anomaly detection based on streaming tweets or comments. For instance, Elzayady et al. (2018) utilized Apache Spark in combination with machine learning models for real-time sentiment analysis, optimizing preprocessing and model selection for improved accuracy. Ahmed et al. (2020) proposed a real-time system using Apache Spark and Apache Kafka to predict heart disease risk from tweets, evaluating multiple models including Decision Trees, SVM, and Random Forest, and applying grid search to select the optimal model. Zaki et al. (2020) developed a framework for collecting, processing, and visualizing Twitter data to analyze the real-time psychological state of Iraqi citizens, while Kilinc (2019) focused on fake account detection on Twitter using Spark MLlib. Another application by Kabir et al. (2020) involved tracking tweets from the United States during the COVID-19 pandemic to examine changes in topics, sentiment intensity, and subjectivity.

Although these studies demonstrate the effectiveness of combining Spark with machine learning for real-time social data analysis, most of them remain limited to overall sentiment classification and do not delve into finer-grained elements such as entities, aspects, or opinions. The proposed study extends this direction by building a real-time EAOS extraction system for Vietnamese YouTube comments. This system not only detects sentiment but also identifies and links textual spans represent-

ing entities, aspect categories, and opinions within each comment. By integrating deep learning models with big data frameworks such as Apache Spark, the system aims to enable fine-grained, real-time sentiment analysis from continuously updated social media streams, offering a novel direction for multidimensional opinion mining in dynamic online environments.

3 Real-Time EAOS Extraction System

The proposed real-time EAOS (Entity–Aspect–Opinion–Sentiment) extraction system consists of two main components: (i) an offline-trained EAOS extraction model and (ii) a streaming comment processing pipeline for applying the trained model to real-world data in real time. The overall system architecture is illustrated in Figure 1.

3.1 Offline Module: EAOS Extraction Model

Given the high velocity and volume of continuously generated social media data, traditional processing systems often fall short in terms of performance, scalability, and real-time responsiveness. This limitation is particularly evident in natural language processing tasks that require instant analysis and feedback upon data arrival. To address these challenges, the EAOS extraction system in this study is built upon Apache Spark, an open-source distributed computing framework that supports in-memory parallel processing and real-time data stream handling through the Spark Structured Streaming library.

Apache Spark is selected for its scalability, high processing speed, and rich ecosystem that supports numerous high-level libraries. In this system, Spark SQL is used for handling structured data, facilitating key text preprocessing steps such as filtering, normalization, and organization before feeding the data into the model. Structured Streaming is employed to construct a continuous data pipeline, ingesting real-time YouTube comments collected via the platform's official API. For deep learning model integration, PySpark, the Python API for Spark, enables efficient connection between the offline-trained PhoBERT model and the data stream being processed within Spark.

Specifically, incoming YouTube comments undergo a cleaning process that includes duplicate removal, text normalization, and the elimination of special characters and emojis. The cleaned data is then passed into the PhoBERT model to simultaneously predict four components: sentiment, aspect (aspect category), and the start and end positions of spans corresponding to entity and opinion expressions. These results are aggregated and visualized in real time, creating a continuous and automated loop from data collection to user feedback analysis on social media.

The use of Apache Spark not only ensures high-speed processing and scalability in the face of increasing data volume but also allows seamless integration with modern deep learning models. The combination of Spark's big data processing capabilities and PhoBERT's strong language representation power results in a stable, accurate, and efficient system tailored to the real-world characteristics of Vietnamese social media data.

3.2 Entity–Aspect–Opinion–Sentiment Extraction Module

3.2.1 Training Dataset

The dataset was constructed within the scope of this research to support the task of EAOS (Entity–Aspect–Opinion–Sentiment) extraction, where each data instance reflects one or more user opinions under popular entertainment videos on YouTube. Input data consists of user comments collected via Google Apps Script, which interfaces with the YouTube API to retrieve the latest comment threads. After collection, the data undergoes a preprocessing pipeline designed to remove noise and normalize the text. Specifically, the system eliminates non-linguistic elements such as hashtags, URLs, emojis, pictographic characters, and special symbols. The text is then converted to lowercase, punctuation is separated, duplicated characters are removed, and `vnCoreNLP` is applied for word segmentation and lexical normalization.

Manual annotation was carried out by a single annotator to maintain consistency, following detailed guidelines. Each comment may contain multiple EAOS quadruples, with each consisting of four components: entity, aspect (aspect category), opinion, and sentiment. Aspect categories are divided into five predefined groups: `APPEARANCE`, `CHARACTERISTIC`, `SPECIALIZE`, `TECHNICAL`, and `OTHER`. Sentiment is labeled as `positive`, `negative`, or `neutral`. For implicit components, the start–end index pair is assigned as $(-1, -1)$. The dataset is stored in tabular format, with each row containing the comment text,

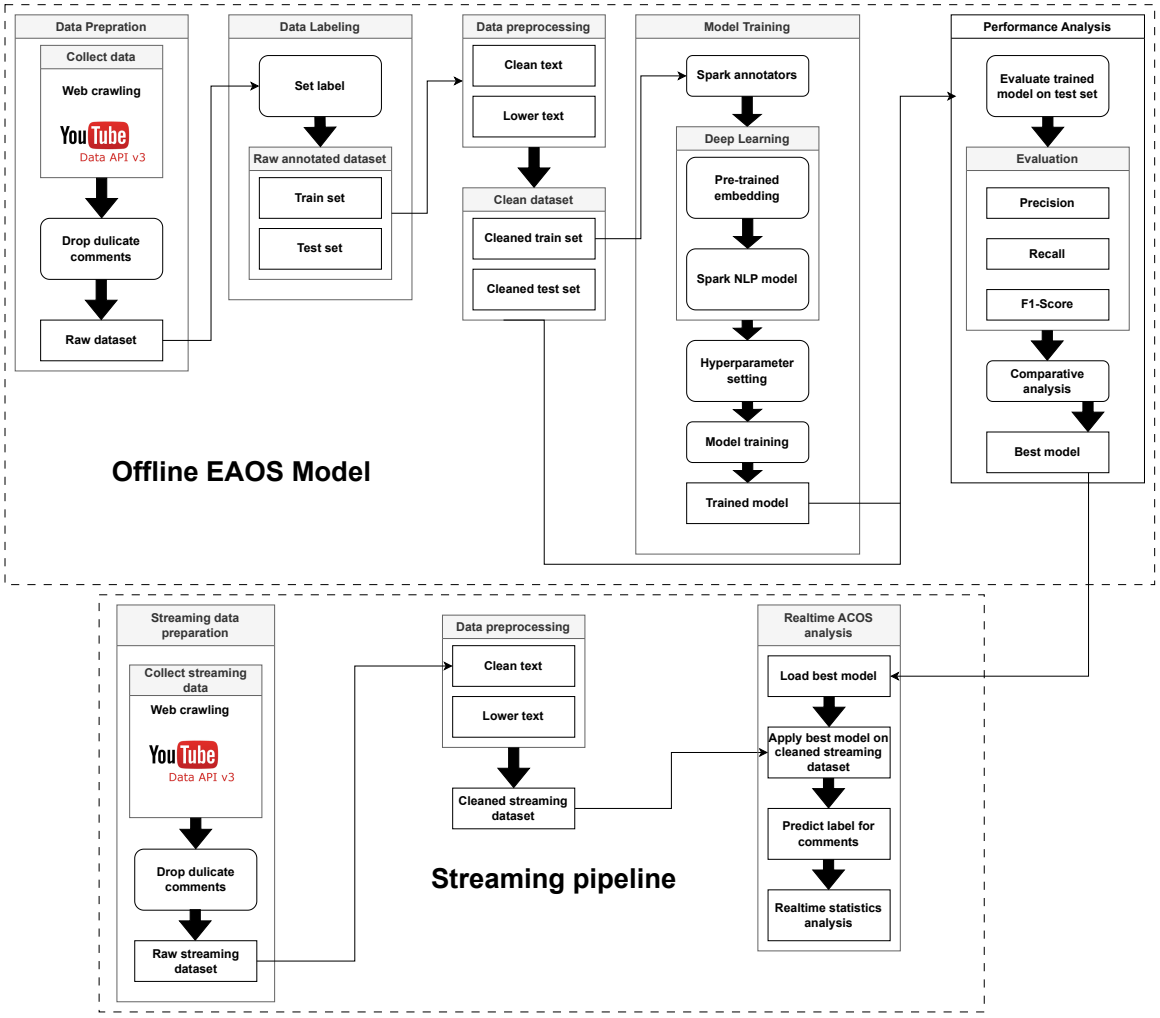


Figure 1: Overall EAOS extraction pipeline: offline training and real-time inference.

span positions, and the corresponding labels.

After annotation and data augmentation, the dataset contains a total of 10,065 comments, split into 9,065 training samples, 500 development samples, and 500 testing samples. Tables 1 and 2 summarize the distribution of sentiment and aspect categories in the training and development sets.

Table 1: Distribution of sentiment labels.

	Positive	Negative	Neutral
Train (9,065)	7,428	3,145	948
Dev (500)	402	142	69

Overall, the training set exhibits a distribution of 64.5% positive, 27.3% negative, and 8.2% neutral sentiment. Among the five aspect categories, "Specialize" accounts for the largest proportion, followed by "Other," while the remaining three categories are more evenly distributed.

Table 2: Distribution of aspect categories.

Aspect Category	Train	Dev
Specialize	3,800	240
Other	2,907	194
Appearance	1,675	106
Technical	1,574	47
Characteristic	1,565	26

3.2.2 Preprocessing

Text preprocessing is essential, especially when dealing with social media data. The preprocessing pipeline removes emojis, mentions, hashtags, URLs, emails, and technical markers. Special characters, numbers, and extra whitespaces are also eliminated. The text is then lowercased, and `vnCoreNLP` is used for word segmentation and lemmatization.

Given the complexity of Vietnamese, which

is monosyllabic and context-sensitive, tools like vnCoreNLP improve segmentation accuracy and grammatical preservation. Comments that are too short or too long are removed to ensure high-quality training data.

3.2.3 Deep Learning Models

LSTM and BiLSTM Baselines. LSTM and BiLSTM are used as baseline architectures for contextual learning. LSTM captures long-term dependencies but is limited to one-directional context. BiLSTM addresses this by processing input in both directions, which benefits tasks like EAOS where components may depend on future or past context.

EAOS Architecture. An advanced model combines PhoBERT, BiLSTM, Multi-Head Attention, and Graph Neural Networks (GNN). Input comments are encoded by PhoBERT, generating contextual vectors for each token. These vectors are passed through BiLSTM and Self-Attention layers.

The model performs three tasks: (1) extract candidate entities, (2) extract candidate opinions, (3) pair them into entity–opinion pairs. Both explicit and implicit expressions are captured. A GNN then models relationships between these pairs and predicts the aspect and sentiment labels.

The model outputs valid quadruples (e_i, a_j, o_k, s_l), where e_i is the entity, a_j the aspect, o_k the opinion, and s_l the sentiment. End-to-end training minimizes error propagation and improves overall accuracy.

This integrated model leverages PhoBERT for Vietnamese semantic encoding, BiLSTM for bidirectional context, Attention for focusing on important tokens, and GNN for relational modeling in EAOS extraction.

3.3 Online Module

3.3.1 Real-Time Comment Collection from YouTube

Google Apps Script is used in conjunction with the YouTube Data API to automatically collect comments from videos belonging to entertainment programs such as Rap Viet, 2 Days 1 Night, and The Masked Singer. The system periodically checks for new videos and downloads newly posted comments, removing duplicates to ensure data uniqueness. These comments are temporarily stored before the next processing stage. The retrieval process operates continuously via HTTP connections and employs OAuth functions for user authentication and secure data access.

3.3.2 Real-Time EAOS Analysis

After data collection, Spark Structured Streaming is used to process and stream comments to the pre-trained model from the offline phase. Each comment undergoes a preprocessing pipeline involving removal of special characters, emojis, and noise terms, normalization via vnCoreNLP, word segmentation, and standard text formatting. The processed text is then tokenized using the PhoBERT tokenizer and transformed into feature vectors.

These vectors are fed into the EAOS model, which may include components such as PhoBERT, BiLSTM, Attention, or GNN, depending on the selected architecture. The model simultaneously predicts: (i) entity and opinion spans, (ii) aspect category labels for each entity, and (iii) sentiment labels indicating positive, negative, or neutral attitudes. The extracted quadruples (e_i, a_j, o_k, s_l) are aggregated and stored in JSON or database format for statistical analysis.

The system then performs real-time statistical analyses such as: comment counts by sentiment type, distribution of user-focused aspects, frequency of specific entities or opinions over time, and shifts in viewer sentiment. These insights offer a comprehensive view of audience feedback on entertainment content and can assist content producers and media managers in making informed decisions.

4 Evaluation and Experimental Results

4.1 Offline Evaluation

Evaluation Metrics. To assess the effectiveness of the EAOS extraction models, we adopt three standard metrics commonly used in classification and sequence labeling tasks: **Precision**, **Recall**, and **F1-score**. These metrics are crucial in tasks with label imbalance and demand precise identification of EAOS components.

- **Precision** measures the proportion of correct positive predictions out of all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

  where TP (True Positive) is the number of correct positive predictions, and FP (False Positive) is the number of incorrect positive predictions.

- **Recall** measures the proportion of actual positives that are correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where FN (False Negative) is the number of missed positive instances.

- **F1-score** is the harmonic mean of precision and recall:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

It balances the trade-off between precision and recall, and is particularly suitable for multi-task, imbalanced settings like EAOS.

These metrics are applied not only to classification components (e.g., Apest and Sentiment) but also to span detection components (Entity and Opinion), where token-level accuracy and span-level coverage are measured.

Model	Accuracy	Precision	Recall	F1-score
LSTM	32.52%	82.64%	32.52%	18.91%
LSTM + Attention	44.97%	52.28%	44.97%	40.56%
BiLSTM	53.14%	52.45%	53.14%	50.27%
BiLSTM + Attention	56.46%	57.76%	56.46%	53.26%
EAOS (Proposed)	**68.14%**	**69.23%**	**68.14%**	**68.28%**

Table 3: Experimental results comparing different models.

The table above presents the performance of different model architectures. The basic LSTM model achieves the lowest accuracy at 32.52%, despite a high precision of 82.64%, indicating a highly conservative prediction style that misses many true cases. Introducing Attention significantly improves recall (from 32.52% to 44.97%) and nearly doubles the F1-score to 40.56%.

The BiLSTM model, benefiting from bidirectional context encoding, shows better recall and F1-score (50.27%). Adding Attention further boosts performance to an accuracy of 56.46% and F1-score of 53.26%, marking the best among RNN-based models.

The final EAOS model, combining PhoBERT, BiLSTM, Attention, and GNN, achieves superior results across all metrics: accuracy of 68.14%, precision of 69.23%, recall of 68.14%, and an F1-score of 68.28%. This confirms the effectiveness of integrating contextualized language models and graph-based relational reasoning for comprehensive EAOS extraction.

In conclusion, the proposed EAOS model demonstrates substantial improvement over traditional architectures, validating the value of combining powerful representation layers, attention mechanisms, and graph structures in complex information extraction tasks.

4.2 Online Evaluation

Real-time Data Collection Speed The system collects comments from YouTube using the platform's API. During deployment, it takes approximately 5–7 minutes to process and store every 100 comments, depending on comment length, API response speed, and real-time data availability. This speed ensures timely data flow for real-time analysis.

EAOS Real-time Statistics From a total of 1,288 EAOS quads extracted in real-time, we compute the distribution of categories under different sentiment labels as shown in Table 2.

Sentiment	Appearance	Characteristic	Specialize	Technical	Other	Total
Negative	3	15	37	13	98	166
Positive	60	34	153	4	658	909
Neutral	0	7	2	4	200	213
Total	63	56	192	21	956	1288

Table 4: Category distribution by sentiment in real-time EAOS extraction

As shown, positive sentiment dominates (909 out of 1,288 EAOS quads, or 70.6%), which aligns with the nature of entertainment content on YouTube. "Other" and "Specialize" are the most frequent categories, indicating vague or skill-related user feedback.

Negative sentiment comments often focus on "Other" and "Specialize", suggesting viewer dissatisfaction tends to center on performance or less-defined categories. Neutral sentiment is mostly associated with the "Other" category, indicating descriptive or unclear expressions.

Overall, these statistics confirm the system's ability to reflect user reactions and trend shifts in real-time, supporting media monitoring and decision-making based on public feedback.

4.2.1 Error Analysis on Aspect Categorization

One notable observation is that a relatively high proportion of aspect labels fall into the `OTHER` category (approximately 25% of all annotations). Upon closer inspection, we found that many of these cases involve comments that are vague, multi-faceted, or outside the predefined set of four concrete categories (`APPEARANCE`, `CHARACTERISTIC`,

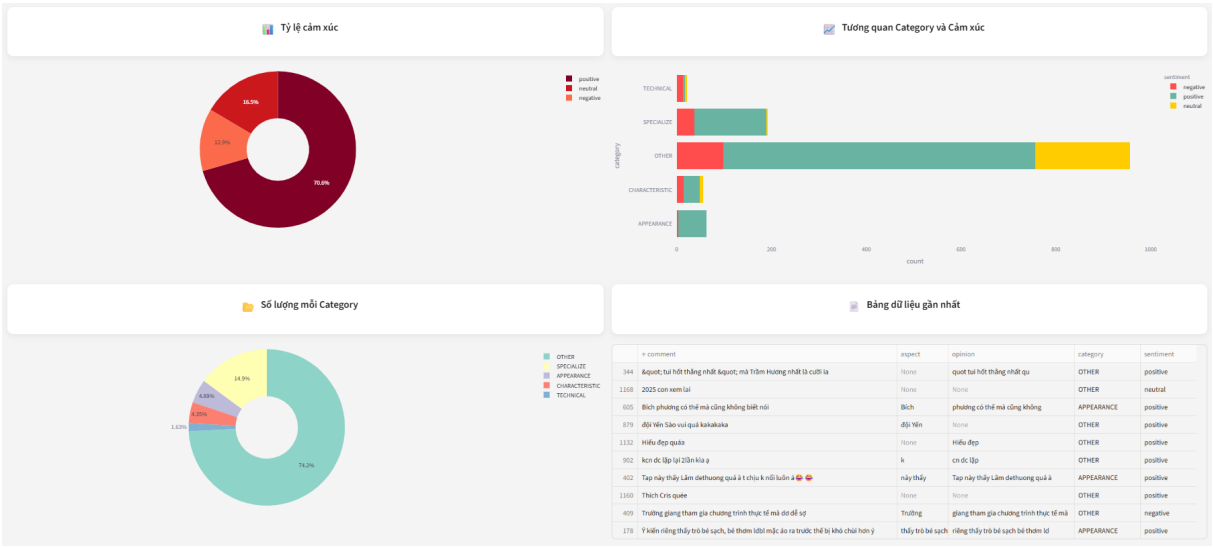


Figure 2: Real-time EAOS system dashboard illustrating sentiment distribution, category frequency, the correlation between sentiment and category, and a table of the most recent extracted data.

SPECIALIZE, TECHNICAL). For example, general praise such as 'Hay quá!" (So good!") or expressions of emotion like Thích quá tri" (Really love it") do not explicitly refer to an identifiable aspect and are therefore placed in OTHER. This phenomenon reduces the usefulness of the system for fine-grained analysis. In future work, we plan to refine the annotation guidelines, introduce sub-categories within OTHER, and perform additional manual annotation to reduce the proportion of ambiguous cases. Such refinements are expected to improve both the interpretability and the practical value of the aspect extraction component.

5 Conclusion and Future Work

This work proposed and implemented a pioneering real-time information extraction system for Vietnamese social media comments, addressing the EAOS (Entity–Aspect–Opinion–Sentiment) task in a unified multi-task framework. To the best of our knowledge, this is the first system for Vietnamese that simultaneously detects entities, aspect categories, opinion spans, and sentiment within a single architecture—contrasting with previous studies that typically focused on individual subtasks such as sentiment analysis or aspect-based sentiment analysis. The system comprises two main components: an offline module for model training and an online module for large-scale, real-time EAOS extraction.

The proposed model leverages PhoBERT's contextual semantic representations combined with task-specific linear layers for category classification, sentiment classification, and span extraction. The dataset—collected from popular entertainment programs, normalized, and manually annotated in the EAOS format—enabled effective fine-tuning of the model. Evaluation results showed that the proposed EAOS model outperformed strong baselines (LSTM, BiLSTM, Attention-based models), achieving an accuracy of 68.14% and an F1-score of 68.28%. These findings confirm the advantages of a transformer-based multi-task architecture for fine-grained sentiment analysis in Vietnamese.

The online component collects new user comments from YouTube via API and processes them using an Apache Spark Streaming–based pipeline, enabling large-scale, continuous data streams to be analyzed in real time. Experiments demonstrated an average latency of 50–60 seconds from comment posting to analysis completion. The system also provides real-time dashboards with statistics on aspects, sentiments, and opinions, offering actionable insights for marketing, brand tracking, and community feedback monitoring. With its flexibility and scalability, the architecture can be extended to other platforms such as Facebook, TikTok, and online forums, serving as a practical tool for decision support.

Limitations and Future Work — Despite these promising results, several limitations remain. First, the dataset is currently restricted to the entertainment domain, which may affect generalizability to other domains such as product reviews or news comments. Second, a relatively high proportion

of aspects were labeled as "OTHER", suggesting the need for finer-grained annotation and further error analysis. Third, annotation was performed by a single annotator, which, although consistent, does not capture inter-annotator agreement. Finally, large language models (LLMs) were not considered in this study due to computational constraints; integrating such models represents an important direction for future work.

Future extensions will therefore focus on (i) expanding the dataset to multiple domains and ensuring higher annotation reliability, (ii) adapting the system to additional social media platforms, (iii) integrating it into real-time monitoring applications such as brand tracking and public opinion analysis, and (iv) exploring multilingual and cross-lingual settings. These directions aim to enhance the system's robustness, coverage, and applicability in real-world scenarios.

Acknowledgement

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D4-2025-14.

References

[1] Hongjie Cai, Rui Xia, Jianfei Yu. Aspect–Category–Opinion–Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL 2021), pp. 340–350. 10.18653/v1/2021.acl-long.29

[2] Dan Ma, Jun Xu, Zongyu Wang, Xuezhi Cao, Yunsen Xian. Entity–Aspect–Opinion–Sentiment Quadruple Extraction for Fine-grained Sentiment Analysis. *arXiv preprint* arXiv:2311.16678 (2023).

[3] Wenya Zhang, Xin Li, Yang Deng, Lidong Bing, Wai Lam. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *arXiv preprint* arXiv:2203.01054 (2022).

[4] Soni S., Rambola A. A Comprehensive Review of ACOSQE with Implicit Aspects and Opinions. *Artificial Intelligence Review* (2023). 10.1007/s10462-023-10633-x

[5] Loris Di Quilio, Fabio Fioravanti. Evaluating the Aspect–Category–Opinion–Sentiment Analysis Task on a Custom Dataset. In: *CEUR Workshop Proceedings* (2022).

[6] Dang Van Thin, Lac Si Le, Vu Xuan Hoang, Ngan Luu-Thuy Nguyen. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. *arXiv preprint* arXiv:2103.09519 (2021).

[7] Dat Quoc Nguyen, Anh Tuan Nguyen. PhoBERT: Pre-trained Language Models for Vietnamese. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037–1042. 10.18653/v1/2020.findings-emnlp.92

[8] Hieu Nguyen, Linh Le, et al. ViSoBERT: A Pre-trained Language Model for Vietnamese Social Media Text Processing. *arXiv preprint* arXiv:2310.11166 (2023).

[9] Omar Elzayady, Xue Li, Mohamed Farouk. Real-Time Sentiment Analysis on Twitter Data Streams using Apache Spark. *International Journal of Data Science*, vol. 5, pp. 45–60 (2018).

[10] Mohammed Ahmed, Sara Al Murl. Real-time Heart Disease Risk Prediction from Twitter Data Using Apache Spark and Kafka. *Health Informatics Journal*, vol. 26, no. 2, pp. 1012–1028 (2020).

[11] Md. Kabir, Jane Smith. COVID-19 Twitter Sentiment Analysis using Apache Spark Streaming. In: *Proceedings of the 8th International Conference on Big Data Analytics* (2020).

[12] Nguyen Luong Chi, Nguyen Thi Minh Huyen, Nguyen Cam Tu, Le Hong Phuong. Vietnamese Open Information Extraction. *arXiv preprint* arXiv:1801.07804 (2018).

On the Role of Contextual Information and Ego States in LLM Agent Behavior for Transactional Analysis Dialogues

Monika Zamojska and Jarosław A. Chudziak
Faculty of Electronics and Information Technology
Warsaw University of Technology
Warsaw, Poland
{monika.zamojska.stud, jaroslaw.chudziak}@pw.edu.pl

Abstract

LLM-powered agents are now used in many areas, from customer support to education, and there is increasing interest in their ability to act more like humans. This includes fields such as social, political, and psychological research, where the goal is to model group dynamics and social behavior. However, current LLM agents often lack the psychological depth and consistency needed to capture the real patterns of human thinking. They usually provide direct or statistically likely answers, but they miss the deeper goals, emotional conflicts, and motivations that drive real human interactions. This paper proposes a Multi-Agent System (MAS) inspired by Transactional Analysis (TA) theory. In the proposed system, each agent is divided into three ego states — Parent, Adult, and Child. The ego states are treated as separate knowledge structures with their own perspectives and reasoning styles. To enrich their response process, they have access to an information retrieval mechanism that allows them to retrieve relevant contextual information from their vector stores. This architecture is evaluated through ablation tests in a simulated dialogue scenario, comparing agents with and without information retrieval. The results are promising and open up new directions for exploring how psychologically grounded structures can enrich agent behavior. The contribution is an agent architecture that integrates Transactional Analysis theory with contextual information retrieval to enhance the realism of LLM-based multi-agent simulations.

1 Introduction

Rapid progress in Large Language Models (LLMs) has enabled the development of conversational agents that are increasingly deployed in areas requiring human-like social interaction (Önder Gürcan, 2024). These include customer service, educational tutoring (Wang et al., 2024b), and healthcare applications (Morrow et al., 2023; Chen et al., 2025). The potential to extend these capabilities into social simulations is significant and offers a range of benefits to researchers (see Figure 1). However, even as the agents' abilities are impressive (Mittelstädt et al., 2024), they still exhibit responses that lack the psychological depth and behavioral consistency characterizing human communication (Frisch and Giulianelli, 2024). These agents typically generate statistically probable responses based on their training data, but they fail to capture the underlying emotional motivations, internal conflicts, and unconscious behavioral patterns that are necessary for authentic social interactions (Bail, 2024).

To address this gap, this paper proposes a novel Multi-Agent System (MAS) architecture that integrates principles from Transactional Analysis (TA), a well-established psychological framework for un-

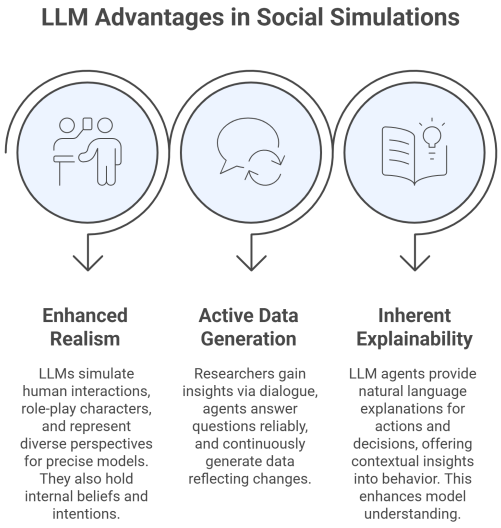


Figure 1: Key advantages of using LLM-based agents for social simulations, summarized from the analysis (Önder Gürcan, 2024).

derstanding human behavior and interpersonal communication (Stewart and Joines, 2012). The proposed approach models each agent as a complex system consisting of three distinct *ego states* — Parent, Adult, and Child — each representing different knowledge structures (Tosi and Bianchini, 2013; Horowitz, 1991) with their own psychological perspectives and information processing styles. This architecture attempts to incorporate the deeper psychological mechanisms that influence how people interpret social situations, access relevant information, and formulate responses based on their internal emotional states.

The key contribution of this work is the development and evaluation of a framework that combines TA-structured *ego states* with contextual information retrieval mechanisms to improve the psychological realism of LLM-based agent interactions. Using controlled experiments that compare agents with and without access to memory banks (Zhong et al., 2024), the study demonstrates that this approach leads to more complex, emotionally grounded, and psychologically consistent behaviors. The findings suggest that explicit modeling of internal psychological structures, combined with targeted information retrieval, represents a promising direction for developing more human-like conversational agents capable of authentic social interaction.

2 Background and Related Work

Making LLM agents behave more realistically in social interactions involves two key areas of consideration. The first is understanding human thought and communication. The second is developing agent architectures that can effectively reproduce these observed human patterns. The following section discusses these points.

2.1 Transactional Analysis for Structuring Agent Behavior

Transactional Analysis (TA) is a psychological theory offering a structured way to understand human interactions and behavior (Berne, 1958; Stewart and Joines, 2012). While initiated by Eric Berne, TA continues to evolve. Central to TA is the model of three '*ego states*' — Parent, Adult, and Child — each representing distinct patterns of thinking, feeling, and behaving. Other researchers have pointed out that these *ego states* can be seen as structures that hold meaning and integrate knowledge (Tosi and Bianchini, 2013; Horowitz, 1991), store memories (Novey, 1998), and even work like connected neural networks (Joines, 2016; Schiff, 1981). Each of these *ego states* has a unique knowledge and information processing style (see Figure 2):

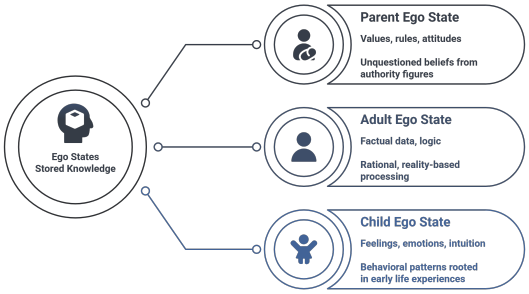


Figure 2: Conceptual model of the three *ego states* — Parent, Adult, and Child — and their associated stored knowledge, as described in Transactional Analysis (Berne, 1958; Stewart and Joines, 2012).

- The Parent *ego state* reflects behaviors, thoughts, and emotions adopted from parental figures. This includes a knowledge base of messages about social rules and moral values.
- The Adult *ego state* acts as rational knowledge processor. It focuses on facts, logical thinking, and understanding the current reality.
- The Child *ego state* consists of behaviors, emotions, and thought patterns developed in early childhood, often based on needs and fears. It draws upon a store of emotional experiences, focusing on feelings and spontaneity.

TA posits that long-term patterns of behavior are often navigated by an individual's '*life script*'. A *life script* is an unconscious life plan, developed in childhood through a complex interplay of factors (Berne, 1972). It guides decisions, shapes relationships, and often manifests in repetitive patterns, reinforcing beliefs about oneself and the world.

In TA, social interactions are called '*transactions*' — exchanges of information that occur between individuals' *ego states*. The nature of these *transactions* significantly impacts communication flow. For example, if a response originates from an unexpected *ego state*, a *crossed transaction* occurs, often causing confusion or conflict. In contrast, when a response comes from the *ego state* that was targeted, the *transaction* is considered *complementary*, and communication typically proceeds

smoothly. Transactions involving a hidden psychological message can lead to psychological '*game*' (recurring patterns of nonconstructive *transactions*) (Berne, 2016).

Another important concept is '*discounting*' - an unconscious process of ignoring or disqualifying certain information. *Discounting* is often linked to certain *ego states*, especially when a person reacts with fear or rigid beliefs. Taken together, TA provides a rich framework for conceptualizing how individuals structure, store, and process information, and how it guides their behavior in social interactions.

2.2 LLM-based Multi-Agent Systems (MAS)

Large Language Models (LLMs) have enabled the creation of intelligent agents capable of engaging in rich, human-like interactions (Gao et al., 2024; Zamojska and Chudziak, 2025b; Wang et al., 2024a). These agents can generate context-aware responses, demonstrate social reasoning, and adapt to evolving conversational dynamics (Dolant and Kumar, 2025; Frisch and Giulianelli, 2024). A Multi-Agent System (MAS) combines multiple such agents, each with its own perspective and role, into a shared environment (Kostka and Chudziak, 2024). Recent work has focused on exploring applications of LLM-based MAS in debate (Taubenfeld et al., 2024; Harbar and Chudziak, 2025), virtual town simulation (Huang et al., 2025; Park et al., 2023), and social network formations (Zhang et al., 2024; Takata et al., 2024).

To achieve realistic interactions, modern architectures incorporate more than just language capabilities. Memory management allows agents to recall past interactions and ensure consistent behavior (Chen et al., 2024). Memory is typically split between short (in the LLM context window) and long-term storage (managed externally using vector databases or similar techniques) (Zhong et al., 2024; Huang et al., 2024a). In addition, reflection and planning modules help agents handle feedback, analyze their memories, and change strategies, based on how humans process information (Yao et al., 2023). These components help ensure that agents can simulate conversations and group dynamics that are more psychologically reliable (Kostka and Chudziak, 2025; Huang et al., 2024b).

3 A TA-Structured Architecture for Simulating Social Dynamics

Our approach to simulating nuanced social dynamics is realized through an agent architecture grounded in Transactional Analysis (Zamojska and Chudziak, 2025a). The agent is created as a system of interacting components. TA's *ego states* (see Section 2.1) are modeled as distinct knowledge-processing modules (Parent, Adult, and Child), each equipped with its own dedicated memory bank. Given a conversational context, each module retrieves the most similar past memory (if exists) and proposes a potential response. Then, a final decision-making process, performed by an overarching LLM agent, guided by the *life script*, selects the most contextually appropriate response from the proposals.

The overall agent behavior can be defined as a function:

$$R = D(\{r_p, r_a, r_c\}, S, C) \tag{1}$$

where:

- R is the final response.
- r_i is the response from the i-th *ego state* ($i \in$ Parent (p), Adult (a), Child (c)$\}$).
- S is the agent's life script.
- C is the current conversational context.
- D is the decision mechanism that selects the response R.

3.1 Ego State Sub-Agents

The foundation of the architecture lies in its representation of an agent's personality through the Parent (E_p), Adult (E_a), and Child (E_c) *ego state* modules. Technically, each module is an independent LLM-powered ReAct agent (Yao et al., 2023), utilizing the GPT-4o model (OpenAI, 2024). Behavior is shaped through a specific system prompt (P_i, where $i \in \{p, a, c\}$) defining its persona and information processing style:

- The **Parent** module (E_p), driven by prompt P_p, reflects authority and rules.
- The **Adult** module (E_a), via prompt P_a, represents logical, objective decision-making.
- The **Child** module (E_c), through prompt P_c, embodies emotions and reacts based on needs and fears.

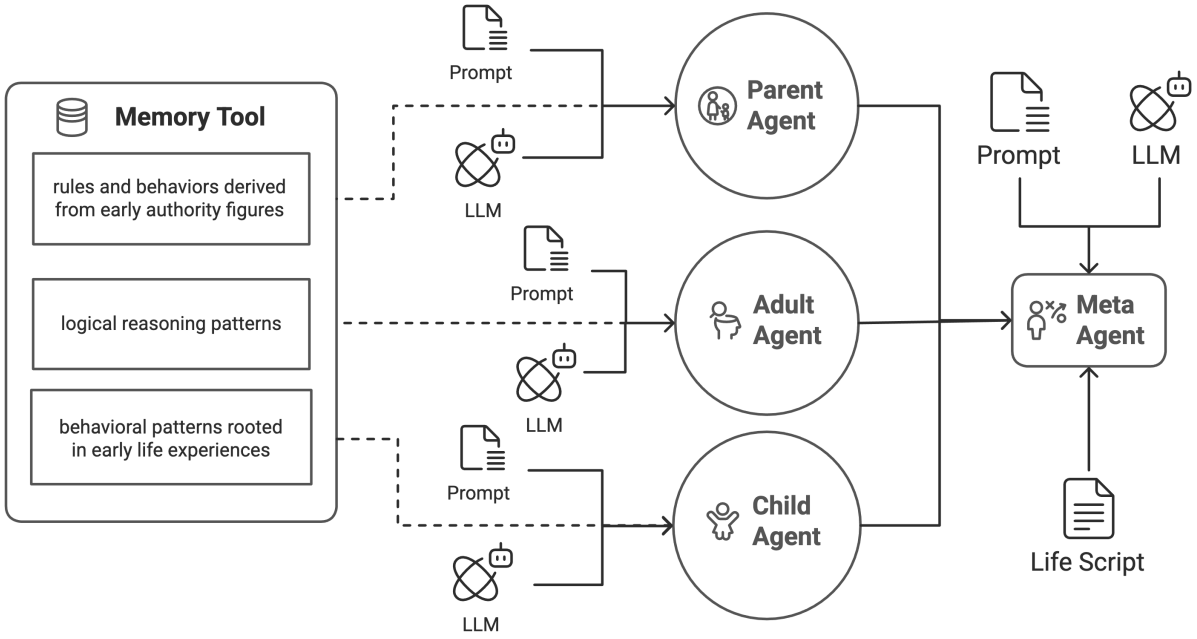


Figure 3: Agent architecture. Each agent consists of three sub-agents — Parent, Adult, and Child — each driven by a distinct prompt and (in the memory-enabled condition) a retrieval-augmented vector memory. At each turn, all sub-agents generate candidate responses based on the current conversational input.

Each *ego state* module E_i generates its potential response r_i, guided by the module's specific system prompt P_i, the current conversational context C, and the relevant information retrieved m_i from its memory bank M_i, see Section 3.2. This is expressed as:

$$r_i = \text{LLM}(P_i, C, m_i) \tag{2}$$

where $\text{LLM}(\cdot)$ signifies the process of generating text conditioned on the provided data.

3.2 Memory as Contextual Information Retrieval

Each *ego state* module E_i ($i \in \{p, a, c\}$) can actively augment its knowledge by querying its dedicated memory bank M_i. This is implemented as a tool available to each *ego state*. The memory banks M_i store information corresponding to its characteristic knowledge base, as detailed in Section 2.1). Each memory item is structured as a JSON record containing context (description of a past situation or interaction), successful reaction, associated emotions, and proper tone of response. The textual context is indexed as embeddings in a FAISS (Facebook AI Similarity Search) vector database (Johnson et al., 2021). The reaction, emotions, and tone are stored as metadata associated with the embedding.

During its ReAct reasoning cycle, an *ego state* module E_i can decide to invoke this tool by formulating a natural language query q_i based on its current conversational context C). The memory retrieval step selects a set of top-k memories:

$$m_i = \arg\max_{m \in M_i} \cos(\text{Embed}(q_i), \text{Embed}(m)) \tag{3}$$

where:

- $\text{Embed}(\cdot)$ represents the embedding function for semantic similarity.
- $\cos(\cdot, \cdot)$ denotes the cosine similarity between the context and memory embeddings.
- m_i are the memory items retrieved for the *ego state* i.
- q_i is query sent by *ego state* i.

The retrieved m_i is returned to E_i and incorporated into its subsequent reasoning and response generation (r_i in Equation 2).

4 Experimental Design

This section outlines the experimental setup designed to evaluate the impact of *ego states* (see Section 2.1) and contextual information on the behavior of LLM agents engaged in dialogues simulating Transactional Analysis (TA) principles. The

```
You are John, a junior software developer with defensive tendencies. Your Life Script is: "I Almost Make It" or "I Never Quite Succeed". You have a pattern of starting well but faltering at crucial moments, often due to internal disorganization or a subconscious fear of success/completion.

Your Life Position is: "I'm Not OK, You're OK." You tend to see others (especially authority figures) as competent and yourself as inherently flawed or less capable, particularly under pressure.

Your Primary Drivers are:

    1. Try Hard: You focus on the effort you put in, sometimes in a disorganized way, rather than efficient completion. You want to be seen as trying.

    2. Please Me: You want to be liked and avoid disapproval, but your other patterns often sabotage this.
```

Figure 4: The prompt defining the life script (S) for the agent John. This script guides the agent's decision-making process, shaping its behavior to align with an "I Almost Make It" pattern and the internal conflict of hiding procrastination.

experiment aims to observe and compare agent responses in a defined scenario under two distinct conditions: with and without memory access.

4.1 Scenario Design

The scenario selected for this experiment is a common workplace interaction designed to underline characteristic *ego state* responses. The setting is a Monday morning project update meeting. The characters involved are Taylor, the Project Lead, whose core motivations are driven by a "Must Be In Control and Perfect" *life script*. She feels like maintaining high standards and managing situations meticulously is only way to feel secure and validated. John, a key team member, operating under an "I Almost Make It" *life script* (see Figure 4). He repeatedly comes close to achieving a goal or success but ultimately falls short at a crucial moment, often due to internal disorganization, self-sabotage, or a subconscious fear of completion. The core conflict arises from John's failure to submit a critical Q3 data analysis report. This non-completion is caused by John's procrastination and lack of focus during the preceding week.

4.2 Experimental Conditions

To evaluate the impact of contextual information on agent behavior, experiments were conducted under two distinct conditions. For each condition, 22 dialogues were simulated, with each dialogue consisting of 4 conversational turns per agent. This resulted in a total of 88 responses per agent being collected for analysis in each setup.

The first condition, **Memory OFF**, involves agents operating without access to the memory bank. The agents (Parent, Adult, Child) will generate responses based only on their initial detailed prompts.

The second condition, **Memory ON**, involves agents utilizing their information retrieval tools (see Section 3.2). In this setup, each of the three *ego state* agents (Parent, Adult, Child) for both John and Taylor has access to its dedicated memory system with predefined memory items.

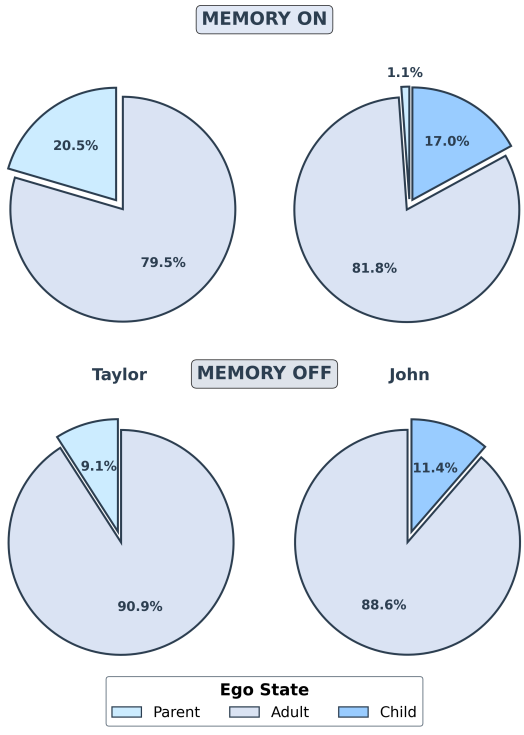


Figure 5: Distribution of Ego State Selection, Parent (P), Adult (A), and Child (C), for Agents Taylor (left column) and John (right column). The top row shows the results for the Memory ON condition, where agents had access to contextual information from their memory banks. The bottom row shows the results for the Memory OFF condition.

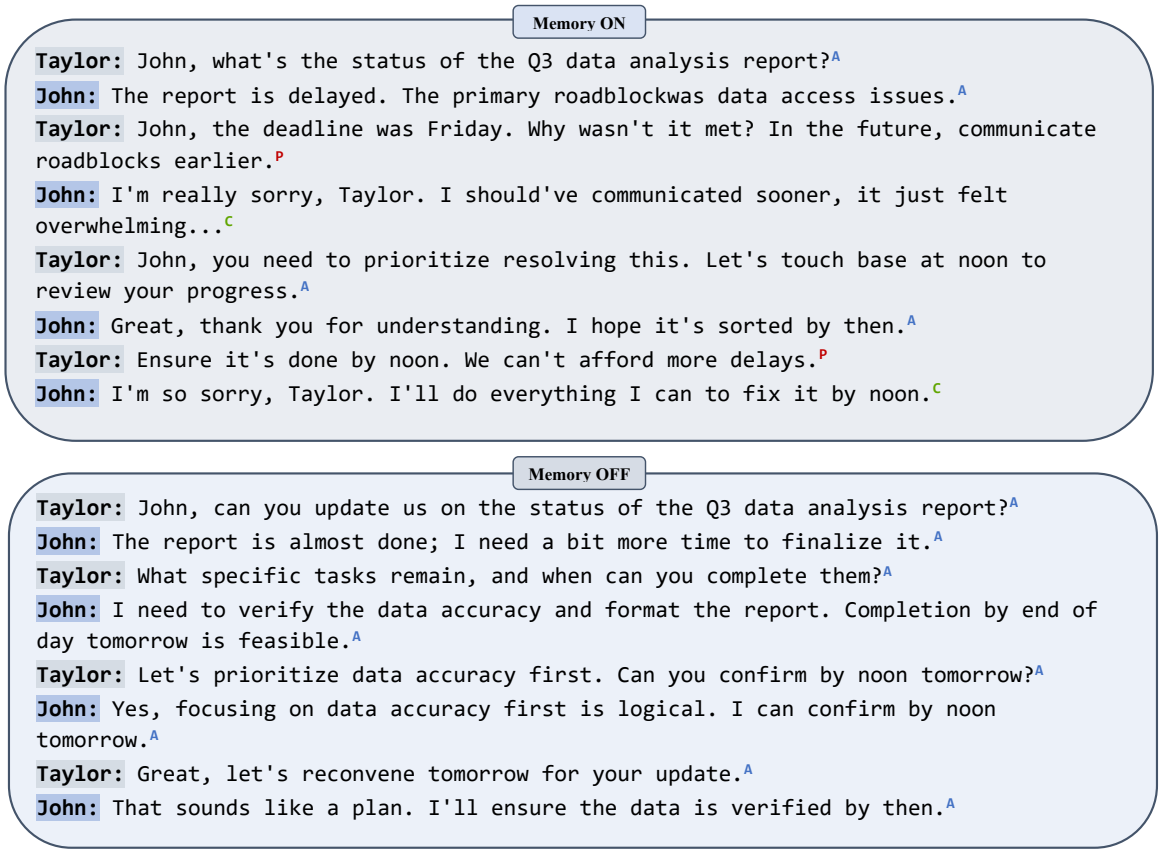


Figure 6: Comparative dialogues illustrating LLM agent behavior with information retrieval enabled (ON, top) versus disabled (OFF, bottom). *Ego state* activations (P: Parent, A: Adult, C: Child) are indicated for each statement.

4.3 Key Findings

A primary observation is that the explicit modeling of Parent, Adult, and Child *ego states* allowed agents to move beyond a default, often rational or solution-oriented, LLM behavior. Even in the **Memory OFF** condition, agents demonstrated the capacity to generate responses reflecting distinct *ego state* perspectives. However, when memory retrieval was enabled (**Memory ON**), agent responses became more nuanced, consistent with their character, and emotionally grounded (see comparative dialogue examples in Figure 6).

Across both **Memory ON** and **OFF** conditions, the Adult *ego state* was frequently selected by the meta-decision process for both Taylor and John. This is likely influenced by the professional work setting of the simulated scenario, where rational communication (characteristic of the Adult state) is often the expected norm. This indicates that while the architecture supports diverse *ego state* expression, the conversational context and the nature of the task heavily influence which *ego state* is chosen.

The ablation study (**Memory ON** vs. **OFF**) highlighted that access to contextual information influenced the distribution of selected *ego states*. Specifically, in the **Memory ON** condition, John exhibited an increase in Child *ego state* responses (from 10 to 15 of his turns), while Taylor's engagement from her Parent more than doubled (from 8 to 18) (see Figure 5 for detailed distributions). This shift suggests that the retrieved information provides stronger, more specific cues for producing a response that is more psychologically consistent and grounded. Such a memory-enhanced response becomes a much more compelling candidate for the meta-decision LLM, as it better aligns with the agent's core *life script*. This leads to a higher selection rate of non-Adult *ego states* and more dynamic interactions.

The increased activation of non-Adult states directly fostered the conditions for a more frequent Parent-to-Child dynamic. For instance, when Taylor communicated from her Parent, her messages were inherently more critical and evaluative. This type of input is a trigger for John's Child state, whose *life script* is centered on feelings of inadequacy. This pattern, where a change in an *ego state*

by one agent prompts a complementary *ego state* shift in the other, was also observable in **Memory OFF** condition, but less frequently due to the limited diversity of ego state selections. Future research will aim to make such *complementary* (and *crossed*, see Section 2.1) *transactions* more explicit within the simulation's logic and analysis.

5 Discussion

While the initial results from applying the architecture are promising, we acknowledge that this research contains certain limitations which provide directions for future research. The current evaluation focuses mainly on qualitative analysis within a single dialogue scenario. This approach restricts how broadly these conclusions can be applied in different types of social interaction. Another important limitation concerns the memory component - these were predefined rather than developed through interaction experiences.

Based on observations, next research efforts will target multiple important areas to improve the proposed system and overcome current shortcomings. We intend to include more Transactional Analysis (TA) concepts like *discounting* (see Section 2.1), *strokes* (small units of recognition that satisfy the need to be noticed), and *stamp collecting* patterns (accumulation of negative emotions) to make agent interactions more psychologically realistic (Stewart and Joines, 2012). The next essential step involves the development of a more transparent, algorithmic mechanism to replace the current LLM-based selection of the final response. This new mechanism could incorporate a weighting system, where the answer is influenced by real-time conversational metrics, like accumulated 'emotional stamps' leading to a build-up of frustration. The final response could then be generated as a fusion of *ego state* outputs, with each contribution proportional to its calculated weight, which would better simulate the internal psychological conflicts of human decision-making.

For improving memory functionality, we want to investigate approaches that enable agents to automatically generate and modify their *ego state* memories during conversations. This might involve using reinforcement learning techniques to determine what experiences should be remembered and how these memories affect future responses. Most importantly, conducting broader testing with different scenarios and possibly including TA practitioner judges will be necessary to properly evaluate the advantages and complexities of this psychology-based agent design.

6 Conclusion

This paper has presented a novel approach to enhance the psychological realism of LLM-based agents through the integration of Transactional Analysis theory with contextual information retrieval mechanisms. The experimental evaluation demonstrates that modeling agents as composite systems of Parent, Adult, and Child *ego states* leads to more nuanced and psychologically grounded interactions compared to traditional LLM agents. The ablation study reveals that memory-enabled agents exhibit more diverse *ego state* activations. While the initial results are promising, several limitations are acknowledged including single scenario validation and reliance on predefined memory content, which present opportunities for future research. The implications of this research extend beyond technical improvements to LLM agents. Grounding agent behavior in established psychological theory opens new possibilities for applications in social science research, educational simulations, and therapeutic contexts.

Acknowledgments

The work reported in this paper was partly supported by the Polish National Science Centre under grant 2024/06/Y/HS1/00197.

References

Christopher A. Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

E. Berne. 1958. Transactional analysis: A new and effective method of group therapy. american journal of psychotherapy. *American Journal of Psychotherapy*, 12(4):735—-743.

E. Berne. 1972. *What Do You Say After You Say Hello?: The Psychology of Human Destiny*. Bantam books. Grove Press.

E. Berne. 2016. *Games People Play*. Penguin Life.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8(1):159.

Antoine Dolant and Praveen Kumar. 2025. Agentic LLM framework for adaptive decision discourse. *Preprint*, arXiv:2502.10978.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11.

Yarolsav Harbar and Jaroslaw A. Chudziak. 2025. Simulating oxford-style debates with LLM-based multi-agent systems. In *Intelligent Information and Database Systems*, pages 286–300, Singapore. Springer Nature Singapore.

Mardi Jon Horowitz. 1991. *Person schemas and maladaptive interpersonal patterns*. The University of Chicago Press.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024a. Emotional RAG: Enhancing Role-Playing Agents through Emotional Retrieval . In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 120–127, Los Alamitos, CA, USA. IEEE Computer Society.

Yizhe Huang, Xingbo Wang, Hao Liu, Fanqi Kong, Aoyang Qin, Min Tang, Song-Chun Zhu, Mingjie Bi, Siyuan Qi, and Xue Feng. 2025. Adasociety: An adaptive environment with social structures for multi-agent decision-making. *Preprint*, arXiv:2411.03865.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, and Xiangliang Zhang. 2024b. Social science meets LLMs: How reliable are large language models in social simulations? *Preprint*, arXiv:2410.23426.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vann S. Joines. 2016. Understanding second-order structure and functioning. *Transactional Analysis Journal*, 46(1):39 – 49.

Adam Kostka and Jarosław A. Chudziak. 2024. Synergizing logical reasoning, knowledge management and collaboration in multi-agent LLM system. In *Pacific Asia Conference on Language, Information and Computation (PACLIC 38)*, Tokyo, Japan.

Adam Kostka and Jarosław A. Chudziak. 2025. Towards cognitive synergy in llm-based multi-agent systems: Integrating theory of mind and critical evaluation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Justin Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14.

Elizabeth Morrow, Teodor Zidaru, Fiona Ross, and 1 others. 2023. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. *Frontiers in Psychology*, 13.

T Novey. 1998. A proposal for an integrated self [letter to the editor]. *The Script*, 28(7):6.

OpenAI. 2024. Hello GPT-4o. Accessed: 2025-04-05.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. Association for Computing Machinery.

JL Schiff. 1981. Ego states. In *Workshop at the Southeast Institute Annual Spring Conference, Raleigh, NC*.

I. Stewart and V. Joines. 2012. *TA Today: A New Introduction to Transactional Analysis*, 2nd edition. Lifespace Publishing, Nottingham.

Ryosuke Takata, Atsushi Masumori, and Takashi Ikegami. 2024. Spontaneous emergence of agent individuality through social interactions in large language model-based communities. *Entropy*, 26(12).

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 251–267. Association for Computational Linguistics.

Maria Teresa Tosi and Susanna Bianchini. 2013. A social-cognitive definition of ego states to implement ta research. *International Journal of Transactional Analysis Research & Practice*, 4(1):107–112.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. 2024a. A survey on human-centric LLMs. *Preprint*, arXiv:2411.14491.

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024b. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252:124167.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *International Conference on Learning Representations (ICLR)*.

Monika Zamojska and Jaroslaw A. Chudziak. 2025a. Games agents play: Towards transactional analysis in LLM-based multi-agent systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Monika Zamojska and Jaroslaw A. Chudziak. 2025b. Simulating human communication games: Transactional analysis in LLM agent interactions. In *Recent Challenges in Intelligent Information and Database System*, pages 173–187, Singapore. Springer Nature Singapore.

H. Zhang, J. Yin, M. Jiang, and C. Su. 2024. Can agents spontaneously form a society? introducing a novel architecture for generative multi-agents to elicit social emergence. *Preprint*, arXiv:2409.06750.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Önder Gürcan. 2024. LLM-augmented agent-based modelling for social simulations: Challenges and opportunities. In *Hybrid Human AI Systems for the Social Good*, pages 134–144.

Enhancing Scientific Title Generation via Optimized Sentence Ordering

Thanh-Thien Khuu[1,2], Thien-Thuan Huynh[1,2], Nam Van Chi[1,2], Tung Le[1,2,*]
[1]Faculty of Information Technology, University of Science, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
{ktthien22,htthuan22}@clc.fitus.edu.vn
{vcnam, lttung}@fit.hcmus.edu.vn
[*]Corresponding author: Tung Le - lttung@fit.hcmus.edu.vn

Abstract

Generating concise titles for machine learning abstracts is essential for navigating complex literature but challenging due to specialized terminology and dense text structures. We propose a novel sentence ordering method that uses a two-stage sequence-to-sequence BART framework, augmented by an auxiliary model that extracts sentences with the highest keyword overlap to the title and ranks candidate sentence permutations. The optimal ordering guides BART to produce coherent and concise titles. We evaluate our method on the test split of a large dataset of over 21,000 machine learning title–abstract pairs from Springer journals. Results show that structured input via optimized sentence ordering improves title quality compared to baseline models. These findings highlight sentence ordering as an underexplored yet effective strategy for enhancing scientific text generation.

1 Introduction

Machine learning (ML) research has surged in recent years, with Springer journals publishing thousands of abstracts that distill cutting-edge advances, from neural architecture search to reinforcement learning. Titles play a pivotal role in this ecosystem, serving not only as entry points for researchers but also as essential elements for indexing and retrieval across academic platforms like SpringerLink. Crafting a title that concisely conveys complex technical contributions remains a non-trivial task, especially when abstracts contain dense, domain-specific terminology (e.g., "attention mechanisms") and exhibit structural dispersion, where key ideas are embedded across multiple, non-adjacent sentences. These characteristics pose significant challenges for automated title generation systems, which must distill and reorganize salient content into coherent and informative summaries.

In this work, we utilize BART, a denoising autoencoder (Lewis et al., 2019), for this title generation task. While previous methods typically treat abstracts as flat, unordered text, our approach introduces a two-stage sequence-to-sequence pipeline that incorporates sentence-ordering cues to improve the semantic coherence and informativeness of generated titles. By doing so, we propose a novel strategy that bridges document structure awareness with abstractive summarization, paving the way for more effective automated indexing of scientific literature.

Our method begins with input structuring, where a RoBERTa-based auxiliary model (Liu et al., 2019) is trained to identify important sentences from each abstract using heuristic labels derived from keyword overlap with the gold title. A permutation scoring module subsequently ranks up to `max_permutation` shuffled sentence orderings according to their similarity to the gold title embedding. The permutation that receives the highest score, which reflects coherence and relevance to the title, is selected as the input. This refined input is then used in the title generation stage, where a BART model is trained to produce scientific titles. By conditioning the generator on input that is more coherent and focused, the model learns to generate concise and accurate titles. Training and evaluation are carried out on a curated dataset of machine learning title and abstract pairs obtained from Springer journals. Our results demonstrate that incorporating sentence reordering not only improves standard metrics such as ROUGE and BERTScore, but also enhances semantic alignment with reference titles. This two-stage approach highlights the value of structural preprocessing in abstractive title generation and offers a practical framework for improving title quality in scientific publishing workflows.

Our contributions are as follows:

- We present a novel two-stage method that integrates sentence selection and ordering using RoBERTa and BART to improve scientific title generation in the machine learning domain.

- We construct and release a new dataset of over 21k machine learning research title-abstract pairs collected from Springer journals.[1] [2]

- We demonstrate that our approach improves over a strong BART baseline in both ROUGE and BERTScore, showing consistent gains in title relevance and coherence.

The paper is organized as follows. Section 2 reviews related work, Section 3 details our methodology, Section 4 presents evaluations, Section 5 analyzes input properties, Section 6 discusses the results, and Section 7 concludes.

2 Related Work

2.1 Scientific Title Generation

Scientific title generation is often framed as an extreme summarization task, aiming to distill the essence of a research article or abstract into a concise and informative title. Early approaches relied on statistical and rule-based methods, such as feature-based classifiers that extracted keywords from abstracts to construct titles (Kupiec et al., 1995). These methods, while computationally efficient, struggled with capturing nuanced semantic relationships and often produced formulaic outputs lacking domain-specific precision.

The advent of neural networks marked a significant shift in title generation. Encoder-decoder architectures, particularly recurrent neural networks (RNNs), were employed to generate abstractive summaries and titles (Nallapati et al., 2016). These models improved fluency but were limited by their dependence on sequential processing, which often failed to capture long-range dependencies in complex scientific texts. To address this, retrieval-based methods, such as k-nearest-neighbors (k-NN) approaches, leveraged word co-occurrence patterns in abstracts to propose candidate titles (Putra and Khodra, 2017). While effective for general-purpose texts, these methods often generated generic or overly broad titles when applied to scientific domains, where precision and specificity are paramount.

More recently, pre-trained transformer models, such as GPT-2 and T5, have been fine-tuned for title generation tasks (Riku and Masaomi, 2022). These models generate multiple candidate titles, which are then ranked or refined using heuristic or learned scoring mechanisms. For instance, fine-tuned GPT-2 models have been used to propose diverse title candidates, with post-processing steps to select the most contextually relevant option. However, transformer-based approaches often struggle with domain-specific scientific terminology and may produce titles that lack the innovative or precise phrasing required in academic contexts. Additionally, these models typically treat the input text as a fixed sequence, ignoring the potential benefits of restructuring the input for better coherence or informativeness.

Recent advances have explored the incorporation of domain knowledge into title generation. For example, some approaches integrate scientific ontologies or citation networks to enhance the relevance of generated titles. Others have experimented with hybrid models that combine neural generation with rule-based constraints to ensure adherence to domain-specific conventions. Despite these advances, a key limitation persists: existing methods do not explicitly optimize the input structure (e.g., sentence ordering) to enhance the quality of generated titles. Our work addresses this gap by introducing a novel pipeline that ranks sentence permutations using an auxiliary RoBERTa scorer to guide BART title generation, ensuring that the input structure maximizes coherence and informativeness for scientific title generation.

2.2 Sentence Ordering in NLP

Sentence ordering is a critical task in Natural Language Processing, aimed at arranging sentences to maximize coherence and logical flow, which is particularly important for text generation, summarization, and question answering. Early approaches to sentence ordering relied on heuristic methods, such as ranking sentences based on lexical cohesion or syntactic patterns. However, these methods often failed to capture deep semantic relationships, leading to suboptimal arrangements in complex texts like scientific abstracts.

Neural approaches have significantly improved sentence ordering. For instance, Gong et al. (2016) introduced an end-to-end pointer network to model

[1] Source: https://link.springer.com

[2] Dataset: https://www.kaggle.com/datasets/tiamatt/springerjournal-450tk-0-5cosine

sentence sequences for summarization, achieving better coherence than traditional methods. Similarly, Logeswaran et al. (2017) used RNNs to treat sentence ordering as a coherence optimization problem, leveraging sequential dependencies. More recently, transformer-based models like BERT have employed sentence-level representations, such as their dedicated [CLS] embeddings, to predict optimal orderings through pairwise comparisons or sequence modeling (Devlin et al., 2019). However, to our knowledge, no prior work has directly used [CLS] embeddings to score and rank sentence permutations specifically for sentence ordering.

Our work proposes a novel application of sentence ordering tailored specifically for scientific title generation. Unlike previous approaches, which mainly focus on coherence in summarization or narrative tasks, we use `roberta-base`'s [CLS] embeddings to rank sentence permutations, optimizing the input structure for a `bart-base` title generation model. This allows the generated titles to be not only coherent but also better aligned with the key contributions and domain-specific content of the abstract, addressing an important gap in the literature.

3 Methodology

We present a novel method to enhance title generation by optimizing the sentence order within scientific abstracts. Our approach consists of three key stages: (1) training a `roberta-base` model to identify salient sentences, (2) generating multiple sentence permutations and ranking them , and (3) fine-tuning a `BART-base` model on the highest-ranked permutations to produce coherent and informative titles. This section outlines the full pipeline (see Figure 1), including model architectures, data pre-processing, and training strategies.

3.1 Sentence Selection

To estimate the relevance of a sentence with respect to the title, we train the `roberta-base` (Liu et al., 2019) encoder as a regression model. Each sentence S_i in the abstract is independently encoded using RoBERTa, with a special [CLS] token prepended to the input sequence for sentence-level representation. A scalar relevance score is predicted via a linear head applied to the [CLS] embedding.

Let $h_i \in \mathbb{R}^d$ be the [CLS] embedding vector

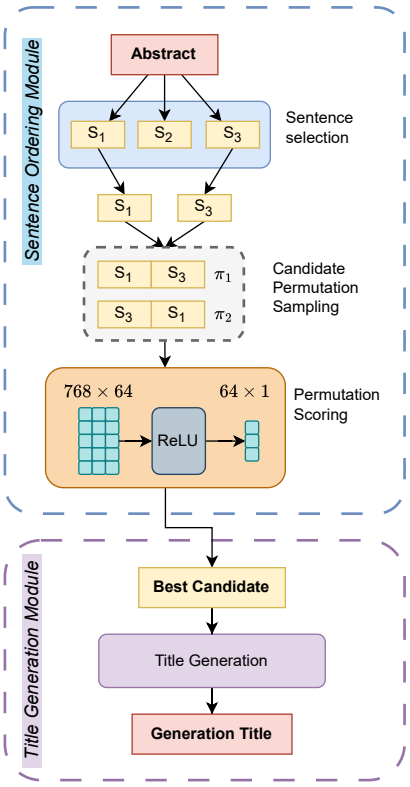


Figure 1: Overview of the title generation architecture with sentence extraction and permutation scoring.

extracted from the last hidden state of RoBERTa for sentence S_i and $y_i \in [0, 1]$ the normalized keyword overlap score (see Equation 8). The model employs a linear head with parameters $W \in \mathbb{R}^{768}$ and $b \in \mathbb{R}$, computing:

$$\hat{y}_i = W^\top h_i + b \tag{1}$$

where the output is a single scalar score without activation to constrain it to $[0, 1]$. The model is trained to minimize the Mean Squared Error (MSE) loss function:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2}$$

where $\hat{y}_i$ is the predicted score and y_i is the ground truth label, and N is the number of sentences in the batch.

We demonstrate our model's sentence selection process using the abstract from Logeswaran et al. (2017) as input:

3.2 Dual Encoder Architecture

Building on the extracted sentences from the previous stage, the dual encoder framework reorders

Table 1: Example of model input and output for sentence extraction in scientific abstracts. The yellow box highlights the original abstract, while the green box contains the sentences selected by the model.

Title:
Sentence Ordering and Coherence Modeling using Recurrent Neural Networks

Original abstract:

Modeling the structure of coherent texts is a key NLP problem. The task of coherently organizing a given set of sentences has been commonly used to build and evaluate models that understand such structure. We propose an end-to-end unsupervised deep learning approach based on the set-to-sequence framework to address this problem. Our model strongly outperforms prior methods in the order discrimination task and a novel task of ordering abstracts from scientific articles. Furthermore, our work shows that useful text representations can be obtained by learning to order sentences. Visualizing the learned sentence representations shows that the model captures high-level logical structure in paragraphs. Our representations perform comparably to state-of-the-art pre-training methods on sentence similarity and paraphrase detection tasks.

Sentences selected by the model:

1. Modeling the structure of coherent texts is a key NLP problem.
2. We propose an end-to-end unsupervised deep learning approach based on the set-to-sequence framework to address this problem.
3. Our model strongly outperforms prior methods in the order discrimination task and a novel task of ordering abstracts from scientific articles.
4. Visualizing the learned sentence representations shows that the model captures high-level logical structure in paragraphs.
5. Our representations perform comparably to state-of-the-art pre-training methods on sentence similarity and paraphrase detection tasks.

these inputs to improve title generation. The architecture integrates a main encoder-decoder for generation and an auxiliary encoder for ranking sentence permutations.

Main Encoder-Decoder. A `bart-base` model serves as the primary sequence-to-sequence component. It takes a linearized sequence of selected sentences as input and generates the corresponding scientific title.

Auxiliary Encoder and Scoring Head. To guide the sentence ordering, a `roberta-base` encoder is paired with a trainable scoring head, which is a two-layer multilayer perceptron (MLP) implemented as $\text{Linear}(768 \to 64) \to \text{ReLU} \to \text{Linear}(64 \to 1)$, where the input is the `[CLS]` embedding of each permutation. The output is a scalar score indicating the estimated quality of the sentence ordering.

Permutation Sampling and Scoring. For each abstract, a set of permutations is sampled from the previously selected sentences by randomly shuffling their order. To maintain computational feasibility, only a fixed number of unique permutations are generated per abstract. Each permutation is then encoded and scored by the auxiliary model to estimate its informativeness and coherence. The permutation receiving the highest score is selected as input to the main encoder-decoder for title generation.

Proxy-Based Ranking Supervision. To supervise the scorer, we use a weak proxy signal based on semantic similarity to the gold title. Each permutation π_i and the gold title t are encoded using the frozen auxiliary encoder, and their `[CLS]` embeddings h_{π_i} and h_t are extracted. The proxy score for each permutation is computed as the cosine similarity:

$$\text{proxy}(\pi_i) = \cos(h_{\pi_i}, h_t) = \frac{h_{\pi_i} \cdot h_t}{\|h_{\pi_i}\| \, \|h_t\|} \quad (3)$$

Despite containing the same content, different sentence orderings lead to distinct contextualized embeddings due to the encoder's positional encodings and attention patterns. Thus, the cosine similarity reflects the degree to which a permutation semantically aligns with the target title.

Prior to computing pairwise supervision signals, both the predicted scores $\phi(h_{\pi_i})$ and the proxy scores $\text{proxy}(\pi_i)$ are sorted in descending order to obtain their respective ranking positions. This ensures that relative pairwise preferences are computed consistently.

To train the scorer, we align its predicted scores with the proxy scores using a differentiable approximation of Kendall's tau rank correlation (Kendall, 1938). We define the pairwise target label for each permutation pair (π_i, π_j) as:

$$y_{ij} = \frac{\text{sign}(\text{proxy}(\pi_i) - \text{proxy}(\pi_j)) + 1}{2} \quad (4)$$

which equals 1 if π_i should be ranked above π_j, and 0 otherwise. The soft prediction of the scorer is:

$$\hat{y}_{ij} = \sigma\left(\frac{\phi(h_{\pi_i}) - \phi(h_{\pi_j})}{\tau}\right) \quad (5)$$

where σ is the sigmoid function, τ is a temperature hyperparameter, and $\phi(h_\pi)$ denotes the scorer's scalar output for permutation π.

The final ranking loss is computed using the binary cross-entropy (BCE) between the predicted and proxy-based pairwise preferences:

$$\mathcal{L}_{\text{kendall}} = \frac{1}{N(N-1)} \sum_{i \neq j} \text{BCE}(\hat{y}_{ij}, y_{ij}) \quad (6)$$

This formulation softly penalizes misalignments between the predicted permutation order and the proxy-induced ranking while remaining fully differentiable.

Overall Training Objective. The final loss combines the main sequence-to-sequence generation objective with the auxiliary ranking supervision. Let $\mathcal{L}_{\text{seq2seq}}$ denote the cross-entropy loss between the generated and gold titles. The complete training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{seq2seq}} + \lambda \mathcal{L}_{\text{kendall}} \quad (7)$$

where λ is a tunable hyperparameter balancing the two components.

4 Experiments and Evaluations

4.1 Dataset

We created a custom dataset for training and evaluating our model by scraping title–abstract pairs from SpringerLink journals in the machine learning domain. The dataset includes 21,545 English-language articles, focusing on subfields like deep learning, neural networks, and natural language processing.

To ensure quality and consistency, we limited abstracts to 128–450 tokens and titles to 8–32 tokens. Poorly aligned title–abstract pairs were filtered out based on cosine similarity between their embeddings (generated via a Sentence-Transformer model), retaining only pairs with a similarity score above 0.5.

For vocabulary and topic consistency across splits, we used stratified sampling based on key machine learning terms (e.g., 'CNN', 'transformers', 'reinforcement learning'). The dataset was split into training (80%, 17,236 samples), validation (10%, 2,154 samples), and test (10%, 2,155 samples) subsets, proportional to each group's size.

Data were sourced from open access metadata and abstracts, with no restrictive licenses from Springer journals prohibiting their use for research at the time of collection.

For sentence extraction, we split each abstract into individual sentences using the Natural Language Toolkit (Loper and Bird, 2002). To determine sentence importance, we adopt a keyword-based scoring method. Specifically, we use KeyBERT (Grootendorst, 2020) to extract keywords from both the gold title and each sentence in the abstract. We define an overlap score between a sentence and the title as:

$$\text{score}(S_i) = \frac{|KW(S_i) \cap KW(T)|}{KW(T)} \quad (8)$$

where $KW(S_i)$ and $KW(T)$ are the sets of keywords extracted for sentence S_i and title T, respectively.

These scores serve as soft labels to supervise a RoBERTa as a scoring model, allowing it to learn to predict sentence relevance in alignment with the title semantics (see Figure 2).

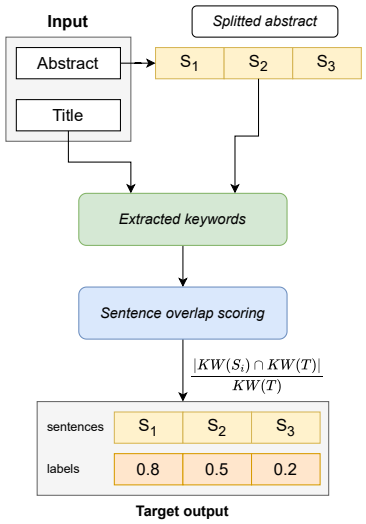


Figure 2: Overview of the data preprocessing pipeline for sentence extraction and labeling.

4.2 Experimental Setup

We perform a manual hyperparameter search using a dataset of 21,545 abstract–title pairs, split into 80% training (17,236), 10% validation (2,154), and 10% testing (2,155), preserving the original distribution. The title generation model is initialized from BART-base, and the sentence scoring model uses a RoBERTa-base encoder. The model is trained for 3 epochs with a batch size of 4, a

learning rate of 2e-5, and gradient accumulation over 8 steps, giving an effective batch size of 32.

For each abstract, the top 7 sentences are selected using the sentence importance module. Up to 30 unique permutations of these sentences are then sampled by random shuffling and scored by the auxiliary encoder. The highest-scoring permutation is used as input to the main encoder-decoder for title generation.

The scorer is trained with soft Kendall's tau loss (see Equation 6), using a temperature $\tau = 0.5$ and a ranking loss weight $\lambda = 1.5$. All experiments are run on a Kaggle-provided Tesla P100 GPU, taking about 6 hours for training the sentence extraction model and 8 hours for the title generation model. To ensure statistical reliability under permutation variability, we repeat each experiment five times without fixed seeds and report the mean and standard deviation of the evaluation metrics.

4.3 Evaluation Metrics

We evaluate generated titles using both lexical overlap and semantic similarity metrics. For lexical evaluation, we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004). ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, respectively, while ROUGE-L captures the longest common subsequence between the generated and reference titles. ROUGE-Lsum is designed for summarization tasks and computes ROUGE-L at the sentence level, which better aligns with abstractive generation tasks such as ours.

To assess semantic similarity, we use BERTScore (Zhang et al., 2020), which computes token-level similarity using contextual embeddings from a pretrained BERT model. We report the precision, recall, and F1 scores, with F1 as the primary semantic metric. BERTScore captures meaning beyond surface form and correlates well with human judgment in generation tasks.

4.4 Results and Analysis

We evaluate the effectiveness of our proposed method by comparing it against both a strong encoder-decoder baseline and several recent instruction-tuned large language models (LLMs). Table 2 reports performance across ROUGE and BERTScore metrics.

Standard BART Baseline. We first compare with a `facebook/bart-base` model trained on full abstracts in their original sentence order, without any sentence filtering or reordering. This model serves as a conventional baseline for scientific title generation. As shown in Table 2, our proposed method consistently outperforms this baseline, demonstrating the benefit of explicitly modeling sentence selection and coherence.

LLM-Based Generation. We compare our method with three instruction-tuned large language models: GPT-4.1 Mini (OpenAI, 2025), Gemini 2.5 Flash (Gemini Team, 2025), and a fine-tuned LLaMA 3.2 (1B) (Grattafiori et al., 2024). For GPT and Gemini, we use a zero-shot approach, prompting each model to create a title from the full abstract with the instruction: *"Write a short, formal and clear title for this scientific research. Return ONLY the title: <abstract>"*.

For LLaMA 3.2 (1B), we evaluate two settings. First, we fine-tune the model on full abstracts without sentence selection or reordering. This improves over zero-shot prompting but remains weaker than BART and our method. Second, we integrate the same sentence selection and ordering strategy as in our proposed method. This variant (*LLaMA 3.2 + Our Method*) yields substantial gains over the plain fine-tuned version, especially in ROUGE and recall-oriented BERTScore, showing that sentence-level control is crucial even for LLMs. However, it still trails the improved BART model, suggesting that encoder-decoder architectures remain better suited for compact title generation under structured input.

As shown in Table 2, the zero-shot LLMs perform significantly worse than the supervised baselines, especially in ROUGE metrics. Their outputs often miss key technical terms or include generic phrasing, which lowers both precision and recall. The fine-tuned LLaMA model performs better than the zero-shot models, showing that task-specific training helps. However, it still lags behind our proposed method in all metrics. This highlights the importance of selecting relevant sentences and presenting them in a coherent order before generation. Our method benefits from this sentence-level control, achieving stronger coverage (high recall) and more accurate phrasing (high precision and F1).

Overall Performance. The results highlight that task-specific supervision and input control (through sentence selection and ordering) are more effective than prompting general-purpose LLMs.

Table 2: Performance comparison on the test set. Results are reported as the mean ± standard deviation over 5 runs. We evaluate using ROUGE-1/2/L and BERTScore (Precision, Recall, and F1).

Model	ROUGE-1	ROUGE-2	ROUGE-L	Precision	Recall	F1
BART-base (baseline)	56.76	36.67	49.33	91.66	90.84	91.23
GPT-4.1 mini	13.41	7.16	10.37	87.83	80.71	84.12
Gemini 2.5 Flash	10.86	6.22	9.00	88.52	80.38	84.24
LLaMA 3.2 1B Instruct	30.39	17.95	26.33	83.00	85.14	83.94
LLaMA 3.2 1B (w/ Method)	44.19	24.89	35.22	87.56	91.28	89.37
Proposed Method	**59.69 ± 0.33**	**38.99 ± 0.24**	**51.66 ± 0.23**	**92.01 ± 0.09**	**91.06 ± 0.06**	**91.52 ± 0.04**

While integrating our method into LLaMA narrows the gap with BART, the encoder-decoder model still achieves the strongest balance of precision, recall, and F1. Our method therefore demonstrates both the utility of structured input and the architectural advantage of supervised sequence-to-sequence learning for scientific title generation.

To complement the quantitative results, we present a qualitative comparison using the same abstract previously introduced in the sentence selection analysis (Section 3.1). The example, shown in the box below (see Table 3), includes the gold title and titles generated by different models. All titles were generated using a maximum length of 32 tokens to ensure a fair comparison across models.

The **gold title** explicitly conveys both the core task (sentence ordering and coherence modeling) and the methodological framework (recurrent neural networks). The **BART baseline** generates a fluent but generic title that lacks task specificity and fails to mention sentence ordering. **Our method (Dual Encoder)** improves on this by directly referencing sentence ordering, thus more accurately reflecting the research focus, albeit in a simpler phrasing.

Among the LLMs, **GPT-4.1 mini** produces the most faithful and specific title. It correctly identifies both the set-to-sequence modeling approach and the sentence ordering task, resulting in a well-structured and informative title that closely aligns with the abstract. **Gemini 2.5 Flash**, while fluent and coherent, shifts focus toward sentence representation learning and omits the sentence ordering aspect, partially reflecting the abstract. **LLaMA-3.2 1B Instruct** captures the main task, sentence ordering, and introduces the concept of coherence, but phrases it in broader terms. While its title is shorter and more abstract than the others, it still reflects key ideas and avoids hallucinating unrelated terms.

The **LLaMA-3.2 1B Instruct + Our Method** variant shows how structured input control influences generation. Its output is more detailed and task-relevant than the plain LLaMA version, explicitly highlighting abstract ordering and positioning the method as unsupervised deep learning. However, by adding elements like "sentence similarity" and "paraphrase detection" not present in the abstract, it becomes informative but less precise than the gold reference.

These examples show the trade-offs across models in terms of task relevance, specificity, and fluency. LLMs like GPT and Gemini generate polished and expressive outputs, but may emphasize secondary elements or reframe the task. LLaMA, while more concise, remains grounded in the core ideas. In contrast, our model offers a more targeted and faithful summary of the task, striking a balance between relevance and simplicity without introducing content outside the source abstract.

5 Input Property Analysis

To study how input characteristics affect model performance, we examined two properties of the abstracts: the number of sentences and the total token count. We used a fixed test set and measured the relationship between these properties and the performance of the model using ROUGE and BERTScore. The results are shown in Table 4.

Across all ROUGE variants, we find a weak but statistically significant negative correlation with both token count and sentence count (r between -0.048 and -0.071, $p < 0.05$). This suggests that longer abstracts tend to slightly reduce lexical overlap with the reference titles. In contrast, BERTScore does not show a significant correlation with either measure ($p > 0.05$), indicating that semantic similarity is largely unaffected by abstract length.

6 Discussion

Our results show that adding sentence selection and ordering to a standard text generation model can

Table 3: Qualitative comparison of generated titles from various models. The yellow highlights denote outputs from task-specific transformer models (e.g., BART baseline and our dual-encoder method), while the green highlights indicate outputs from general-purpose large language models (LLMs), including GPT, Gemini, and LLaMA.

Gold Title:
Sentence Ordering and Coherence Modeling using Recurrent Neural Networks

GPT-4.1 mini:

Unsupervised Deep Set-to-Sequence Modeling for Coherent Text Structure and Sentence Ordering

Gemini 2.5 Flash:

Unsupervised Deep Learning for Coherent Text Structuring and Sentence Representation Learning

LLaMA-3.2 1B Instruct:

Sentence ordering: a new approach to model coherence

LLaMA-3.2 1B Instruct + Our method

Ordering abstracts from scientific articles: an end-to-end unsupervised deep learning approach based on sentence similarity and paraphrase detection tasks.

BART Baseline:

An end-to-end unsupervised deep learning approach for coherent sentences

Our Method (Dual Encoder):

An unsupervised deep learning approach to order sentences

improve the quality of generated scientific titles. The full model performs better than the baseline across all metrics, though the gains are not significant. This suggests that transformer models like BART already do a good job, but guiding them with more structured input can still help.

Looking at the generated examples, our method produces titles that are more relevant to the task and better grounded in the input abstract. In comparison, large language models (LLMs) generate fluent and polished titles, but sometimes add terms that were not mentioned in the input. This makes them less reliable in settings where accuracy matters. Among the LLMs, GPT produces the most specific and faithful output, while Gemini tends to generalize or shift focus slightly. LLaMA produces a concise and mostly relevant title, but its phrasing is more abstract.

Table 4: Pearson correlation between abstract length and evaluation metrics. "Sent." refers to the number of sentences in the abstract, and "Tok." refers to the abstract token count. r is Pearson's correlation coefficient, and p is the corresponding significance value.

Metric	Sent. r	Sent. p	Tok. r	Tok. p
ROUGE-1	-0.048	0.025	-0.051	0.017
ROUGE-2	-0.070	0.001	-0.060	0.005
ROUGE-L	-0.071	0.001	-0.065	0.003
BERTScore F1	-0.037	0.089	-0.034	0.115

The input property analysis shows small but consistent negative correlations between abstract length and ROUGE scores, meaning longer abstracts tend to have less lexical overlap with the reference titles. Correlations with BERTScore are weaker and not significant, indicating that semantic similarity is mostly unaffected. This suggests that longer inputs may add wording variation without reducing the ability to capture the main meaning, supporting the role of sentence selection in removing less relevant content.

6.1 Why BART Outperforms Larger Models and LLMs

Despite the emergence of larger and more sophisticated language models, our results consistently show that the BART-based approach with structured input processing outperforms both general-purpose LLMs (GPT-4.1, Gemini 2.5) and even LLaMA 3.2 enhanced with our proposed method. We hypothesize several key factors underlying this counterintuitive finding.

Task-specific architectural advantage. BART's encoder-decoder architecture is specifically designed for text generation tasks that require distilling and restructuring information. Unlike decoder-only models (GPT, LLaMA), BART can explicitly separate the encoding and decoding phases, allowing for better control over input representation and output generation. This separation enables the model to better focus on the most relevant parts of the input during encoding while maintaining generation fluency during decoding.

Supervised fine-tuning vs. general instruction following. Our BART model is fine-tuned directly on the scientific title generation task with thousands of abstract-title pairs from the target do-

main. In contrast, general-purpose LLMs rely on instruction-following capabilities acquired during pre-training and instruction tuning across diverse tasks. While this makes LLMs more versatile, it may dilute their focus on the specific constraints and conventions of scientific title generation, such as maintaining technical precision while achieving conciseness.

Input control and structured processing. The combination of sentence selection and ordering creates a more focused and coherent input representation that plays to BART's strengths. Our analysis shows that even when this structured input approach is applied to LLaMA (yielding substantial improvements), it still falls short of BART's performance. This suggests that the encoder-decoder architecture is better suited to leverage structured inputs for generation tasks, as it can dedicate the entire encoder to processing the ordered sentences before generating the title.

Precision-recall balance in constrained generation. Scientific title generation requires a delicate balance between covering key concepts (recall) and avoiding extraneous information (precision). Our results show that while LLMs excel at fluency and creativity, they often introduce terms not present in the source abstract or generalize concepts beyond what is warranted. BART with structured input achieves a better precision-recall trade-off, generating titles that are both comprehensive and faithful to the source content.

Training data alignment and domain specificity. Our BART model is trained specifically on scientific abstracts from machine learning journals, allowing it to learn domain-specific patterns in terminology, structure, and style. While LLMs have seen vast amounts of text during pre-training, their knowledge is distributed across many domains and tasks, potentially making them less attuned to the specific requirements of scientific title generation in this domain.

7 Conclusion

In this work, we studied how to improve scientific title generation by adding sentence selection and ordering to a transformer-based model. These steps help the model focus on the most important parts of the abstract and arrange them in a more logical way. Our experiments, both with automatic metrics and sample outputs, showed that this extra structure makes the generated titles more relevant and aligned with the input.

Although the improvements over the baseline were not significant, both selection and ordering gave us more control over the content. This is especially useful in fields where accuracy and clarity matter. The input property analysis showed that abstract length impacts lexical overlap, supporting the role of sentence selection in enhancing title relevance, even if the overall impact is modest. Our results suggest that adding simple structure to input can make models more controllable without needing more compute. In the future, these ideas could be used as planning steps for LLMs, applied to other types of text, or combined with human feedback to make better decisions. This work gives useful insights into how to balance structure and fluency in text generation.

Limitations

Although our method improves title generation by optimizing sentence ordering, it has several constraints. The proxy supervision signal—cosine similarity between abstract sentences and title embeddings—relies on pretrained encoders and may fail to capture subtle semantics, especially for metaphorical or abstract titles, leading to noisy ranking guidance. The cap of 30 sampled permutations limits the search space and risks overlooking better sentence orderings in content-rich abstracts. Our evaluation is confined to machine learning abstracts from Springer journals, whose relatively uniform rhetorical structures may not reflect the variability found in biomedical, humanities, or informal domains, raising concerns about generalization. Furthermore, the absence of human evaluation restricts interpretability, as automated metrics alone cannot fully assess fluency or informativeness. Finally, the need for auxiliary scoring and multiple forward passes adds computational overhead, which may hinder scalability to large datasets or real-time applications unless optimized further.

Acknowledgments

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *Preprint*, arXiv:1611.04953.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Julian Kupiec, Jan O. Pedersen, and Francine R. Chen. 1995. A trainable document summarizer. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2017. Sentence ordering and coherence modeling using recurrent neural networks. *Preprint*, arXiv:1611.02654.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *Preprint*, arXiv:cs/0205028.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Jan Wira Gotama Putra and Masayu Leylia Khodra. 2017. Automatic title generation in scientific articles for authorship assistance: A summarization approach. *Journal of ICT Research and Applications*, 11(3):253–267.

Matsumoto Riku and Kimura Masaomi. 2022. A title generation method with transformer for journal articles. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1115–1120.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Dependency-Aware Word Prediction Integrated with Incremental Parsing

Hiroki Unno[1], Tomohiro Ohno[2], Koichiro Ito[1], Shigeki Matsubara[1,3]
[1]Graduate School of Informatics, Nagoya University
[2]Graduate School of Science and Technology for Future Life, Tokyo Denki University
[3]Information Technology Center, Nagoya University
unno.hiroki.t9@s.mail.nagoya-u.ac.jp
ohno@mail.dendai.ac.jp
{ito.koichiro.z5, matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

Abstract

This paper proposes a method for dependency-aware word prediction performed simultaneously with word input to enhance the performance of real-time natural language processing tasks, such as complementary response generation in dialogue systems and simultaneous machine interpretation. The characteristics of our method are as follows: 1) the targets of the word prediction are not the immediate next words but non-inputted words that have a dependency relation with any of the inputted words, and 2) the word prediction is integrated with incremental dependency parsing. We performed experiments on predicting non-inputted words that have a dependency relation with any of the inputted words, and compared the results with human performance, which confirmed the feasibility of our method. Furthermore, to verify the usefulness of our method for complementary response generation, we evaluated the agreement between actual complementary responses and the words predicted by our method. In addition, we compared the results with those obtained by a large language model (LLM). The results demonstrated that our method can predict main parts of actual complementary responses with higher performance than the LLM, which indicates that our method can provide informative cues with little noise for the complementary response generation.

1 Introduction

Several natural language processing tasks, such as dialogue systems (Nakano et al., 2000; Chiba and Higashinaka, 2025; Liu et al., 2022) and simultaneous machine interpretation (Wang et al., 2024; Ryu et al., 2006; Gu et al., 2017), require real-time responses, and a common requirement of such systems is to execute processing simultaneously with time-continuous input of sentence components. Previous studies have investigated ways to improve the performance of these real-time tasks by predicting non-inputted parts of a sentence through next word prediction (Alinejad et al., 2018; Tsunematsu et al., 2020; Ouyang et al., 2025; Ekstedt and Skantze, 2021).

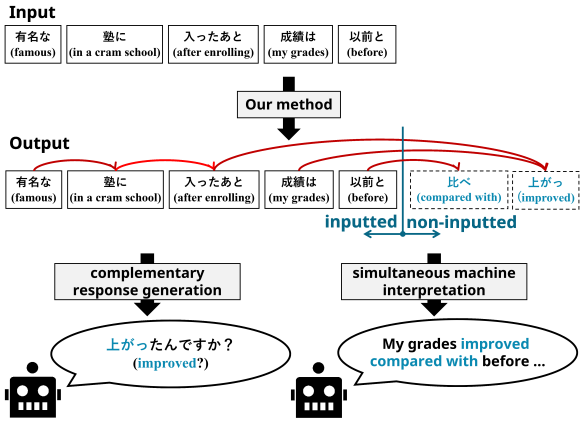


Figure 1: Examples demonstrating the effectiveness of predicting non-inputted words that have dependency relations with any of the inputted words.

In contrast, among these tasks, especially in tasks such as complementary response generation for Japanese and simultaneous machine interpretation between languages with divergent word order (e.g., Japanese-English), situations arise where predicting words that have a dependency relation with any of the inputted words is more beneficial than next word prediction (Oda et al., 2015). For example, in complementary response generation, it is necessary to concisely complement the speaker's utterance; thus, it is important to predict non-inputted words that have a dependency relation with any of the inputted words. In addition, in simultaneous Japanese-English interpretation, predicting predicates that have a dependency relation with any of the inputted words at an early stage enables faster interpretation (Matsubara et al., 2000; Grissom II et al., 2016; Li et al., 2020). As shown in Figure 1, predicting a non-inputted word "上が

っ” (improved), which has dependency relations with the inputted words “入った後” (after enrolling) and “成績は” (my grades), allows systems to generate a complementary response or perform machine interpretation simultaneously. However, to the best of our knowledge, no previous study has focused on non-inputted words that have a dependency relation with any of the inputted words as prediction targets.

Thus, this paper proposes a method for dependency-aware word prediction, i.e., the prediction of non-inputted words that have a dependency relation with any of the inputted words, simultaneously with word input for Japanese. In our method, the word prediction is integrated with incremental dependency parsing, which identifies dependency relations between the inputted and non-inputted words, and both processes are performed jointly in an end-to-end manner. As shown in Figure 1, given a sequence of inputted words, our method identifies dependency relations and the non-inputted words involved in those relations. This approach is based on the following hypothesis: when humans predict non-inputted parts of a sentence, they implicitly anticipate both the syntactic structure and the non-inputted words simultaneously.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 explains our method in detail. Section 4 describes experiments conducted to evaluate the performance of our method, and Section 5 examines how accurately the method can predict the heads of complementary responses. Finally, the paper is concluded in Section 6.

2 Related Work

2.1 Next Word Prediction

Numerous previous studies have investigated ways to improve the performance of real-time natural language processing tasks by predicting the non-inputted parts of a sentence based on next word prediction (Alinejad et al., 2018; Tsunematsu et al., 2020; Ouyang et al., 2025; Ekstedt and Skantze, 2021). For example, focusing on real-time tasks, Tsunematsu et al. (2020) were the first to formalize the task of completing the remaining word sequence given the first 25%, 50%, and 75% of a sentence, and they proposed a method for this completion. In addition, in the simultaneous machine interpretation context, Alinejad et al. (2018) proposed a method to predict the next word of an inputted sequence of words based on an RNN, and Ouyang et al. (2025) proposed a method to predict multiple word sequence candidates coming after the inputted word sequence based on a large language model (LLM). Focusing on dialogue systems, Ekstedt and Skantze (2021) proposed a method to predict a specified number of words following the inputted word sequence by GPT-2.

2.2 Application-Side Requests

As discussed in Section 1, in various tasks, including complementary response generation and simultaneous interpretation between languages with different word orders, it is particularly important to predict non-inputted words that have a dependency relation with any of the inputted words. However, the above-mentioned approaches based on next word prediction merely predict a sequence of non-inputted words by repeatedly predicting the next word, without explicitly identifying which of the predicted words has a dependency relation with any of the inputted words. In addition, the approach based on next word prediction has a known issue, where the prediction accuracy tends to degrade due to error propagation (Zhang et al., 2023; Qian et al., 2025).

To address these issues, this study takes an approach that directly predicts non-inputted words that have a dependency relation with any of the inputted words. To the best of our knowledge, no previous study has focused on non-inputted words that have a dependency relation with any of the inputted words as the prediction targets. However, as a related approach, there exist some studies that have proposed methods to predict the final verb of a sentence (Matsubara et al., 2000; Grissom II et al., 2016; Li et al., 2020). For example, Matsubara et al. (2000) performed early prediction of verbs based on noun phrases and reported the effectiveness of this method in simultaneous interpretation. Additionally, Grissom II et al. (2016) compared human performance with a statistical model in the verb prediction using incomplete Japanese and German sentences, and Li et al. (2020) proposed a method to predict the final verb using a neural model. These methods focus solely on the final verb; however, our approach differs from these existing approaches because it aims to predict all non-inputted words that have a dependency relation with any of the inputted words.

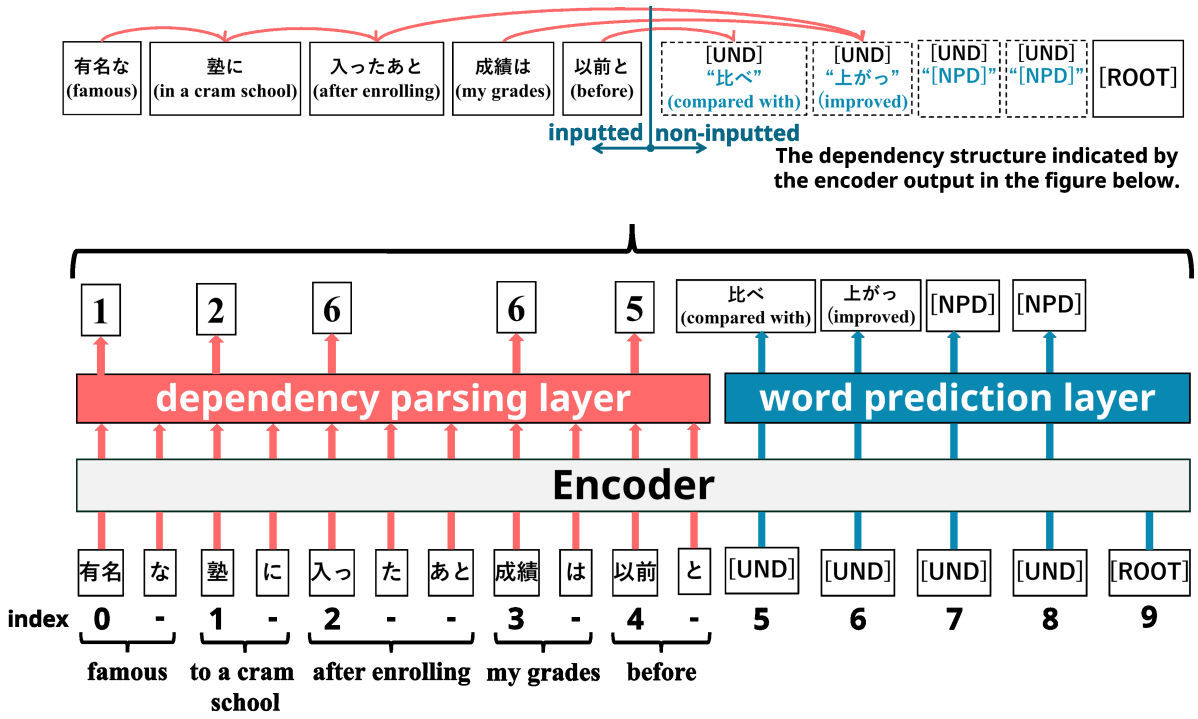


Figure 2: Overview of our method.

3 Word Prediction Integrated with Incremental Dependency Parsing

This section describes our method to predict non-inputted words, i.e., head words[1] of non-inputted *bunsetsus* that have a dependency relation with any of the inputted bunsetsus, for Japanese sentences. A bunsetsu is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu comprises one independent word and zero or more ancillary words. In Japanese, a dependency relation is a modification relation in which a modifier bunsetsu is dependent on a modified bunsetsu, which is expressed as “a modifier bunsetsu → a modified bunsetsu” in this paper.

3.1 Overview of Our Method

Our method end-to-end identifies a structure, as shown in the output of Figure 1 whenever a bunsetsu b_i $(1 \leq t < n)$ is inputted for a sentence comprising n bunsetsus $b_1 \cdots b_n$. In other words, our method identifies not only the dependency structure, which also includes the dependency relations between inputted and non-inputted bunsetsus, but also each head word of the non-inputted modified bunsetsus[2].

An overview of our method is shown in Figure 2. The input to the encoder is the inputted bunsetsu sequence $B_t = b_1 \cdots b_t$. More precisely, the token sequence of B_t is inputted with additional special tokens [UND] and [ROOT], which are described later. We append two independent fully connected layers to the final layer of the encoder. Here, the first fully connected layer parses the modified bunsetsu of each inputted bunsetsu, and the second fully connected layer predicts the head word of each non-inputted modified bunsetsu.

The model was trained using multi-task learning, which jointly optimizes word prediction and dependency parsing. In addition, the total loss is calculated as the weighted sum of the losses for each task with the weighting coefficient λ as follows:

$$Loss = \lambda \times Loss_{\text{word\_prediction}} + (1 - \lambda) \times Loss_{\text{dependency\_parsing}} \quad (1)$$

Here, we adopt the cross-entropy loss as the loss for each task.

[1]Our definition of a head follows that of Uchimoto et al. (1999), except that auxiliary verbs are excluded. In addition, non-independent words are generally not considered head words but are accepted as the heads when no other candidate exists within the bunsetsu.

[2]Non-inputted modified bunsetsus are those that have dependency relations with any of the inputted bunsetsus because we assume that no dependency is directed from right to left, which is almost true in Japanese.

3.2 Word Prediction

Following the previous study (Yoshida and Kawahara, 2022), our method introduces a special token [UND], which denotes an undetermined token as a head word of a non-inputted modified bunsetsu. This special token is added as many as the number of inputted bunsetsus, assuming each inputted bunsetsu is dependent on a different non-inputted bunsetsu.

To predict each [UND] token, the model computes the probability distribution over words in the vocabulary and selects the word with the highest probability. Here, to process a special token [UND] that has no dependency relation with any of the inputted bunsetsus, we introduce a special token [NPD] and train the model to output [NPD] as the token for such [UND] tokens. This design is intended to prevent unnecessary word predictions.

3.3 Incremental Dependency Parsing

Following Shibata et al. (2019), our method formulates dependency parsing as a head-selection problem (Zhang et al., 2017). Specifically, for each head word of an inputted bunsetsu, the model predicts the index of the corresponding modified bunsetsu. We model the dependency relations between bunsetsus through word-level, more precisely, token-level parsing.

We add a [ROOT] token to the end of the input sequence and set [ROOT] as the modified bunsetsu for the head word of the sentence-level root bunsetsu. The index of a [UND] token, which represents a non-inputted bunsetsu, is selected when the model decides that the modified bunsetsu has not yet been inputted.

4 Experiment

To evaluate the feasibility of our method to predict non-inputted modified bunsetsus, we conducted experiments using Japanese lecture speech transcripts and compared the performance of our method with that of a human.

4.1 Dataset

In this study, we used Japanese lecture speech from the Simultaneous Interpretation Database (SIDB) (Matsubara et al., 2002). This dataset is annotated with morphological tags, bunsetsu boundaries, clause boundaries, and dependency relations, all of which have been corrected manually. Note that the morphological tags are annotated based on the IPA dictionary.

We performed 16-fold cross-validation to evaluate the performance of our method. Here, in each fold, one of the 16 lectures served as the test set, and the remaining 15 lectures were used for training. This procedure was repeated once for each lecture. We used two of the 16 lectures as a development set and evaluated the model's performance on the remaining 14 lectures.

We created the training data incrementally. In other words, whenever a new bunsetsu was inputted, we generated a single instance corresponding to the inputted bunsetsu sequence to create the training data. Thus, each sentence produced as many instances as it has bunsetsus. Using all instances to train our model may lead to overfitting. In this experiment, following Yoshida and Kawahara (2022), we prevented overfitting by limiting the incrementally created data used for training to 5% of all generated instances.

4.2 Evaluation Metrics

For performance evaluation, we compared the performance of our model with that of a human (Unno et al., 2024). Here, a single annotator predicted the non-inputted modified bunsetsus and performed incremental dependency parsing on the test set. In other words, the annotator predicted the structure shown in Figure 1.

We evaluated the word prediction performance for non-inputted modified bunsetsus using recall, precision, and F1 score. Here, recall is defined as the percentage of modifier bunsetsus whose modified bunsetsu's head word was predicted correctly out of all modifier bunsetsus whose modified bunsetsus were non-inputted in the gold dependency structure. Precision is defined as the percentage of modifier bunsetsus whose modified bunsetsu's head word was predicted correctly out of all modifier bunsetsus whose modified bunsetsus were non-inputted in the parsed dependency structure. Figure 3 shows an example of the evaluation metrics calculation for word prediction. Even if the top-1 prediction was incorrect, the appearance of the correct word in the higher-ranked list can benefit downstream modules or human users. Thus, we adopted top-k ($k = 1, 2, 3, 4, 5$) and evaluated performance by determining whether the correct words were included in the top-k outputs of our method.

To evaluate the incremental dependency pars-

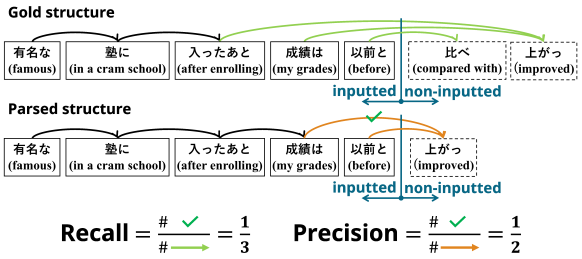


Figure 3: Example of evaluation metrics calculation for word prediction.

Table 1: Results for word prediction of non□inputted modified bunsetsus.

	Recall	Precision	F1
ours (top-1)	7.76	8.04	7.90
ours (top-2)	11.46	11.87	11.66
ours (top-3)	14.42	14.94	14.67
ours (top-4)	16.52	17.11	16.81
ours (top-5)	18.34	19.00	18.67
human (top-1)	11.26	10.91	11.08

ing, we classified each dependency according to whether its modified bunsetsu had already been inputted or was still non-inputted, and we computed the recall, precision, and F1 score values separately for the two categories. Here, recall is defined as the percentage of correctly parsed dependency relations out of all dependency relations in the gold dependency structure. Precision is defined as the percentage of correctly parsed dependency relations out of all dependency relations in the parsed dependency structure.

4.3 Implementation

We implemented our model using PyTorch[3], and we employed the publicly available pre-trained Japanese BERT model released by Tohoku University[4] as the encoder. As a result of hyperparameter tuning on the development set, we set the batch size to 4, the learning rate to 1e-5, the number of training epochs to 10, and the loss-weighting coefficient λ to 0.6. For this evaluation, we varied the random seed, trained five models, and computed the average value of each evaluation metric.

4.4 Experimental Results

4.4.1 Results for Word Prediction

Table 1 shows the evaluation results for word prediction of non-inputted modified bunsetsus. As can be seen, the F1 score obtained by our method increased consistently as the number of candidates increased from top-1 to top-5. In addition, from top-2 to top-5, our method achieved F1 scores that exceeded the human-level performance, which confirms the feasibility of our method.

In contrast, with our method, the F1 score of the top-1 was 3.28 points lower than that of the human, which indicated that our method still falls short of human-level accuracy when limited to a single output. Possible improvements include refining the prediction mechanism and re-ranking candidate outputs. We leave these directions for future work.

4.4.2 Results for Incremental Dependency Parsing

Table 2 shows the evaluation results for incremental dependency parsing. As shown, our method achieved similar performance in cases where the modified bunsetsus were non-inputted (F1 score: 72.30) and where they were already inputted (F1 score: 70.53). These results demonstrate that our method can identify dependency relations involving for non-inputted bunsetsus.

Compared with human performance, the gap in F1 score was 6.19 points when the modified bunsetsu was non-inputted; however, this gap widened to 17.49 points when the modified bunsetsu was already inputted. This result confirms that there is room to improve our method.

4.5 Discussion

Our method formulates word prediction and incremental dependency parsing as two separate tasks, and it accomplishes both tasks in a single framework by solving them simultaneously through multi-task learning. However, the outputs of the two tasks are not always mutually consistent. As shown in Figure 4, two types of inconsistencies can occur. First, our method may output [NPD] as a head word of a non-inputted bunsetsu even though it has determined that the non-inputted bunsetsu has dependency relations with any of the inputted bunsetsus. Second, our method may output

[3] https://pytorch.org/

[4] https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

Table 2: Results for incremental dependency parsing.

	inputted			non-inputted		
	Recall	Precision	F1	Recall	Precision	F1
ours	71.35	69.73	70.53	71.05	73.60	72.30
human	87.77	88.28	88.02	79.74	77.29	78.49

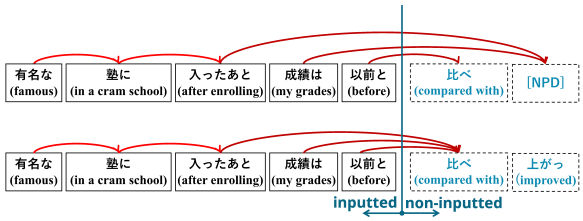


Figure 4: Examples where word prediction and incremental dependency parsing are inconsistent.

a token other than [NPD] as the head word for a non-inputted bunsetsu, that is determined to not have such a relation. In both cases, the outputs of the word prediction and incremental dependency parsing do not align with each other.

To examine the consistency of the outputs of the two tasks in our method, we calculated the following two percentages.

- The percentage of [UND] tokens predicted to be tokens other than [NPD] among those determined parsed to have dependency relations with any of the inputted bunsetsus.
- The percentage of [UND] tokens parsed to have dependency relations with any of the inputted bunsetsus among those predicted to be tokens other than [NPD].

We found that the former and latter percentages were 97.53% and 95.77%, respectively, which indicates that the outputs of the two tasks are mainly consistent. However, a portion remains inconsistent. Addressing such cases will be the focus of future work.

5 Application of Our Method

This section evaluates whether predicting non-inputted modified bunsetsus using the our method is helpful for downstream tasks. Here, we focus on complementary response generation in dialogue systems as a specific downstream task.

A complementary response is a type of responsive utterance that conveys attentive listening (attentive listening response) to a narrative and complements the speaker's narrative. Generating a complementary response requires understanding the content of the speaker's narrative and predicting what the speaker will say; thus, producing such responses is known to be highly effective in conveying an attentive listening attitude. Figure 5 shows an example of a complementary response. In this example, a listener has predicted the non-inputted word "犬派" (a dog person) on the basis that the adversative conjunction "けど" (but) is followed by the clause "猫派だった" (I used to be a cat person), and produced a complementary response. As shown in this example, the element that the listener predicts and provides as a complementary response (i.e., "犬派" (a dog person) in this case) is frequently not the immediate next word. More frequently, it is a modified bunsetsu that has a dependency relations with an inputted bunsetsu, e.g., "猫派だったんですけど" (I used to be a cat person, but) in this figure, and appears later in the speaker's narrative, after at least one intervening word (refer to Appendix A). Thus, in complementary response generation, our method to predict non-inputted modified bunsetsu is expected to be effective.

Accordingly, we assessed the effectiveness of our method for complementary response generation by examining how accurately it predicts the head word of the final bunsetsu that forms the complementary response (referred to as the head word of the complementary response). Specifically, we evaluated the agreement between the head words of the non-inputted modified bunsetsu predicted by our method and the head words of the complementary responses. We demonstrate that our method predicts the head words of the complementary responses with higher accuracy by comparing with an LLM-based method that predicts the non-inputted parts by repeating the next word prediction.

5.1 Responsive Utterance Corpus

In this experiment, we used the Responsive Utterance Corpus (Ito et al., 2022)[5], which contains

[5]We did not use the Responsive Utterance Corpus (Ito et al., 2022) in the experiments discussed Section 4 because

Figure 5: Example of a complementary response.

Prompt
System Prompt: A partial narrative is provided as input. Predict the sentence that follows this narrative. Output only your prediction. Do not repeat the input narrative. Do not add any character decoration or markup. User Prompt: {narrative}

Figure 6: Prompt used in the experiment (English translation).

148,952 attentive listening responses produced by 11 annotators while listening to the narrative audio. JELico (Aramaki, 2016), which is a corpus of narratives produced by elderly people, was employed for the narrative audio. The collected attentive listening responses are classified into 16 categories, including complementary responses, and the corpus contains a total of 614 complementary responses. Note that annotators produced their responses offline while listening to the pre-recorded narrative audio; thus, the content of those responses could not influence the subsequent progression of the narratives. As a result, the speaker never omits or withholds the content of the complementary responses.

5.2 Experimental Settings

In this experiment, we divided the 2,156 narrative sentences in the Responsive Utterance Corpus into 447 sentences for testing, 446 sentences for development, and 1,263 sentences for training. Note that the annotators of the corpus did not always produce a complementary response for each sentence. The 447 test sentences contain 141 complementary responses, and the 446 development sentences contain 123. In this experiment, we treated each point at which a complementary response occurred as a prediction point. Here, we used the narrative from the beginning of the sentence up to the point where the complementary response was produced as the input at each point. Both our method and the LLM-based method predicted the non-inputted parts. For this evaluation, we calculated the agreement between the predicted head words and the head words of the complementary responses for each prediction method.

To train our method, we further fine-tuned the model trained with SIDB in the experiment described in Section 4, using the training part of the Responsive Utterance Corpus. We generated data instances for each of the 1,263 training sentences[6] using the procedure described in Section 4.1, and we trained the model on these data. Based on tuning with the development data, we set the hyperparameters of our method to a batch size of 5, learning rate of 1e-5, three training epochs, and a loss-weight coefficient λ of 0.7.

5.3 LLM-based Method

We employed GPT-4.1[7] as the compared method and instructed it to predict the sentence in the narrative that follows the point at which the complementary response occurs. Table 6 shows the prompt used in this evaluation. Here, we first processed each sentence generated by the LLM using CaboCha (Kudo and Matsumoto, 2003) to apply morphological analysis and bunsetsu segmentation. Then, we determined the head word of each bunsetsu from the resulting parse.

5.4 Evaluation Metrics

We used two evaluation metrics, i.e., recall and precision. Here, recall is defined as the proportion of the 141 complementary responses in the test set whose head words were predicted correctly. A prediction was considered correct if at least one of the head words predicted by each method for the non-inputted bunsetsus matched the head word in the complementary response. Precision is defined as the proportion of the predicted non-inputted bunsetsus whose head word matches the head word of the complementary response. In the evaluation, we used five predictions results obtained by each

it lacks manually annotated gold-standard dependency information.

[6]Here, we used the dependency information parsed by using CaboCha because this data lacks the manual annotation.

[7]https://openai.com/index/gpt-4-1/

Table 3: Results for predicting head word of the complementary responses.

Method	Recall					Precision				
	top-1	top-2	top-3	top-4	top-5	top-1	top-2	top-3	top-4	top-5
ours	**7.56**	**10.08**	**11.76**	**11.76**	**13.45**	**6.87**	**9.16**	**10.69**	**10.69**	**12.21**
LLM	**7.56**	7.56	10.08	**11.76**	12.61	1.43	1.65	2.09	2.31	2.42

method. Specifically, the predictions from the top-1 to top-5 with higher probabilities were selected for our method, and we set the temperature to 0.5 for the LLM and performed five predictions.

5.5 Experimental Results

Table 3 shows the experimental results. As can be seen, the recall values obtained by our method were at least as high as that of the LLM across all top-k levels, which indicates that our method can predict complementary responses with higher coverage than the LLM.

The gap was even greater for precision than for recall; our method achieved considerably higher scores. In addition, we found that increasing k did not narrow the gap between the two methods, and our method consistently retained its advantage across all top-k levels. Since an LLM is based on the iterative next word prediction, it generates not only words that have a dependency relation with any of the inputted bunsetsus but also other many words that do not have a dependency relation with them. Thus, simply using the LLM lowered its precision. In contrast, our method focused on only bunsetsus that have a dependency relation with any of the inputted bunsetsus, effectively discarding irrelevant candidates and identifying the head word of the complementary response more accurately. Thus, compared to simply using an LLM, our method can provide informative cues with little noise for the complementary response generation.

Overall, our method, which directly predicts the head words of non-inputted modified bunsetsus, achieved higher agreement with the head words of the complementary responses than the LLM based on iterative next word prediction. These results highlight the strength of our method in providing relevant, low-noise cues that are useful for complementary response generation.

6 Conclusion

This paper has proposed a method that parses dependency structures incrementally to identify dependency relations between inputted and non-inputted bunsetsus and simultaneously predicts the non-inputted bunsetsus involved in those relations. Our method was evaluated in experiments to predict non-inputted bunsetsus that had a dependency relation with any of the inputted bunsetsus, and compared the results with human performance. The results confirmed both the feasibility of our method and remaining challenges. In addition, applying our method to predicting the head words of complementary responses yielded higher performance than a LLM based on iterative next word prediction. These results further verify the usefulness of our method for the complementary response generation.

Limitations

First, the current model can sometimes produce word predictions that are inconsistent with its dependency parsing outputs because the two tasks are only integrated via a weighted loss, and we impose no constraints to enforce consistency between the word prediction and incremental dependency parsing. Second, the evaluation performed in the current study was limited to Japanese lecture-style narratives (from the SIDB). In other words, we did not evaluate generalization to conversational narratives, other domains, or other languages. Finally, the reported results rely on the top-1 to top-5 candidates. We did not examine selection strategies, e.g., re-ranking, for our method, thereby leaving a gap in terms of practical deployment.

Acknowledgments

This work was supported by JSPS KAKENHI Grand Number JP24K15076.

References

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction Improves Simultaneous Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3022–3027, Brussels, Belgium.

Eiji Aramaki. 2016. Japanese Elder's Language Index Corpus v2. https://figshare.com/articles/dataset/Japanese_Elder_s_Language_Index_Corpus_v2/2082706.

Yuya Chiba and Ryuichiro Higashinaka. 2025. Investigating the Impact of Incremental Processing and Voice Activity Projection on Spoken Dialogue Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3687–3696, Abu Dhabi, UAE.

Erik Ekstedt and Gabriel Skantze. 2021. Projection of Turn Completion in Incremental Spoken Dialogue Systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 431–437, Singapore and Online.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental Prediction of Sentence-final Verbs: Humans versus Machines. In *Proceedings of the 20th Conference on Computational Natural Language Learning*, pages 95–104, Berlin, Germany.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to Translate in Real-time with Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain.

Koichiro Ito, Masaki Murata, Tomohiro Ohno, and Shigeki Matsubara. 2022. Construction of Responsive Utterance Corpus for Attentive Listening Response Production. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7244–7252, Marseille, France.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-Based Text Analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31, Sapporo, Japan.

Wenyan Li, Alvin Grissom II, and Jordan Boyd-Graber. 2020. An Attentive Recurrent Model for Incremental Prediction of Sentence-final Verbs. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2020*, pages 126–136, Online.

Chang Liu, Xu Tan, Chongyang Tao, Zhenxin Fu, Dongyan Zhao, Tie-Yan Liu, and Rui Yan. 2022. ProphetChat: Enhancing Dialogue Generation with Simulation of Future Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 962–973, Dublin, Ireland.

Shigeki Matsubara, Keiichi Iwashima, Nobuo Kawaguchi, Katsuhiko Toyama, and Yasuyoshi Inagaki. 2000. Simultaneous Japanese-English Interpretation based on Early Prediction of English Verbs. In *Proceedings of Symposium on Natural Language Processing*, volume 4, pages 268–273.

Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2002. Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 153–159, Las Palmas, Canary Islands - Spain.

Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa. 2000. WIT: A Toolkit for Building Robust and Real-time Spoken Dialogue Systems. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*, page 150–159, USA.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 198–207, Beijing, China.

Siqi Ouyang, Oleksii Hrinchuk, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, Lei Li, and Boris Ginsburg. 2025. Anticipating Future with Large Language Model for Simultaneous Machine Translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5547–5557, Albuquerque, New Mexico.

Junlang Qian, Zixiao Zhu, Hanzhang Zhou, Zijian Feng, Zepeng Zhai, and Kezhi Mao. 2025. Beyond the Next Token: Towards Prompt-Robust Zero-Shot Classification via Efficient Multi-Token Prediction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7093–7115, Albuquerque, New Mexico.

Koichiro Ryu, Shigeki Matsubara, and Yasuyoshi Inagaki. 2006. Simultaneous English-Japanese Spoken Language Translation Based on Incremental Dependency Parsing and Transfer. In *Proceedings of the International Conference on Computational Linguistics/Annual Meeting of the Association for Computational Linguistics 2006 Main Conference Poster Sessions*, pages 683–690, Sydney, Australia.

Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2019. BERT ni yoru nihongo koubun kaiseki no seido koujou (In Japanese). In *Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing*, pages 205–208. Improved accuracy of Japanese parsing with BERT [Translated from Japanese.].

Kazuki Tsunematsu, Johanes Effendi, Sakriani Sakti, and Satoshi Nakamura. 2020. Neural Speech Com-

pletion. In *Proceedings of Interspeech 2020*, pages 2742–2746.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 1999. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 196–203, Bergen, Norway.

Hiroki Unno, Tomohiro Ohno, Koichiro Ito, and Shigeki Matsubara. 2024. Human Performance in Incremental Dependency Parsing: Dependency Structure Annotations and their Analyses. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 697–706, Tokyo, Japan.

Minghan Wang, Thuy-Trang Vu, Jinming Zhao, Fatemeh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2024. Simultaneous Machine Translation with Large Language Models. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 89–103, Canberra, Australia.

Airi Yoshida and Daisuke Kawahara. 2022. Kouzouteki aimaisei ni motozuku yomizurasa no kenshutsu (In Japanese). In *Proceedings of 28th Annual Meeting of the Association for Natural Language Processing*, pages 425–429. Detecting readability based on structural ambiguity [Translated from Japanese.].

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency Parsing as Head Selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 665–676, Valencia, Spain.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. 2023. PLANNER: Generating Diversified Paragraph via Latent Language Diffusion Model. In *Advances in Neural Information Processing Systems*, volume 36, pages 80178–80190.

A Analysis of Complementary Responses

To evaluate whether prediction of non-inputted modified bunsetsus can function effectively in complementary response generation, we analyzed using the Responsive Utterance Corpus (Ito et al., 2022). The corpus contains 614 complementary responses in total. We randomly sampled 300 complementary responses from the entire set for analysis.

The analysis showed that out of the 300 sample complementary responses, there were 95 responses whose contents appeared in the narrative utterances after the responses were produced. Among these 95 responses, 25 responses ($26.32\% = 25/95$) contained content that appeared as the immediate next words in the subsequent narrative. Furthermore, among the remaining 70 responses, those where the content appeared after at least one intervening word 56 responses ($80.00\% = 56/70$) contained content that appeared as the modified bunsetsu of one of the inputted bunsetsus. These findings confirm that predicting the non-inputted modified bunsetsus is more valuable than simply predicting the immediate next word when generating complementary responses.

How "empirical" is corpus linguistics?: A meta-analysis using formal concept analysis

Kazuho KAMBARA[1,2], **Yuki Sugawara**[3], **Norihisa TAKAHASHI**[2]
[1] National Institute of Information and Communications Technology / Kyoto, Japan
[2] Ritsumeikan University / Shiga, Japan
[3] Osaka University / Osaka, Japan
Correspondence: kambara-k@nict.go.jp

Abstract

Traditionally, corpus linguistics has been positioned as a field that provides methodologies for observing linguistic phenomena and verifying hypotheses derived from linguistic theories (Fillmore, 1992; Gries, 2010b). McEnery and Brezina (2022) summarised the features of corpus linguistic enquiries as 48 principles and treated hypothesis verification as a central feature. However, this characterisation is still open to discussion. Using formal concept analysis (Ganter and Wille, 1999; Ganter et al., 2005) on the abstracts of *International Journal of Corpus Linguistics*, we show that corpus linguistics is a discipline that often aims at hypothesis generation rather than hypothesis testing.

1 Introduction

Since the dawn of corpus linguistics, its effectiveness in observing linguistic phenomena has been recognised. Along with its advantages, its status as a "theory" has also been discussed (McEnery and Hardie, 2012, 147–164). McEnery and Brezina (2022) elegantly summarised 48 principles as foundations of corpus linguistics and treated hypothesis testing as its central notion. However, its validity remains debatable. In this paper, we argue that corpus linguistics is NOT a field based on hypothesis testing but on hypothesis generation. This result does not diminish the scientificity of corpus linguistics in any way.

This paper is structured as follows: Section 2 overviews the proposals in McEnery and Brezina (2022) and introduces our research question. Section 3 explains the methods and procedures employed in our study. Section 4 reports the results of formal concept analysis and shows that hypothesis testing is not necessarily a central notion in corpus linguistics. Section 5 concludes and overviews some possible developments.

2 Towards a philosophy of corpus linguistics

2.1 Corpus linguistics as theories

This section briefly summarises the long-standing debate on the theoreticality of corpus linguistics and introduces the basic tenets of McEnery and Brezina (2022). The theoreticality of corpus linguistics (i.e., how corpus linguistics connects with (or isolates from) linguistic theories) has been debated (cf. Gries, 2010b; McEnery and Hardie, 2012; Tognini-Bonelli, 2001; Teubert, 2007), leading to the discussion of the scientificity of corpus linguistics. McEnery and Brezina (2022) take the debate on theoreticality to the next level by borrowing the notion of falsifiability (Popper, 1972, 1975). Drawing from Popper's ideas, McEnery and Brezina characterised corpus linguistics as a science of hypothesis testing.

The role of corpus linguistics has been among the topics most discussed by many scholars. A well-known illustration of corpus linguists dates back to Fillmore (1992). In contrast to armchair linguists, corpus linguists are portrayed as someone analysing large datasets to draw quantitative generalisations without paying much attention to theoretical details. However unlikely Fillmore's portraits of corpus linguists are (cf. Gries, 2010b), defining characteristics of corpus linguists has been discussed seriously.

One of the central points in this debate is whether corpus linguistics is a theory or not. Some scholars (cf. Gries, 2010b; McEnery and Hardie, 2012) emphasised the role of corpus linguistics as "tools" of linguistic theories, while others (cf. Tognini-Bonelli, 2001; Teubert, 2007) argued for the theoreticality of corpus linguistics. Although it is highly controversial to assume a set of data and methodology can qualify as "theories", corpus linguistic enquiries often appear theory-independent ones since the authentic data (almost always) "be-

tray" our intuitions, which leads to a more accurate understanding of our authentic language use.

The debates of the theoreticality of corpus linguistics ultimately can lead to the philosophy of science. Philosophy of science deals with ontological and epistemological problems: The former corresponds to the question of "what something is" and the latter to "how we know what something is" (cf. Dennett, 1996; Kambara and Yamanaka, 2023). More specifically, philosophers of science attempt to reveal the kinds of targets of corpus linguistic enquiries (i.e., ontological enquiries) and how corpus linguists know those entities (i.e., epistemological enquiries).

McEnery and Brezina (2022) attempted to construct a full-fledged philosophy of science following the ideas of Popper (1972, 1975). Popper's philosophy of science is perhaps best known for positioning falsifiability as the central notion of scientific enquiries. Falsifiability is the ability to falsify (prove to be false) a hypothesis based on a single observation (McEnery and Brezina, 2022, 42). For instance, someone argued, "Martians' telepathic thought is the strong predictors of distinguishing synonymous pairs (e.g., *sofa* vs. *couch*)". Linguists would not even consider the statement's validity since it cannot be confirmed objectively in any way imaginable. The notion of falsifiability plays a crucial role in deciding which hypothesis is worthy of serious contemplation. Unlike the hypothesis regarding the Martians' interruptions to our daily communications, the hypothesis "A register/genre is a strong predictor of distinguishing synonymous pairs (e.g., *sofa* vs. *couch*)" is much more appealing to corpus linguists since its validity can be examined using various corpus linguistic techniques.

The notion of falsifiability assumes that the heart of scientific enquiries is in verifying hypotheses. Borrowing this idea from the philosophy of science can help us understand the corpus linguistic endeavour more precisely. Due to its theoretical importance, various favourable reviews (Curry, 2023; Levin, 2023; Wu, 2023) have been published, suggesting that many corpus linguists agree with the assumption that corpus linguistics is a science of hypothesis verification.

2.2 Scientificity of corpus linguistics

This section challenges the view that corpus linguistics is a science of hypothesis verification. We argue that this perspective, which centralises hypothesis verification, is problematic because it inherits the limitations of conventional philosophy of science (including Popper's), whose understanding of scientificity heavily depends on the traditional framework of physics. We briefly review current advances in the philosophy of science to argue that positioning hypothesis verification as the central notion of corpus linguistics is debatable.

Philosophers of science in the first half of the 20th century divided scientific inquiry into (i) **the context of justification** and (ii) the **context of discovery** (Reichenbach, 1938). The former was not regarded as a process of scientific inquiry because of psychological factors, and the latter was regarded as the central notion in scientific enquiry. This division was influenced by the position of **logical positivism** in the first half of the 20th century. Logical positivism attempted to create the movement for **Unity of Science** aiming to understand science as a whole employing ideas of mathematics, logic, and physics (Cat, 2024). Logical positivism eliminated the context of discovery, which is an "illogical" process, and emphasised the importance of context of justification in the scientific enquiry of the philosophy of science (Schickore, 2022).

In contrast, philosophers of science in the latter half of the 20th century, the target of the philosophy of science expanded beyond the field of science initially attempted by logical positivism. The philosophy of special sciences (i.e., philosophy of individual scientific fields), which closely examines the case of a specific field, became the central endeavour (Fodor, 1974). This movement was embodied in the 1990s by a movement titled **Disunity of Science** by the Stanford School led by John Dupré, Ian Hacking, Peter Galison, Patrick Suppes, and Nancy Cartwright (Galison and Stump, 1996). The Stanford School rejected logical positivism's attempt to describe a unified world of science and helped to redirect analysis toward describing fragments of individual science. Descriptive science, such as biology, has been adopted as an object of analysis in the philosophy of individual science, which does not necessarily emphasise the context of justification, unlike physics.

For instance, evolutionary biologists are likely to describe the different shapes of beaks in various ecological niches (Skipper and Millstein, 2005). Scientists do not regard these works as irrelevant just because they do not (in a strict sense) verify a hypothesis. Moreover, neuroscientists are more interested in the mechanisms of humans' neural

networks (Machamer et al., 2000; Craver, 2007). Again, these enquiries are "scientific" enough even though they do not aim to verify a hypothesis. These cases suggest that it is not realistic to build scientific foundations for a given field just by borrowing ideas that are originated from the framework of physics.

From a linguistic point of view, it is well-known that methodologies of generative grammar are inspired by those of physics (Harris, 2021, 11). As often discussed, the guiding principles of corpus linguistics are far from those of generative grammar since corpus linguists emphasise the importance of solidifying observational foundations (Leech, 1992). For this reason, it is debatable if the core enterprise of scientific enquiries resides in the verification of hypotheses.

In this paper, we aim to observe the qualitative characteristics of corpus linguistics empirically, which can confirm the generalisation made by McEnery and Brezina. If our discussion is on the right track, hypothesis verification should not be observed often in published corpus linguistic research papers. In this sense, our enquiry can be positioned as a meta-analysis of corpus linguistics.

3 Methods

This section explains the methods used in this study. To observe the characteristics of corpus linguistics, we extracted the abstracts of *International Journal of Corpus Linguistics* (IJCL) and manually annotated them to conduct formal concept analysis (FCA). In the following, after explaining the data extraction procedure, we overview the characteristics of FCA in Section 3.1, introduce the annotation strategies in Section 3.2, and describe the procedure of analysis in Section 3.3.

3.1 Formal Concept Analysis (FCA)

Formal Concept Analysis (FCA) is a method developed by Ganter and Wille (1999). It was developed as a lattice theory in applied mathematics. It deals with qualitative data in the form of $i \times j$. FCA provides a powerful way to visualise the structure of a given data, especially their implicational structures. It has been applied in language studies (Priss, 1998, 2005; Hasebe and Kuroda, 2009; Kuroda, 2015).

For instance, let us say we are interested in the semantic relations among person-denoting nouns (i.e., *person*, *adult*, *child*, *man*, *boy*, *woman*, *girl*). These nouns can be analysed in the form of Table 1. Each semantic feature is coded in a binary fashion. Note that attributes with the value "1" indicate that the noun has the given attributes (i.e., *girl* is YOUNG and FEMALE). Therefore, *person*, the most general term, is coded as 0 for all attributes, representing the absence of specific features (YOUNG, OLD, MALE, FEMALE).

Table 1: A formal concept of person-denoting nouns

	YOUNG	OLD	MALE	FEMALE
person	0	0	0	0
adult	0	1	0	0
child	1	0	0	0
man	0	1	1	0
boy	1	0	1	0
woman	0	1	0	1
girl	1	0	0	1

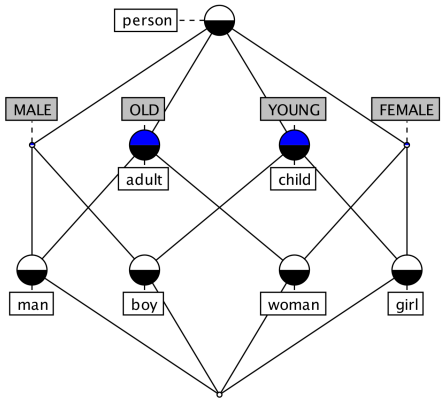


Figure 1: The lattice structure of person-denoting nouns in English

We obtain the lattice in Figure 1 by importing the data to Concept Explorer. This lattice is called concept lattice and represents the class-inclusion relations. White boxes show the names of objects (e.g., *person*, *adult*, *child*, ...), and grey boxes show the names of attributes that classify the given objects. The lattice shows the following implicational relations in (1) from Table 1.

(1) a. The referent of *person* subsumes those of the other six nouns.
b. The nouns *man* and *woman* are special cases of *adult* and of *person* (i.e., The nouns *man* and *woman* are hyponyms of *adult* and *person*).
c. The nouns *boy* and *girl* are special cases of *child* and of *person* (i.e., The nouns boy and girl are hyponyms of *child* and *person*).

d. Unlike the attributes OLD and YOUNG, the attributes MALE and FEMALE do not possess any unique objects, and they are distributed to the special cases of objects possessing OLD or YOUNG.

The structure visualised in Figure 1 is known as a **Hasse diagram**.

Since the conceptual structure of data in Table 1 is relatively straightforward, the interpretations of the concept lattice in Figure 1 are pretty simple. However, identifying and understanding the structure of a large table with a substantial number of rows and columns is labour-intensive. Using FCA can mediate such processes. This paper aims to identify and understand the nature of corpus linguistics using such techniques. The attributes used in this study are explained in Section 3.2.

3.2 Annotation strategies

Because the conceptual structure in a given table is not necessarily evident in advance, the resulting classification may produce a non-optimal lattice. Discarding an object or attribute can help create an optimal concept lattice in FCA. In discarding variables, such actions should be justified by theoretically probable reasons. To achieve this goal, we devised the following attributes to classify the given corpus linguistic research.

(2) Types of goals:
   a. `is_theoretical`: 1 iff the given paper's goal was motivated theoretically, 0 otherwise.
   b. `is_educational`: 1 iff the given paper's goal was motivated educationally, 0 otherwise.
   c. `is_methodological`: 1 iff the given paper's goal was motivated methodologically, 0 otherwise.

(3) Types of "methods":
   a. Types of approach:
      (i) `verifies_hypothesis`: 1 iff the given paper's goal is to verify a hypothesis, 0 otherwise.
      (ii) `presents_hypothesis`: 1 iff the given paper aimed to generate or present a new hypothesis via describing certain phenomena, 0 otherwise.
   b. Kinds of targets:
      (i) `target_is_spoken`: 1 iff the target of analysis was spoken, 0 otherwise.
      (ii) `target_is_written`: 1 iff the target of analysis was written, 0 otherwise.
   c. Characteristics of targets:
      (i) `corpus_is_balanced`: 1 iff the analysed corpus (or its fragments) was balanced in its own right, 0 otherwise.
      (ii) `corpus_is_representative`: 1 iff the analysed corpus (or its fragments) was representative, 0 otherwise.
   d. Originality of targets:
      (i) `introduces_new_dataset`: 1 iff the author(s) of the given paper devised a new dataset, 0 otherwise.
      (ii) `dataset_is_shared`: 1 iff `introduces_new_dataset` is 1, AND the author(s) of the presented paper made the new dataset public, 0 otherwise.

(4) Types of phenomena:
   a. `target_is_micro`: 1 iff the analysed target was a specific expression (e.g., word, phrase, construction), 0 otherwise.
   b. `target_is_cross-linguistics`: 1 iff the analysed target was cross-linguistic, 0 otherwise.
   c. `target_is_variation`: 1 iff the analysed target was a variation of some kind (e.g., genre, gender, place), 0 otherwise.

(5) The year of publication:
   a. `is_in_90s`: 1 iff the given paper was published in the 1990s, 0 otherwise.
   b. `is_in_00s`: 1 iff the given paper was published in the 2000s, 0 otherwise.
   c. `is_in_10s`: 1 iff the given paper was published in the 2010s, 0 otherwise.
   d. `is_in_20s`: 1 iff the given paper was published in the 2020s, 0 otherwise.

Though some attributes may seem redundant, the finalised design is intentional. For instance, as for Types of approach, we intentionally devised both `verifies_hypothesis` and

presents_hypothesis to code the purpose of a given paper that verifies and presents a new hypothesis at the same time. Kinds of targets have four possible combinations, as shown in Table 2. Characteristics of targets are certainly nuanced. The attribute corpus_is_balanced is evaluated on whether the author(s) employed a balanced corpus. Representativeness of corpora is evaluated on how exhaustive the authors collected the given data. For instance, if an author decided to analyse the language use in e-mail exchanges and collect only a handful, corpus_is_balanced is coded as 0.

3.3 Procedures

We first extracted abstracts of all articles published in *International Journal of Corpus Linguistics* (IJCL) from 1996 to 2023. We excluded a total of 53 book reviews and contributions in special issues, as the abstracts of these articles could not be retrieved automatically. As a result, we chose 440 articles for exhaustive qualitative analysis.

For each of the 440 articles, we manually and semi-automatically annotated the features introduced in Section 3.2. After standardising the article names, we used the final file to input Concept Explorer 1.3 for formal concept analysis. However, using the whole data significantly slowed the program's execution, so we randomly sampled 50 articles for formal concept analysis. All 440 annotated abstracts are available on Open Science Framework (OSF).

4 Analysis

4.1 Classification without optimisation

This section reports the results of FCA and their interpretations. As explained, the "uncompromised" lattice can yield non-optimal classification, as shown in Figure 2 with red colliding lines, which is typical when many attributes are included. To arrive at an optimal solution, we can discard either (i) some objects (for possible misclassifications) or (ii) some attributes. For the purpose of achieving a more readable and interpretable lattice, and under the assumption that the object classification is sound, we proceeded by selectively removing attributes introduced in Section 3.2.

4.1.1 Classification based on the goals

Removing all the attributes other than the types of goals produces a simplified lattice as in Figure 3. It shows that (i) the attributes is_theoretical, is_methodological, and is_educational each possess unique generating objects (meaning each goal type is the sole characteristic of at least one research paper), and (ii) no articles were classified as theoretical and educational, while the mixture of methodological motivations with theoretical or educational motivations was observed. As the node in the lattice shows, the most frequent motivations are methodological ($38\% = 19/50$), which aligns with the perspective that corpus linguistic enquiries are often viewed as "tools" for theories.

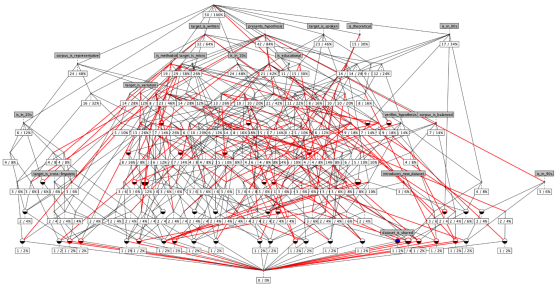


Figure 2: The "uncompromised" lattice of sampled articles using all attributes (with the proportions instead of article ids)

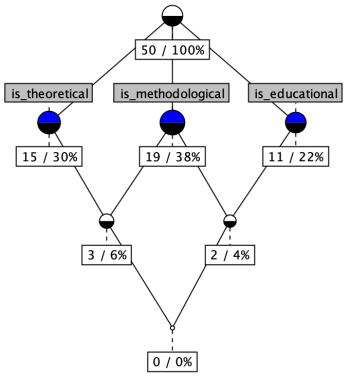


Figure 3: The concept lattice of goals (with the proportions instead of article IDs)

4.1.2 Classification based on the "methods"

Since the number of attributes related to "methods" is quite large, the classification lattice becomes more complex than the other types. Figure 4 is the lattice based on the Types of "methods", in which only two clear implications are read:

(6) a. If a given research paper's data set is shared (in the sense of dataset_is_shared), it introduces a new data set (i.e., corpora) which is a collection of written language (target_is_written), and it

Table 2: Possible combinations of `target_is_spoken` and `target_is_written` and their examples

	`target_is_spoken` = 1	`target_is_spoken` = 0
`target_is_written` = 1	A corpus of written and spoken language	A corpus of written language
`target_is_written` = 0	A corpus of spoken language	A corpus of other language (e.g., sign language), or NA

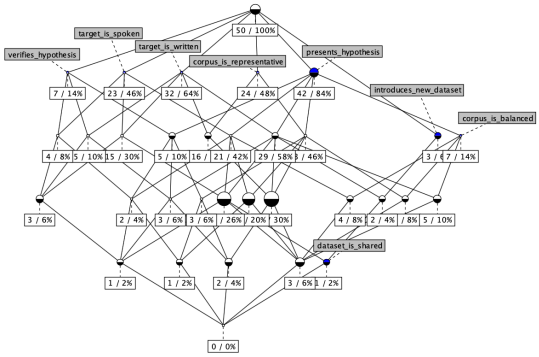


Figure 4: The concept lattice of "methods" (with the proportions instead of article IDs)

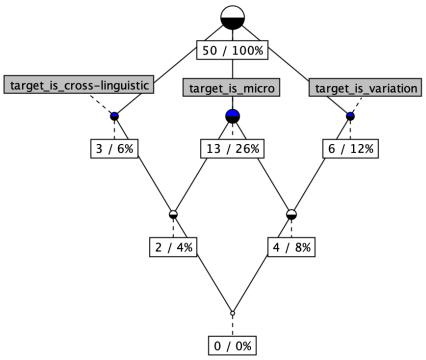


Figure 5: The concept lattice of phenomenon types (with the proportions instead of article IDs)

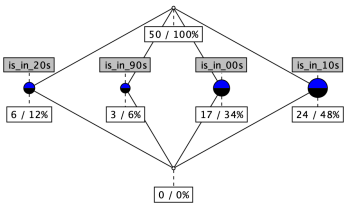


Figure 6: The concept lattice of publication year (with the proportions instead of article IDs)

presents a hypothesis

b. If a given research paper utilises a balanced corpus (i.e., `corpus_is_balanced`), it presents a hypothesis.

Though the first implication is interesting enough, constructing a corpus of written language can be due to the effect of random sampling. However, it is likely for corpus linguists to construct a corpus of written language since it is more accessible than spoken ones. In addition, the latter parts of both implications arrive at presentations of hypotheses, suggesting that verifying hypotheses in corpus linguistic enquiries is not as central as conventionally assumed.

4.1.3 Classification based on types of phenomena

Similar to the classification lattice in Figure 3, the classification lattice based on the types of phenomena is easy to understand. Figure 5 is the lattice using only the phenomenon types for its attributes. As can be read from the lattice, `target_is_cross-linguistic` and `target_is_variation` are mutually exclusive. Since most of the variation research focuses on the distributions of particular expression(s), it is technically challenging to combine cross-linguistic enquiries with variation research. However, if we ignore such variations, we can compare the corresponding expressions cross-linguistically. The lattice shows some of the practical constraints in a corpus linguistic research.

4.1.4 Classification based on the years

Figure 6 shows that the concept lattice based on publication periods is not as "interesting" as the others. This is because all attributes are mutually exclusive, and it only shows the proportions of each time period. It shows that most of the investigated articles were published in the 2010s, reflecting our extraction procedures' limitations. As can be read from the lattice, the raw frequency of `is_in_20s` is larger than that of `is_in_90s`, suggesting that the frequency of publication is accelerated, considering our data only contains articles published in recent years.

4.2 Hypotheses testing in corpus linguistics

As the concept lattice in Section 4.1.2 shows, the role of hypothesis verification is less central than conventionally assumed. This tendency is suggested by the fact that the node labelled `verifies_hypothesis` in Figure 4 does not solely define a unique concept and accounts for only 14 articles (14%). Instead, the lattice strongly shows that presentations of hypotheses are more dominant in the given dataset since the attribute presents_hypotheses possesses unique objects and has 42 unique objects (82%).

The characterisation of corpus linguistics by McEnery and Brezina (2022) relies heavily on the characterisation of scientific enquiries by Popper (1972, 1975). As already pointed out in Section 2.2, the centrality of hypothesis verification in corpus linguistics can be debatable. The simple summarisation of attributes suggests that hypothesis generation (as captured by the `presents_hypothesis` attribute) is more widespread than hypothesis verification.

However, this fact does not diminish the scientificity of corpus linguistics. Not all scientific fields aim to verify hypotheses; some simply emphasise the importance of describing the nature of a target in interest. In some subfields of biology, analysts do not always have an overall understanding of investigated creatures, which usually motivates their empirical enquiries (Kampourakis and Uller, 2020). Like these biologists, corpus linguists often begin without knowing precisely how a given expression behaves in a specific discourse. Instead, they usually devise a systematic procedure for observing various instances of authentic language use. If linguists could know the "inside" of corpus data, encountering unexpected instances becomes impossible. Fillmore (1992, 35) pointed out that corpus data allows linguists to observe data without unnecessary biases.

The descriptive tendency of corpus linguistics invites gap-spotting approaches (Alvesson and Sandberg, 2013), in which researchers identify the "gap" in previous studies to construct their research questions[1]. As discussed, corpus linguists do not know the contents of the investigated data, which easily allows them to create a research question. For instance, if a researcher finds a frequently discussed topic in theoretical or applied linguistics, she can ask how actual speakers realise such a phenomenon. For instance, Gries (2006) discussed the polysemous network of the verb *run*. A lexical item's polysemy network has been discussed in cognitive linguistic literature (Lakoff, 1987). However, how such a network is structured from attested cases had not been clarified. Gries identified senses of the verb *run* and demonstrated how corpus-linguistic techniques contribute to identifying the quantitative and qualitative aspects of polysemous words. This work is a typical case of the gap-spotting approach because linguists did not know the behaviour of the word *run*.

However, as repeatedly emphasised, the descriptive nature of corpus linguistics does not diminish the scientificity of corpus linguistics. Instead, it suggests that corpus linguistic enquiries should be seen as descriptive science like biology (or maybe even ecology). These fields of descriptive science contribute to a realistic understanding of entities in the real world. Since corpus linguists have emphasised the importance of observing attested cases, it is more natural to assume that corpus linguistics is a science of discovery rather than a science of verification.

4.3 Corpus linguistics as a "method" (all over again)

In qualitative analysis using FCA, analysts must carefully select the appropriate attributes that represent some significant characteristics of the target. As discussed, uncompromised classification results in non-optimal resolution (See Figure 2). Based on the discussion that hypothesis verification is not the central notion in corpus linguistics, we selected four attributes: (i) `is_theoretical`, (ii) `is_methodological`, (iii) `presents_hypothesis`, and (iv) `target_is_micro`. As Figure 7 results in an optimal classification, these attributes can represent a typical research project in corpus linguistics.

Figure 7 shows that all research projects are classified into three cross-cutting categories by the above-mentioned attributes `is_theoretical`, `presents_hypothesis`, and `is_methodological`. Among these major attributes, the attribute `presents_hypothesis` is the most widespread (84%), and some projects are purely theoretical (1 unique object) or

[1] Alvesson and Sandberg (2013) criticise the overuse of gap-spotting approaches in social science because such approaches do not invite a novel researches. We refrain from stating that it is preferable or non-preferable for corpus linguists to follow such practices.

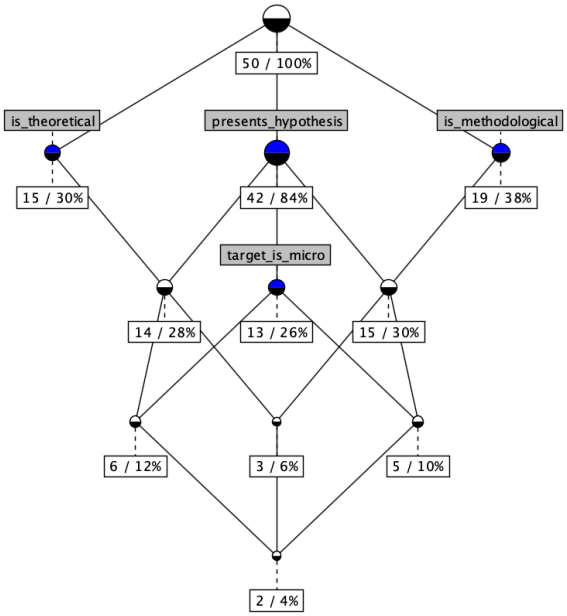


Figure 7: The concept lattice of goals and their approaches

methodological (4 unique objects).

One widely acknowledged advantage of using corpora is that they offer an objective method to observe language use (Fillmore, 1992). The lattice in Figure 7 seems to support the conception of "corpus linguistics as a methodology" because most research projects aim to present a novel hypothesis based on observation of attested data. Some scholars argue that corpus linguistics is a theory of its own, while others treat it as a methodology (McEnery and Brezina, 2022, 147–162). Our analysis empirically suggests that the latter characterisation (corpus linguistics as a methodology/science of discovery) is more fitting than the former.

We repeatedly emphasised that observation of attested data and developing observational tools are central to corpus linguistics, which accords with the statement that hypothesis presentation/generation is more pervasive than hypothesis verification. This tendency was also confirmed in (Gilquin and Gries, 2009). The fact that corpus linguists do not verify hypotheses informed by linguistic theories as often as expected does NOT diminish the scientificity of corpus linguistics. Developments of observational tools cannot be separated from the developments of science. For instance, scientists could not have arrived at a better understanding of creatures' microstructures without the help of electron microscopes. Likewise, without the help of corpora, linguists could have never understood how our intuitions are "betrayed" by attested data.

However, the notion of corpus-as-method does not entirely depend on a specific linguistic theory. In developing concordance tools and constructing a new set of corpora, corpus linguists borrow various notions from neighbouring fields (e.g., Natural Language Processing; NLP). In analysing the given set, analysts exploratorily annotate the given data (cf. Gries, 2010a; Kambara et al., 2023) to see if any combinations of the given variable significantly contribute to the analysis of the given phenomenon. These practices are not deductively derived from the predictions of linguistic theories. Instead, they embrace the "irregularities" found in the data, which accords with the pragmatistic conceptions of science (cf. Quine, 1960, 1961). The pragmatic conceptions of science refer to the gradual progress of scientific knowledge employing all the available resources (Kambara and Yamanaka, 2023; Nefdt, 2023).

In this context, it can be said that the central role of corpus linguistics is to discover **real patterns** hidden in the data (Dennett, 1991). This is the task of systematically capturing the complexity and diversity of actual language use, something that theorists can overlook. If one of the important goals of linguistics is to understand the complex phenomenon of language, then the inventories of corpus linguistic techniques to discover patterns provides an indispensable contribution to the entire field of linguistics. Corpus linguists as discoverers of real patterns can provide a more sophisticated understanding to the debates on corpus-as-method.

5 Conclusion

In this paper, we empirically analysed abstracts published in *International Journal of Corpus Linguistics* (IJCL) and applied Formal Concept Analysis (FCA) to gain a deeper understanding of the field's characteristics. The FCA results strongly suggest that corpus linguistics operates as a science of discovery (akin to descriptive fields like biology) rather than fundamentally as a science of verification. While these findings partially align with the "corpus-as-method" perspective, we argue that this descriptive, discovery-oriented nature necessitates recognising two vital points: (i) corpus linguistics constitutes a distinct scientific field, and (ii) the development of observational tools and procedures is central to its scientific endeavour.

Limitations

Two issues remain unsolved.

First, as previously noted in Section 3.2, relying solely on abstracts for annotation risks distorting the authors' actual intentions and the full scope of their research. Future work should develop a more comprehensive strategy, such as analyzing the full-text content, to address this limitation.

Secondly, our study focused on the qualitative conceptual structure of the papers, rather than conducting a broad quantitative analysis. For our characterisation of corpus linguistics to be fully robust, future studies should aim to quantitatively replicate similar tendencies across a wider population of articles, including those published in related journals.

Acknowledgement

We thank the comments from the three anonymous reviewers. The earlier version of this paper was read at THE JAECS CONFERENCE 2023. Portions of this work were supported by JSPS KAKENHI Grant Number 24K16143. All remaining errors are ours.

References

Mats Alvesson and Jörgen Sandberg. 2013. *Constructing Research Questions: Doing Interesting Research*. Sage Publications, London.

Jordi Cat. 2024. The unity of science. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Carl F. Craver. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.

Niall Curry. 2023. Review of McEnery & Brezina (2022): *Fundamental Principles of Corpus Linguistics*. *International Journal of Corpus Linguistics*, 28(2):278–283.

Daniel C Dennett. 1991. Real patterns. *Journal of Philosophy*, 88(1):27–51.

Daniel C. Dennett. 1996. *Kinds of Minds: Toward an Understanding of Consciousness*. Basic Books, New York.

Charles J Fillmore. 1992. "Corpus linguistics" vs. "computer-aided armchair linguistics". In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, pages 35–66. Mouton de Gruyter, Berlin.

Jerry A. Fodor. 1974. Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2):97–115.

Peter Galison and David J. Stump, editors. 1996. *The Disunity of Science: Boundaries, Contexts, and Power*. Stanford University Press, Stanford.

Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors. 2005. *Formal Concept Analysis: Foundations and Applications*. Springer, Dresden.

Bernhard Ganter and Rudolf Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, Heidelberg.

Gaëtanelle Gilquin and Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Stefan Th. Gries. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Th. Gries and Anatol Stefanowitsch, editors, *Corpora in Cognitive Linguistics: Corpus-Based Approach to Syntax and Lexis*, pages 57–99. Mouton de Gruyter, Berlin.

Stefan Th. Gries. 2010a. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*, 5(3):323–346.

Stefan Th. Gries. 2010b. Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics*, 15(3):327–343.

Randy Allen Harris. 2021. *The Linguistics Wars: Chomsky, Lakoff, and the Battle over Deep Structure*. Oxford University Press, Oxford.

Yoichiro Hasebe and Kow Kuroda. 2009. Extraction of English ditransitive constructions using formal concept analysis. In *23rd Pacific Asia Conference on Language, Information and Computation*, pages 678–685.

Kazuho Kambara, Hajime Nozawa, and Takeshi Takahashi. 2023. Differentiating valence patterns: A quantitative analysis based on formal and semantic attributes. *Constructions*, 15(2).

Kazuho Kambara and Tsukasa Yamanaka. 2023. Philosophy of data science for corpus linguistics: A pragmatistic point of view. *Annals of the Japan Association for Philosophy of Science*, 32:47–73.

Kostas Kampourakis and Tobias Uller, editors. 2020. *Philosophy of Science for Biologists*. Cambridge University Press, Cambridge.

Kow Kuroda. 2015. Formal concept analysis meets grammar typology. In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, pages 329–332.

George Lakoff. 1987. *Woman, Fire and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago.

Geoffrey Leech. 1992. Corpora and theories of linguistic performance. In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, pages 105–122. Mouton de Gruyter, Berlin.

Magnus Levin. 2023. Tony McEnery and Vaclav Brezina. Fundamental principles of corpus linguistics. Cambridge: Cambridge University Press, 2022. 313 pp. ISBN 978-1-1071-1062-5. *ICAME Journal*, 47(1):141–143.

Peter Machamer, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Tony McEnery and Vaclav Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge University Press, Cambridge.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.

Ryan M. Nefdt. 2023. *Language, Science, and Structure: A Journey into the Philosophy of Linguistics*. Oxford University Press, Oxford.

Ryan M. Nefdt. 2024. *The Philosophy of Theoretical Linguistics: A Contemporary Outlook*. Cambridge University Press, Cambridge.

Barbara H. Partee, Alice Ter Meulen, and Robert E. Wall. 1987. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dresden.

Karl R. Popper. 1972. *Objective Knowledge: An Evolutionary Approach*, revised edition. Oxford University Press, Oxford.

Karl R. Popper. 1975. *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5 edition. Routledge, London.

Uta Priss. 1998. The formalization of WordNet by methods of relational concept analysis. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 179–196. MIT Press, Cambridge, Mass.

Uta Priss. 2005. Linguistic applications of formal concept analysis. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis: Foundations and Applications*, pages 149–160. Springer, Dresden.

Willard Van orman Quine. 1960. *Word and Object*. MIT Press, Cambridge, Mass.

Willard Van orman Quine. 1961. *From a Logical Point of View: 9 Logico-Philosophical Essays*. Harper & Row Publishers, New York.

Hans Reichenbach. 1938. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press, Chicago.

Jutta Schickore. 2022. Scientific discovery. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Robert A. Skipper and Roberta L. Millstein. 2005. Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):327–347.

Anna Teubert, Wolfgang Cermakova. 2007. *Corpus Linguistics: A Short Introduction*. Continuum International Publishing Group Ltd., London.

Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. John Benjamins, Amsterdam.

Chenghui Wu. 2023. Book review: *The Fundamental Principles of Corpus Linguistics*. Tony McEnery and Vaclav Brezina. *Digital Scholarship in the Humanities Digital Scholarship in the Humanities*, 38(2):916–917.

A A mathematical characterisation of lattice structures

This appendix provides a formal mathematical characterisation of the lattice structures and concepts (such as the Hasse diagram) used in this paper's Formal Concept Analysis (FCA). Our presentation focuses on illustrating the basic properties of lattice structures and is simplified for explanatory purposes, thus differing slightly from the more formal treatment found in the original characterisation by Ganter and Wille (1999). See Partee et al. (1987) for an introductory explanation of related mathematical concepts.

Definition 1 (partial order and partially ordered set)**.** A binary relation R on a set X satisfies the following conditions 1, 2 and 3.

1. $\forall\, x \in X,\ x\,\mathrm{R}\,x$
2. $\forall\, x, y, z \in X,\ (x\,\mathrm{R}\,y \wedge y\,\mathrm{R}\,z) \Longrightarrow x\,\mathrm{R}\,z$
3. $\forall\, x, y \in X,\ (x\,\mathrm{R}\,y \wedge y\,\mathrm{R}\,x) \Longrightarrow x = y$

Then, the relation R is referred to as **partial order**, the pair (X, R) as a **partially ordered set**.

Definition 2 (lattice)**.** Let $(X, \preceq)$ be a non-empty finite partially ordered set. If $(X, \preceq)$ satisfies following conditions 1 and 2, then $(X, \preceq)$ is called a **lattice**.

1. $\forall x, y \in X$, $\exists z \in X$ such that z satisfies the conditions (a) and (b).
   (a) $z \preceq x \wedge z \preceq y$
   (b) $\forall w \in X$, $w \preceq x \wedge w \preceq y \Longrightarrow w \preceq z$
2. $\forall x, y \in X$, $\exists z \in X$ such that z satisfies the conditions (a) and (b).
   (a) $x \preceq z \wedge y \preceq z$
   (b) $\forall w \in X$, $x \preceq w \wedge y \preceq w \Longrightarrow z \preceq w$

Example 3. Let X be a non-empty finite set and $\preceq$ the binary relation on the powerset $\mathfrak{P}(X)$ defined, for $A, B \in \mathfrak{P}(X)$, by $A \preceq B \overset{\text{def}}{\iff} B \subseteq A$. Then $\preceq$ is a partial order on $\mathfrak{P}(X)$. Thus, for any subset Y of $\mathfrak{P}(X)$, $(Y, \preceq)$ is a partially ordered set. Moreover, for any $A, B \in Y$, we assume that $A \cup B \in Y$ and $A \cap B \in Y$ hold. Then $A \cup B$ and $A \cap B$ satisfy the conditions of the definition 2 with respect to $\preceq$. Therefore $(\mathfrak{P}(X), \preceq)$ is a lattice.

Definition 4 (cover relation)**.** Let $(X, \preceq)$ be a partially ordered set. For $x, y \in X$ with $x \prec y$ (that is, $x \preceq y$ and $x \neq y$), we say that y **covers** x if there is no $z \in X$ such that $x \prec z \prec y$.

Definition 5 (Hasse diagram)**.** Let $(X, \preceq)$ be a non-empty finite partially ordered set. If a graph satisfies the following three conditions, then the graph is called the **Hasse diagram** of $(X, \preceq)$:

(i) The vertex set is X.

(ii) If $x \prec y$ holds, then the vertex y is positioned above the vertex x.

(iii) If y covers x, then give the edge from y to x.

Example 6. Consider the set $X = \{1, 2, 3, 4, 6, 12\}$ of positive divisors of 12. We define a partial order $\preceq$ on X by divisibility: for $x, y \in X$, $x \preceq y \overset{\text{def}}{\iff} x$ divides y.

The cover relations in this partially ordered set $(X, \preceq)$ are

$$1 \prec 2,\ 1 \prec 3,\ 2 \prec 4,\ 2 \prec 6,\ 3 \prec 6,\ 4 \prec 12,\ 6 \prec 12.$$

The corresponding Hasse diagram is given in Figure 8.

It is easy to see that $(X, \preceq)$ forms a lattice, since any two elements of X admit a greatest common divisor and a least common multiple within X.

Example 7 (Application to linguistic categorisation)**.** As shown in Table 1, some person-denoting nouns can be classified using the two properties, (i) age (Old/Young) and gender (Male/Female). The subsets of four attributes $\{$Young, Old, Male, Female$\}$ can represent the semantics of each noun.

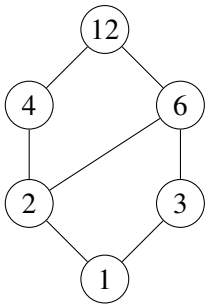


Figure 8: Hasse diagram of the divisors of 12 ordered by divisibility

$$\text{person} = \emptyset,$$

$$\text{adult} = \{\text{Old}\},$$

$$\text{child} = \{\text{Young}\},$$

$$\text{man} = \{\text{Old, Male}\},$$

$$\text{boy} = \{\text{Young, Male}\},$$

$$\text{woman} = \{\text{Old, Female}\},$$

$$\text{girl} = \{\text{Young, Female}\}.$$

Now, for the finite set $X = \{\text{Male, Female, Young, Old}\}$, if we define the subset Y of $\mathfrak{P}(X)$ by

$$Y = \left\{ \begin{matrix} \text{person}, X, \{\text{Male}\}, \{\text{Female}\}, \\ \text{adult, child, man, boy, woman, girl} \end{matrix} \right\},$$

and the binary relation $\preceq$ on Y by the similar manner in Example 3, then $(Y, \preceq)$ is a lattice.

Therefore, the term "person" admits a lattice structure, which can be visualised using the Hasse diagram as shown in Figure 1.

MahaParaphrase: A Marathi Paraphrase Detection Corpus and BERT-based Models

Suramya Jadhav[1,4], **Abhay Shanbhag**[1,4], **Amogh Thakurdesai**[1,4], **Ridhima Sinare**[1,4], **Ananya Joshi**[2,4], and **Raviraj Joshi**[*3,4]
[1]Pune Institute of Computer Technology, Pune
[2]MKSSS' Cummins College of Engineering for Women, Pune
[3]Indian Institute of Technology Madras, Chennai
[4]L3Cube Labs, Pune

Abstract

Paraphrases are a vital tool to assist language understanding tasks such as question answering, style transfer, semantic parsing, and data augmentation tasks. Indic languages are complex in natural language processing (NLP) due to their rich morphological and syntactic variations, diverse scripts, and limited availability of annotated data. In this work, we present the L3Cube-MahaParaphrase Dataset, a high-quality paraphrase corpus for Marathi, a low resource Indic language, consisting of 8,000 sentence pairs, each annotated by human experts as either Paraphrase (P) or Non-paraphrase (NP). We also present the results of standard transformer-based BERT models on these datasets. The dataset and model are publicly shared at https://github.com/l3cube-pune/MarathiNLP.

1 Introduction

Paraphrasing is the task of generating semantically equivalent sentences with different wording or structure. (Bhagat and Hovy, 2013) defines paraphrases as "different surface realizations of the same semantic content, while Barzilay and McKeown (2001) describes paraphrases as "textual expressions that share the same meaning but differ in form". It plays a crucial role in various natural language processing (NLP) applications and is inherently familiar to speakers of all languages (Madnani and Dorr, 2010). Paraphrasing can be a vital tool to assist language understanding tasks such as question answering, style transfer (Krishna et al., 2020), semantic parsing (Cao et al., 2020), and data augmentation tasks (Gao et al., 2020). Interestingly, paraphrase identification can also be effectively implemented for plagiarism detection (Hunt et al., 2019).

The MRPC [1], an English paraphrase corpus, is one such dataset that set a benchmark in creating paraphrase datasets. Since then, a wide variety of techniques, as mentioned in Zhou and Bhat (2021), Madnani and Dorr (2010), Gadag and Sagar (2016) have been developed. However, many of such developments have been around the English language, which for a long time now has been a high-resource language. With plenty of corpora spanning multiple domains like news, sentiment analysis, etc., the preliminary source of sentences becomes rich in diversity, making paraphrase data generation easier. Moreover, models used for detecting semantic and lexical relations between sentence pairs are extensively being developed and put into use, as in Khairova et al. (2022). Developments have also been around languages like Vietnamese (Phan et al., 2022) and Finnish (The Turku Paraphrase (Kanerva et al., 2024)). The ParaCotta corpus (Aji et al., 2022) consists of a paraphrase dataset for around 17 languages and also illustrates how Sentence Transformers like S-BERT can be effectively used for generation as well as evaluation.

While the most important thing to build any model or task-specific dataset (a paraphrase dataset in this case) is having a diverse corpus of scraped and manually verified data, this is severely lacking in the case of Indic languages. This is because of the complexity of Indic languages due to their rich morphological and syntactic variations, diverse scripts, and limited availability of annotated data. However, there has been significant progress in Indic NLP research due to the AI4Bharat-IndicNLP project and IndicNLPSuite (Kakwani et al., 2020), who provide corpora and resources like pretrained models for 10 Indian languages across tasks like sentiment analysis and news headline classification. The Amritha corpus is a paraphrase dataset focused on 4 languages: Hindi, Malayalam, Punjabi, and Tamil (Anand Kumar et al., 2016). The BanglaParaphrase (Akil et al., 2022) focuses on using IndicBART for curating the Bangla paraphrase

*Correspondence: ravirajoshi@gmail.com
[1]Microsoft Research Paraphrase Corpus

corpus.

Very few research groups such as L3Cube [2] are focusing on regional, low resource Indic languages like Marathi. They have also demonstrated that using LLMs for dataset curation (for annotations) has not shown promising results (Jadhav et al., 2024). Moreover, handling paraphrases in Marathi is tricky due to its lexical syntax, complex linguistic features, and the influence of various dialects (Lahoti et al., 2022), (Dani and Sathe, 2024). L3Cube's MahaNLP project (Joshi, 2022b), focused specifically on the Marathi language by developing a Marathi corpus across multiple domains, which helps Marathi NLP.

Contributing to the same project, in this work, we present the L3Cube-MahaParaphrase[3] Dataset, a robust Marathi paraphrase corpus with each sentence pair annotated and manually curated as Paraphrase (P) or Non-paraphrase (NP) with a total of 8K sentence pairs. We further divide the dataset into 5 buckets based on the increasing degree of paraphrase with word overlap and semantic accuracy as factors, giving future research a chance to explore based on varying degrees of paraphrase. The 2-label annotation approach employed is thoroughly described. Furthermore, we also present the results of standard transformer-based BERT models on these datasets. Our key contributions are as follows:

- Created a gold standard 8K Paraphrase corpus for Marathi with labelled sentences pairs as P or NP (4K each for P and NP).
- We divide the MahaParaphrase corpus into multiple buckets based on lexical (word-level) overlap and semantic similarity, thereby capturing varying degrees of paraphrastic and non-paraphrastic relationships between sentence pairs.
- We evaluate existing models like Muril, mBERT, IndicBERT as well as L3Cube's MahaBERT for benchmarking. Additionally, we release MahaParaphrase-BERT[4], a fine-tuned version of MahaBERT trained on the MahaParaphrase corpus.

[2] https://github.com/l3cube-pune/MarathiNLP

[3] https://huggingface.co/datasets/l3cube-pune/MahaParaphrase

[4] https://huggingface.co/l3cube-pune/marathi-paraphrase-detection-bert

2 Literature Review

Research on paraphrase detection and generation has been extensively explored in high-resource languages like English. Different techniques have emerged for paraphrase generation and detection, ranging from using Bi-LSTM with pre-trained GLoVe word vectors (Shahmohammadi et al., 2021) to fine-tuning T5 models (Kubal and Palivela, 2021; Palivela, 2021), and using advanced transformer models like GPT and BERT (Natsir et al., 2023). Combined techniques like variational sampling with hashing sampling, an unsupervised method, have been used for phrase-level and sentence-level paraphrase detection (Hejazizo, 2021). Gangadharan et al. (2020) demonstrated how word vectorization can convert textual data into numerical representations for paraphrase detection and analysis, exploring Count Vectorizer, Hashing Vectorizer, TF-IDF Vectorizer, FastText, ELMo, GloVe, and BERT.

It is important to note that much of this research primarily used English paraphrase corpora for experimentation. Experimentation for other languages is limited due to the lack of quality datasets.

As far as low-resource paraphrase datasets are concerned, Kanerva et al. (2024) introduced a comprehensive dataset, 'Turku Paraphrase,' for the Finnish language. The OpenParcus Dataset consists of paraphrases for six European languages. The ParaCotta Corpus (Aji et al., 2022), which includes around 17 languages, including Hindi, is one of the most diverse datasets spanning a wide variety of languages.

Talking about Indic languages, generation of paraphrases becomes difficult because of rich morphological and syntactic variations and diverse scripts. Moreover, all Indic languages fall under the low-resource category due to the lack of annotated data. The Bangla Paraphrase (Akil et al., 2022) uses IndicBART to synthetically generate paraphrases. In Anand Kumar et al. (2016), a significant milestone was achieved with the release of the Amritha paraphrase corpus for four Indic languages: Hindi, Malayalam, Punjabi, and Tamil, as part of the DPIL@FIRE2016 Shared Task, enabling participants to experiment further.

Another notable effort is the IndicParaphrase Dataset by AI4Bharat(Kumar et al., 2022), which includes 11 Indic languages: Assamese (as), Bengali (bn), Gujarati (gu), Kannada (kn), Hindi (hi), Malayalam (ml), Marathi (mr), Oriya (or), Pun-

jabi (pa), Tamil (ta), and Telugu (te). This dataset provides input and target sentences, as well as a reference list of five sentences with different levels of lexical correlation.

While the Marathi subset in IndicParaphrase is huge for Marathi, it is important to note that it consists only of paraphrased sentence pairs. The same applies to many of the above-mentioned corpora, like BanglaParaphrase, Turku Paraphrase, and OpusParcus. However, the Amritha corpus for the DPIL@FIRE2016 shared task includes labeled sentence pairs as P (Paraphrase) and NP (Non-paraphrase) but does not include Marathi. To date, there is no Marathi paraphrase dataset that consists of both P and NP sentence pairs with 5 varying paraphrastic levels.

3 Dataset

This section provides information on how the dataset was collected. We created the paraphrase dataset in three phases: gathering sentences from MahaCorpus, categorizing them into P and NP using both cosine similarity and back-translation approaches, and then these sentences were manually verified for errors by four native Marathi human annotators. Finally, the sentences were divided into five equally distributed buckets based on word overlaps. Each of these steps is discussed in the following subsections and represented in Figure 1.

3.1 Collection

The required Marathi sentences were taken from the MahaCorpus dataset by L3 Cube, which spans a wide range of topics, including news, sentiment, and hate speech. These sentences were collected from various news sources from the Maharashtra region.

We randomly selected 1 million sentences from this corpus as our primary dataset. In this section, we elaborate on the annotation process for labeling sentences as Paraphrase (P) and Non-Paraphrase(NP).

3.2 Annotations

The collected sentence pairs were annotated using 2 approaches so as to get a mixture of both real and synthetic data. We now explain the two approaches used to categorize sentences as P or NP.

3.2.1 Approach 1: Cosine Similarity

In this approach, we calculated the cosine similarity for every pair of sentences from the 1 million collected sentences. Since contextualized token embeddings have been shown to be effective for paraphrase detection (Devlin et al., 2019), we use BERTScore (Zhang et al., 2020) to ensure semantic similarity between the source and candidates. To do this, we used the sentence transformer MahaSBERT (Joshi et al., 2023) to generate sentence embeddings. Then, we calculated the cosine similarity scores between the embeddings of each pair of sentences.

Based on the scores, we categorized the sentences as follows:

- If the cosine similarity (C.S) score was less than 0.8, the sentences were labeled as NP.
- If the cosine similarity score was between 0.8 and 0.99, the sentences were labeled as P.

3.2.2 Approach 2: Back Translation

In this approach, we used the back-translation technique to generate paraphrase sentences. The process involves:

- Translating a Marathi sentence (S1) into English (S2) using Google Translator.
- Translating it back from English (S2) to Marathi (S3) using Google Translator.

This gives us a pair of sentences: the original sentence (S1) and the back-translated sentence (S3), which we consider as paraphrases.

Filter: To ensure that S3 is not identical to S1, we applied a filter after translation. We used a sentence transformer to calculate the cosine similarity between the sentences and enforced the following rules:

- If the cosine similarity (C.S) score was less than 0.8, we discarded the pair (indicating the meaning might have changed).
- If the cosine similarity score was greater than 0.99, we discarded the pair (indicating the sentences were too similar, likely identical, and not valid paraphrases).

For Marathi-to-English translations (i.e. S1 to S2), we used IndicSBERT (Deode et al., 2023), and for Marathi-to-Marathi comparisons, we used MahaSBERT (Joshi et al., 2023) to compute the cosine similarity score between the two sentences (i.e S1 and S3) in the filter.

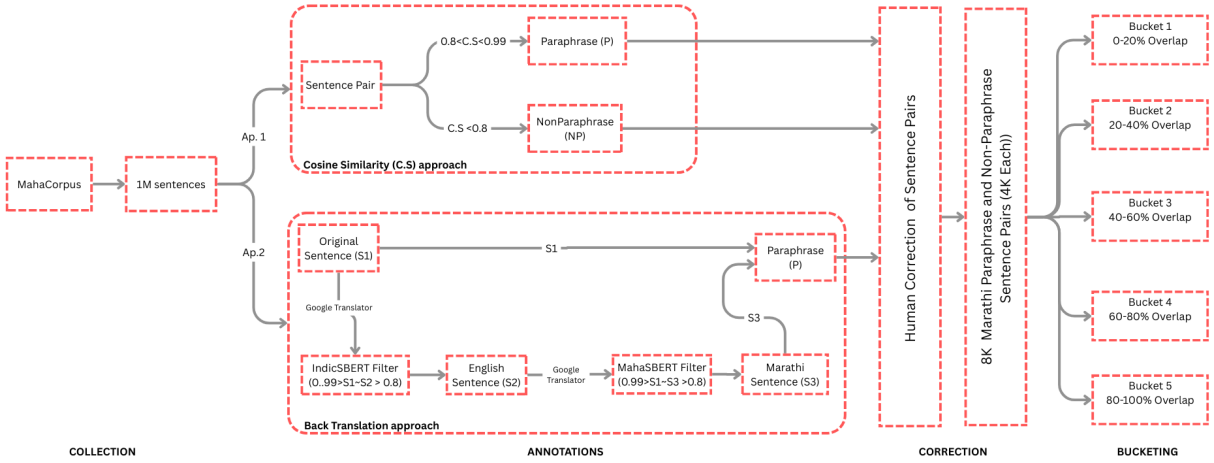


Figure 1: MahaParaphrase Dataset Curation Workflow.

Combining Approaches: Approach 1 provides real data, as both sentences are directly taken from the MahaCorpus (Joshi, 2022a). On the other hand, Approach 2 generates new sentences, which are synthetic. To maintain balance, we used an equal number of sentences from both approaches.

3.3 Human Correction

To ensure that all sentences were correctly classified, the entire dataset was manually verified by native Marathi speakers proficient in reading and writing Marathi.

Any errors were corrected by manually modifying the sentences to ensure accuracy and consistency.

3.4 Bucketing

We further categorized the P and NP data into five buckets for each category, based on word overlap.

- Bucket B5: 80-100% word overlap
- Bucket B4: 60-80% word overlap
- Bucket B3: 40-60% word overlap
- Bucket B2: 20-40% word overlap
- Bucket B1: 0-20% word overlap

Word overlap is calculated as:

$$\text{Word Overlap}(A, B) = \frac{|W(A) \cap W(B)|}{|W(A) \cup W(B)|} \quad (1)$$

where $W(A)$ and $W(B)$ denote the sets of words in sentences A and B respectively.

For example, a pair of sentences in B5 of the NP dataset will be semantically different (and hence categorized as NP), even though they have around 80% or more word overlap. This is significant because it highlights that even with high word overlap, sentences can have different meanings, emphasizing the importance of word order when considering paraphrase pairs.

Now consider another example: a pair of sentences in B1 of the P dataset. These sentences have low word overlap but are still considered paraphrases. This shows that sentences with different words (such as synonyms) can also form valid paraphrase pairs.

This categorization into buckets makes the dataset robust and versatile for evaluation across different scenarios, such as high-overlap non-paraphrase sentence pairs or low-overlap paraphrase sentence pairs. Refer table 1 for bucket-wise examples from the dataset.

4 Dataset Statistics

The dataset contains 4000 rows of sentence pairs labeled as paraphrase (P), and 4000 rows of sentence pairs labeled as non-paraphrase (NP). Each row in the dataset contains a sentence along with a paraphrase or non-paraphrase label.

The average word count for sentences in the dataset, as well as the difference between the averages, is given in Table 2.

Figure 2 shows the distribution of sentence lengths for both Paraphrase (P) and Non-paraphrase (NP) classes in our dataset. Both distributions follow a similar pattern, with the highest frequency occurring for sentence lengths between

Bucket	P	NP
0–20	तर बहिणीला वस्त्राभुषणाची भेट देऊन तिला सुख, शांतता, सौभाग्य, समृद्धी प्राप्त व्हावी, यासाठी भाऊराया प्रार्थना करतो (So that the sister receives happiness, peace, good fortune, and prosperity, the brother prays while giving her clothes and ornaments.) रक्षाबंधन आटोपल्यावर भावाने वस्त्र, आभुषणे किंवा इच्छित भेटवस्तू बहिणीला देऊन तिच्या सुखी जीवनासाठी प्रार्थना करावी (After Raksha Bandhan, the brother should give clothes, ornaments, or a desired gift to the sister and pray for her happy life.) **(19.35%)**	विरोधी भाजपताराणी आघाडीकडून महापौरपदासाठी अर्ज दाखल न झाल्याने महापौरपद निवडीची केवळ औपचारिकता राहिली आहे (As no nomination was filed for mayoral post from the opposition BJP alliance, the selection has become a mere formality.) काम पूर्ण न झाल्याने मंत्रालय पातळीवर नाराजी असल्याने फिआफच्या परिषदेसाठी कुणाला पाठवले गेले नसल्याचे बोलले जात आहे (Due to incomplete work and dissatisfaction at the ministry level, no one was sent for the FIAF conference, as per reports.) **(19.35%)**
20–40	या प्रकरणाची केस डायरी, घटना स्थळाचे फोटो, ऑटोप्सी रिपोर्ट, मुंबई पोलिसांचा फॉरेन्सिक रिपोर्ट आणि नोंदवलेल्या साक्षीदारांचे जबाब लवकरात लवकर पाठवण्याची विनंती सीबीआयकडून करण्यात येणार आहे (CBI will request to send the case diary, crime scene photos, autopsy report, forensic report from Mumbai Police, and recorded witness statements as soon as possible.) याशिवाय सीबीआय मुंबई पोलिसांकडून केसची डायरी, ऑटोप्सी रिपोर्ट, क्राइम सीनचे फोटो, मुंबई पोलिसांचे फॉरेन्सिक रिपोर्ट, पोस्टमॉर्टेम रिपोर्ट, साक्षीदारांचे नोंदवलेले जबाब यांच्या प्रती घेणार आहे (Additionally, CBI will collect copies of the case diary, autopsy report, crime scene photos, forensic report from Mumbai Police, postmortem report, and witness statements.) **(39.13%)**	गेल्या सलग दोन वर्षांत ठाण्यात लायसन्ससाठी अर्ज केलेल्या उमेदवारांपैकी १७ टक्के अर्जदारांना वाहतूक चिन्हांबाबत माहितीच नसल्याचे स्पष्ट झाले आहे (In the past two years, 17% of license applicants in Thane were found to lack knowledge of traffic signs.) या पार्श्वभूमीवर, गेल्या २ वर्षभरातील अर्जदारांच्या चाचणीच्या निकालाची माहिती मिळवली असता, तब्बल १७ टक्के उमेदवारांना वाहतूक चिन्हांची माहितीच नसल्याचे उघड झाले आहे (Data from applicants' tests over the past 2 years revealed that 17% of candidates had no knowledge of traffic signs.) **(39.02%)**
40–60	म्हणून लवकरच अर्थव्यवस्था पूर्वपदावर येतील आणि इंधन मागणी वाढेल, अशी अपेक्षा सौदी अरामकोचे मुख्य कार्यकारी अधिकारी अमीन नासिर यांनी सांगितले (The economy is expected to recover soon and fuel demand will rise, said Amin Nasser, CEO of Saudi Aramco.) म्हणून अर्थव्यवस्था लवकरच पूर्वपदावर येईल आणि इंधनाची मागणी वाढेल, असे सौदी आरामकोचे सीईओ अमीन नासेर यांनी सांगितले (The economy will soon return to normal, and fuel demand will rise, said Saudi Aramco CEO Amin Nasser.) **(59.46%)**	या उपोषणात कार्यकर्त्यांनी सहभागी व्हावे, असे आवाहन भाजपचे शहर जिल्हाध्यक्ष भगवान घडमोडे व ग्रामीण जिल्हाध्यक्ष एकनाथराव जाधव यांनी केले आहे (Workers should participate in this protest, appealed BJP's city district president Bhagwan Ghadmode and rural president Eknathrao Jadhav.) भाजपचे पदाधिकारी, कार्यकर्ते व नागरिकांनी या आंदोलनात सहभागी व्हावे, असे आवाहन भगवान घडमोडे, विजय साळवे यांनी केले आहे (BJP officials, workers, and citizens should join the protest, appealed Bhagwan Ghadmode and Vijay Salve.) **(59.46%)**
60–80	तसेच, जीवे मारण्याची धमकी देऊन त्यांच्याकडील ७०० रुपयांची रोख रक्कम आणि दुचाकी असा ४० हजार ७०० रुपयांचा ऐवज चोरून नेला (Threatening to kill, he stole □700 in cash and a bike totaling □40,700.) तसेच, जीवे मारण्याची धमकी देत त्यांच्याकडील ७०० रुपयांची रोख रक्कम आणि दुचाकी असा ४० हजार ७०० रुपयांचा ऐवजाची चोरी केली (He threatened and stole □700 in cash and a bike worth □40,700.) **(78.95%)**	पिंपरीतील शंभर टक्के निकालस्टेर्लिंग हायस्कूल, प्रियदर्शनी, स्वामी समर्थ (भोसरी), कमलनयन बजाज, गीता-माता, (चिंचवड), एसएनबीपी, अल्फोन्सा, निर्मल बेथनी, रॉजर्स इंग्लिश, डी (100% results in schools like Sterling High School, Priyadarshani, etc. in Pimpri) शंभर टक्के निकाल लागलेल्या शाळा स्टेर्लिंग हायस्कूल, प्रियदर्शनी, स्वामी समर्थ (भोसरी), कमलनयन बजाज, गीता माता, (चिंचवड), एसएनबीपी, अल्फोन्सा, निर्मल बेथनी, रॉजर्स इंग्लिश, डी (Schools with 100% result: Sterling High School, Priyadarshani, etc.) **(79.07%)**
80–100	भाद्रपद महिन्यातील शुद्ध अष्टमीला राधादेवीचा जन्म झाला (Radhadevi was born on Shuddha Ashtami in the month of Bhadrapada.) राधादेवीचा जन्म भाद्रपद महिन्यातील शुद्ध अष्टमीला झाला (Radhadevi's birth occurred on Shuddha Ashtami in the month of Bhadrapada.) **(100.00%)**	मी खरी भविष्यवाणी केली नव्हती काय? (Didn't I make the correct prediction?) मी खरी भविष्यवाणी केली नव्हती काय? (Didn't I make the correct prediction?) **(100.00%)**

Table 1: P and NP Sentence Pairs with Overlap Percentages by Bucket. Every cell consists of S1 and S2 along with their english translations followed by the word overlap percentages (in **bold**). The examples choosen are pairs with max word overlap in that particular bucket.

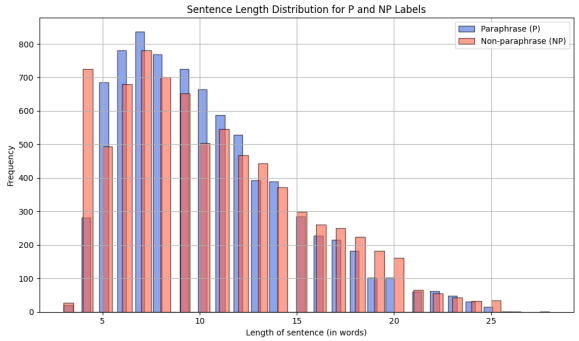


Figure 2: Sentence length distribution for Paraphrase (P) and Non-paraphrase (NP) classes. The x-axis shows sentence length in words, and the y-axis indicates the frequency of those lengths.

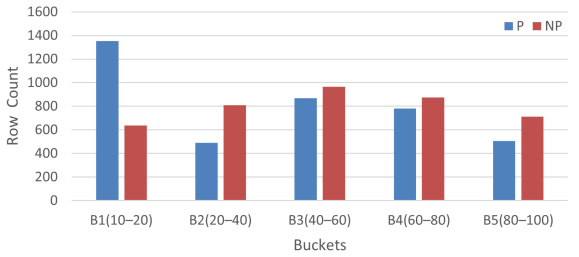


Figure 3: Bucket Wise Distribution. The values in brackets are the word overlap percentages for each bucket.

5 and 15 words.

Figure 3 shows the bucket wise row count distribution for both Paraphrase (P) and Non-paraphrase (NP) classes. While the total count of P and NP are same (i.e 4000 each), their distribution across buckets is varied.

Dataset	Sentence 1 Avg.	Sentence 2 Avg.	Avg. Diff.
Paraphrase	10.43	9.99	1.45
Non-Paraphrase	9.52	11.22	3.36

Table 2: Average word counts and average difference in sentence lengths per record for Paraphrase and Non-Paraphrase datasets.

5 Baseline Models

5.1 Muril

MuRIL is a language model built especially for Indian languages and trained entirely on a large volume of Indian language text (Khanuja et al., 2021). The dataset contains both translated and transliterated document pairings in order to introduce supervised cross-lingual learning during training.

5.2 MBERT

A BERT-based model called Multilingual BERT (mBERT) was trained using text in 104 distinct languages (Devlin et al., 2019). It is trained with masked language modeling (MLM) and next sentence prediction (NSP) objectives, and it supports a variety of downstream applications, including sentiment analysis.

5.3 IndicBERT

Based on the ALBERT architecture (Lan et al., 2020), IndicBERT is a language model that was trained on a huge corpus of 12 major Indian languages, including Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu, and Assamese. It employs a combined training technique and makes use of data from the IndicCorp dataset (Kunchukuttan et al., 2020) in order to better accommodate low-resource languages. There are two versions of the model: IndicBERT (MLM+TLM) and IndicBERT (MLM alone). IndicBERT-MLM is trained by masking random tokens in monolingual text and predicting them using context. IndicBERT-TLM uses parallel sentences in different languages, masking tokens and predicting them using both languages.

5.4 MahaBERT

A multilingual BERT model called MahaBERT was refined using the L3Cube-MahaCorpus and additional publically accessible Marathi monolingual datasets (Joshi, 2022a).

6 Result

The baseline models described above were fine-tuned and evaluated on our dataset, and the results are presented in Table 3. Among the models that were evaluated, MahaBERT was the most accurate model, followed by IndicBERT (MLM + TLM), Muril, IndicBERT (MLM only) and MBERT.

Model	Score
MahaBERT	88.7
IndicBERT (MLM+TLM)	87.1
Muril	86.9
IndicBERT (MLM only)	85.9
MBERT	84.59

Table 3: Model Performance Comparison. MLM stands for Masked Language Modeling and TLM stands for Translation Language Modeling.

7 Conclusion

In this paper, we present the MahaParaphrase dataset, comprising of 8,000 labeled pairs of both paraphrase and non-paraphrase sentences. The entire dataset was manually verified by four native Marathi speakers and is further divided into five buckets based on word overlap. These bucketed subsets capture varying degrees of paraphrasing intensity, which can support more nuanced research in this domain.

Furthermore, we evaluate the MahaParaphrase dataset using five baseline models, with MahaBERT achieving the highest performance—an F1 score of 88.7%.

By providing this low-resource paraphrase dataset, we aim to equip researchers and practitioners with a valuable resource to advance further research in Marathi NLP.

8 Limitations

Compared to paraphrase dataset for high-resource languages, this dataset is relatively small (8K pairs). Moreover, the presence of code-mixed sentences introduces minor noise especially when using BERT models trained specifically using Marathi. Additionally, the dataset evaluation was limited to BERT-based models; incorporating LLMs could offer a more comprehensive assessment.

Acknowledgement

This work was carried out under the mentorship of L3Cube, Pune. We would like to express our gratitude towards our mentor, for his continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022b).

References

Alham Fikri Aji, Tirana Noor Fatyanosa, Radityo Eko Prasojo, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2022. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *Preprint*, arXiv:2205.04651.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: a high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

M Anand Kumar, Shivkaran Singh, B Kavirajan, and KP Soman. 2016. Shared task on detecting paraphrases in indian languages (dpil): An overview. In *Forum for Information Retrieval Evaluation*, pages 128–140. Springer.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational linguistics*, 39(3):463–472.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6806–6817, Online. Association for Computational Linguistics.

Asang Dani and Shailesh R Sathe. 2024. A review of the marathi natural language processing. *arXiv preprint arXiv:2412.15471*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ashwini Gadag and BM Sagar. 2016. A review on different methods of paraphrasing. In *2016 International conference on electrical, electronics, communication, computer and optimization techniques (ICEECCOT)*, pages 188–191. IEEE.

Veena Gangadharan, Deepa Gupta, L Amritha, and TA Athira. 2020. Paraphrase detection using deep neural network based word embedding techniques. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 517–521. IEEE.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Ali Hejazizo. 2021. Combining variational sampling and metropolis–hastings sampling for paraphrase generation.

Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay

Dahal, and 1 others. 2019. Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104. IEEE.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On limitations of llm as annotator for low resource languages. *arXiv preprint arXiv:2411.17637*.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2023. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. In *Science and Information Conference*, pages 1184–1199. Springer.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. *arXiv preprint arXiv:2202.01159*.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and 1 others. 2024. Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for finnish. *Natural Language Engineering*, 30(2):319–353.

Nina Khairova, Anastasiia Shapovalova, Orken Mamyrbayev, Nataliia Sharonova, and Kuralay Mukhsina. 2022. Using bert model to identify sentences paraphrase in the news corpus. In *COLINS*, pages 38–48.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

D Kubal and H Palivela. 2021. Unified model for paraphrase generation and paraphrase identification.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicnlg suite: Multilingual datasets for diverse nlg tasks in indic languages.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *Preprint*, arXiv:2005.00085.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Affan Hilmy Natsir, Indriana Hidayah, and Teguh Bharata Adji. 2023. Deep learning in paraphrase generation: A systematic literature review. In *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 118–123. IEEE.

Hemant Palivela. 2021. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(2):100025.

Quoc Long Phan, Tran Huu Phuoc Doan, Ngoc Hieu Le, Ngoc Bao Duy Tran, and Tuong Nguyen Huynh. 2022. Vietnamese sentence paraphrase identification using sentence-bert and phobert. In *International Conference on Intelligence of Things*, pages 416–423. Springer.

Hassan Shahmohammadi, MirHossein Dezfoulian, and Muharram Mansoorizadeh. 2021. Paraphrase detection using lstm networks and handcrafted features. *Multimedia Tools and Applications*, 80(4):6479–6492.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online

and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Computational Linguistic Approach to Empathy and its Language Communication Pattern

Xinyi Wang[1], Mingyu Wan[1], Chu-Ren Huang[1]
[1] Department of Language Science and Technology, The Hong Kong Polytechnic University
xinyi1109.wang@polyu.edu.hk; mingyu.wan@polyu.edu.hk; churen.huang@polyu.edu.hk
Correspondence: churen.huang@polyu.edu.hk

Abstract

Research on empathy computation within the context of disaster narratives is the primary focus of this study. We aim to model the emotional dimensions of empathy while systematically exploring its cognitive and social aspects through linguistic features. By analyzing empathy in social media data, we provide a theory-grounded account of its communication patterns, drawing on principles of embodiment and trust concerning social behaviors. Findings reveal that individuals with higher levels of empathy are more likely to use concrete language to convey intentions that foster connection and broader social engagement. Our work demonstrates how language, emotion, and social cognition interact, offering a computational investigation of empathy that may contribute to new perspectives in language technology and communication.

1 Introduction

Empathy, as a significant aspect of human communication, is a complex socio-emotional behavior concerning how we understand others. It interacts with advanced cognitive processes (Coplan and Goldie, 2011; Omdahl, 1995), yet can be effectively conveyed through linguistic devices.

However, over two decades after Picard (1997)'s seminal work on Affective Computing, conceptualizing and measuring empathy remains challenging. On one hand, previous studies have focused mostly on empathy's emotional dimension, while broader facets (e.g., cognitive reasoning and interpersonal tendencies) leave scope for further inquiry. On the other hand, while benefiting from the progress of large language models (LLMs), empathetic modeling remains confined to tasks such as detection and response generation in two-person counseling settings, lacking theoretical interpretability of its linguistic communication patterns at scale.

Motivated by these gaps, our goal is to broaden the boundary of understanding empathy, extract its linguistic representations, and thereby enhance empathic modeling accuracy. Specifically, we developed a replicable language-anchored method using natural language processing technology to compare linguistic differences between empathic and non-empathic expressions in digital contexts. Our key contributions are as follows:

- We propose a theory-grounded body-cognition framework to guide linguistic feature design for empathy modeling.
- We improve empathy classification performance by integrating cognitive, perceptual,and syntactic features.
- We draw on principles of embodiment and trust to account for the communication patterns of empathy.

2 Theoretical Framework

Integrating emotional, cognitive, and social perspectives, we introduced two key theories: the Stereotype Content Model (SCM) (Fiske et al., 2002) and Embodied Appraisal Theory (EAT) (Prinz, 2004).

SCM identifies *Warmth* (refined into *Trust* and *Sociability*) (Abele and Wojciszke, 2014) and *Competence* as core social cognition dimensions, enabling analysis of empathy's cognitive-interpersonal aspects (MacDonald, 1992). *Trust* is the basis for believing in others' good intentions, which is necessary for emotional resonance; *Sociability* reflects active affiliation with others, supporting the interpersonal nature of empathy. *Competence* refers to the perceived ability of others to provide help or pose threats, guiding decisions to collaborate or avoid.

EAT posits that emotions are not solely cognitive products but direct bodily responses to environmental stimuli, which aligns with multiple viewpoints. For instance, the Greek philosophy

of mind's "qualia" concept holds human mind and cognition roots in sensory experiences. Enactive emotion theories (Hutto, 2012) view emotions as dynamic blends of body, cognition, and environment. Embodied cognition theories dismantle mind-body dualism, asserting cognition arises from body-environment coupling. In communication studies, the concept of "embodied presence" is proposed to describe how people immerse themselves in digital environments (Lindemann and Schünemann, 2020) .

Together, we tentatively infer that empathy may relate to "embodied imaginative resonance", rather than mere psychological projection.

3 Related Work

Here, we focus on works that extract linguistic features from empathetic expressions, alongside empathy prediction and classification tasks.

Linguistic Feature Extraction In early studies, researchers such as Gibson et al. (2015) relied on n-gram and Linguistic Inquiry and Word Count to extract linguistic features. They found that empathic therapists used more abstract perceptual language and reflective phrases such as "it sounds like". Herlin and Visapää (2016) identified via qualitative conversation analysis that a more prominent symmetric reference corresponds to greater emotional sharing, exemplified by the Finnish pronouns "se" (English "it") and "toi" (English "that"). Alam et al. (2016) captured 10k trigram, acoustic and psycholinguistic features from customer service calls, boosting Unweighted Agreement by 31% via majority voting. Similarly, Abdul-Mageed et al. (2017) found 10K unigrams and 50K bigrams optimal for identifying pathogenic empathy. Kann (2017) revealed that empathizers favored self-focused language, while sympathizers preferred other-focused language linked to charitable behaviors. Lee et al. (2024) showed idiom and metaphor features improved RoBERTa-twitter-sentiment performance in figurative empathy recognition.

Empathy Prediction and Classification Tasks Key empathy modeling studies have focused on framework development, model optimization, and task-specific performance improvement: Sharma et al. (2020) proposed the "Epitome" framework (dividing empathy communication into emotional reactions, interpretations, explorations) and developed a RoBERTa-based dual-encoder multitask model (with attention mechanism) for empathy recognition and rationale extraction. Buechel et al. (2018) predicted news-triggered empathy and personal distress, where the Convolutional Neural Network (CNN) achieved Pearson correlations of 0.404 for empathy and 0.444 for distress with human ratings, outperforming Ridge regression and Feed-Forward Network. Guda et al. (2021) proposed the demographic-aware EmpathBERT framework, yielding test set accuracies of 64.73% (male) and 64.56% (female). Dey and Girju (2022) enhanced BERT with FrameNet semantic features, achieving significant improvements in classifying cognitive, affective, and prosocial empathy in premed students' narrative essays. In follow-up work, Dey and Girju (2023) applied Construction and Systemic Functional Grammar theories to doctor-patient prose texts. Their findings showed that the Body Part + Process construction (e.g., "Her eyes welled up"), an important linguistic indicator, improved the BERT model's F_1 score by 7%.

To our knowledge, existing studies cover surface-to-semantic features and integrate text-based and cross-modal data, yet overlook empathy's cognitive-social essence and social media contexts. Methodologically, they prioritize prediction and dialogue generation over classification, while neural networks and LLM-oriented approaches, though widely adopted, face trade-offs between computational cost and the interpretability of linguistic communication mechanisms.

Therefore, our key contributions lie in expanding comprehensive empathy feature design for classification tasks and identifying empathy's underlying linguistic patterns.

4 Dataset

We collected 8,000 tweets using the retrieval hashtag **#California wildfires** (January 1–March 26, 2025), filtered noise and short posts ($<$ 2 words) to retain 6,246 samples, and pre-annotated them via DeepSeek-R1 (configured with a temperature of 0.1 and max_tokens of 1) using empathy's three component rules (emotion, cognition, behavior) (Hoffman, 1984) with labels: 0 (irrelevant), 1 (no empathy), 2 (empathy).

To verify the agreement between LLM pre-annotations and human annotations, we used Cohen's Kappa coefficient (via scikit-learn in Python) on 189 random samples, yielding Kappa $\approx$ 0.68. Both the LLM and the manual annotator followed the same annotation scheme (Appendix A). Taking

human annotations as the gold standard, we further evaluated LLM's classification performance and report precision/recall/F_1, with 0.70 overall accuracy, 0.91 precision for empathy (Category 2), and 0.75 recall for no empathy (Category 1) (Appendix 7). Subsequently, the trained researcher systematically reviewed and corrected all LLM-generated labels. Dataset distribution (Table 1) shows relative balance across empathy categories.

5 Lexical Analysis

We analyzed lexical differences between empathetic and non-empathetic texts, laying groundwork for precise feature design in later stages.

Inspired by the distributional consistency framework for distinguishing core lexicons Huang et al. (2005), as well as the application of normalized deviation of proportions (DP_{norm}) in fake news detection Wan et al. (2022), this study also employs DP_{norm} to quantify lexical usage differences, defined as:

$$DP_{\text{norm}} = \frac{DP}{1 - \min_i(s_i)}$$

DP is calculated as:

$$DP = \left(\sum_{i=1}^{n} |s_i - v_i| \right) /2$$

Here, s_i denotes the relative size of texts (proportion of text length in subset i), v_i the observed relative frequency of a lemma in subset i. The higher value shows stronger association with the subset.

We retained 113 discriminative lemmas with a threshold of $DP_{norm} \geq 0.7$, and present word clouds of selected lemmas: empathetic in Figure 1 and non-empathetic in Figure 2. Complete lists of lemmas are provided in Appendices 6 and 7.

Figure 1: Top 30 Empathetic Lemmas (sorted by DP_{norm} descending)

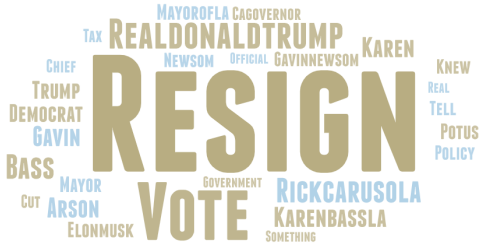


Figure 2: Top 28 Non-Empathetic Lemmas (sorted by DP_{norm} descending)

Key preliminary findings are summarized:

Empathy Targets: Empathetic texts consistently mention vulnerable groups (e.g., "community", "child") and blur in-group/out-group boundaries. Non-empathetic texts emphasize "power" and "authorities": their transactional, emotionally disengaged language aligns with Lewin (2013)'s approach-avoidance theory (suppressed "approach-connection" motives).

Empathy Triggers: Empathetic lexicons include solidarity terms ("pray", "give", "donate"), social process words, and moral-altruistic vocabulary, often tied to blessing and charity. Non-empathetic texts cluster around contentious topics (e.g., politics), which highlight in-group/out-group divisions (Vanman, 2016) and likely dampen empathy toward out-groups.

Despite lexical meaning being context-dependent, these patterns spark semantic feature design (e.g., sociality) in subsequent analyses.

6 Feature Design

From text data, we extracted three feature types: Cognitive, Perceptual, and Syntactic Features.

6.1 Cognitive and Perceptual Features

Sociability and Trust As introduced in SCM (*Warmth* (including *Sociability* and *Trust*) and competence are the two core dimensions of social cognition, which we rely on to evaluate individuals and groups. Here, we hypothesize that lexical use tied to high sociability and trust positively correlates with empathetic expression.

Trust and *Sociability* scores were computed by aggregating word-level scores from Words of Warmth norms (Mohammad, 2025), which quantify the inherent social-cognitive attributes of over 26,000 English words. (e.g., "prayer" : *Sociability score* [S] = 0.952, *Trust score* [T] = 0.848, competence [C] = 0.208; "resign" : S = -0.333, T = 0.273,

Table 1: Dataset distribution

Characteristics	Empathy	No Empathy	Irrelevant
Number of Samples	2,113 (33.83%)	2,995 (47.95%)	1,138 (18.22%)
Total Tokens	54,997 (40.15%)	59,204 (43.21%)	22,819 (16.64%)
Total Sentences	3,949 (33.17%)	5,639 (47.37%)	2,309 (19.46%)
Avg. Tokens/Sample	26.03	19.77	20.05
Avg. Sentences/Sample	1.87	1.88	2.03

Note: Total samples = 6,246; Total tokens = 137,020; Total sentences = 11,897

Text	Trust	Sociability
God bless.	0.870	0.955
God help!	0.741	0.939
Stupid,Ridiculous,Dangerous,Wasteful.	**-0.418**	**-0.678**
No conspiracy theorists without conspiracy terrorists.	**-0.455**	0.000

Table 2: Sentence-level score examples: Trust and Sociability dimensions

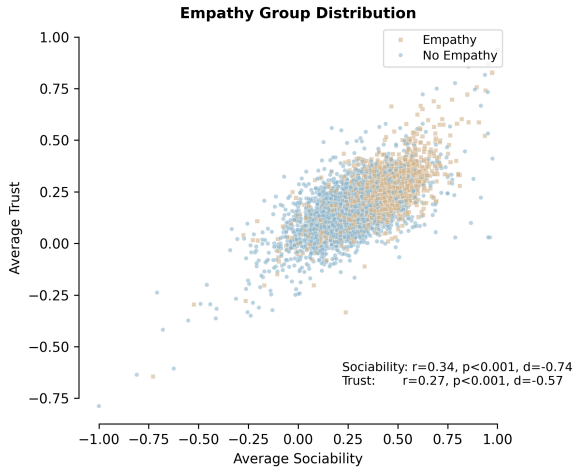


Figure 3: Dimensional distribution between S and T

C = -0.454). Sentence-level examples appear in Table 2. Figure 3 is the sentence-level distribution: *Sociability* (horizontal axis) exhibits a larger effect size (Cohen's d = 0.74) than *Trust* (vertical axis, Cohen's d = 0.57). This upward aggregation process is detailed in Appendix Figure 8.

Embodied Strength It reflects the strength of perceived embodied presence, defined as how strongly one feels physically immersed in others' environment. Grounded in EAT, we hypothesize that more specific sensory details in text will strengthen this embodied presence, thereby eliciting higher empathy.

Embodied scores were computed using Lancaster Sensorimotor Norms (Lynott et al., 2020), which provide sensory ratings for over 40,000 English words across six dimensions (Interoceptive, Auditory, Gustatory, Olfactory, Visual, and Haptic). The norms were used to compute embodied scores via a weighted formula, where coefficients (coef) determined via linear regression. Results (t = 5.36, p < 0.005, Cohen's d = 0.16) indicate that empathic individuals express solidarity by constructing vivid situational contexts that foster a sense of co-presence, as shown by example phrases (Figure 4, Table 3).

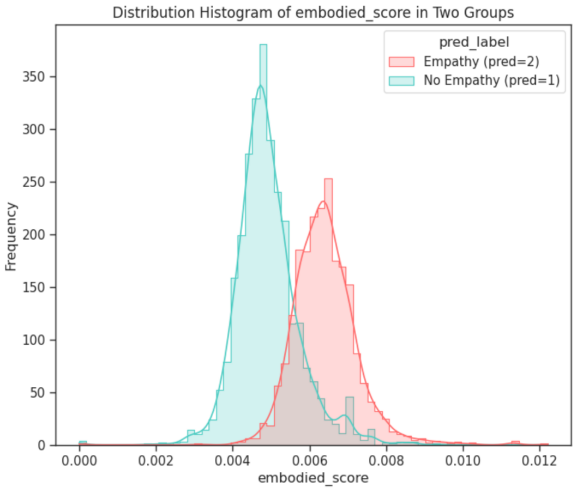


Figure 4: Distribution of embodied score in two groups

Table 3: Sentence-level embodied scores examples

High Embodied score (Top 3)
1. Indigenous man art pieces survived California fires.
2. News accidentally shows human skeleton.
3. I got a briefing at the Command Post and saw firsthand the devastation on Sunset Boulevard and Pacific Coast Highway. Let's come together to help the thousands of Angelenos who lost their homes.
Low Embodied score (Bottom 3)
1. What The Future For Los Angeles?
2. Forget the donate to the Dems.
3. What started the California wildfires?

6.2 Syntactic Features

Parts of Speech Using the Penn Treebank POS tagger, 36 POS tag ratios were generated. The most impactful feature, proper noun singular (NNP),

appears more frequently in non-empathetic texts (pred = 1). These include specific references such as names, places, and institutions (e.g., ARMAGEDDON, MAGA, NYFD), many of which are cultural symbols whose mention may trigger audience identification or controversy. By contrast, the empathy group (pred 2) uses more deindividuated language and vague references such as "those" (referring to victims) or "what they went through" (Figure 5).

This pattern aligns with Markedness theory (Francis, 2007), which distinguishes two types of linguistic choices: "marked" (deviant and attention-requiring) and "unmarked" (default and conventional). In empathetic communication, the unmarked strategy prioritizes de-individualization and shared emotional resonance, reducing individual specificity to broaden its reach.

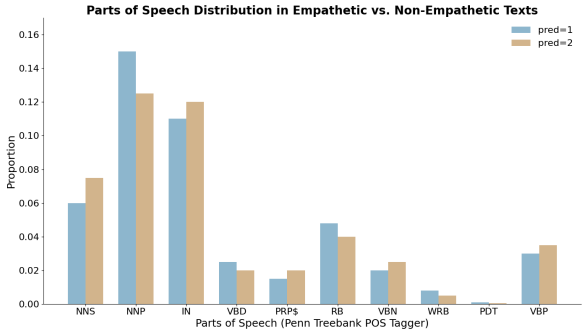


Figure 5: Distribution of parts of speech between empathetic and non-empathetic groups

Sentence Structures We analyzed 10 key sentence types using Stanford CoreNLP. Three showed significant group differences ($p < 0.01$), with imperatives most notable: the empathetic group used more imperatives (e.g., "Please help them") to emphasize action in emergencies, while the non-empathetic group preferred interrogatives (e.g., "Do you trust the foundation?") to focus on questioning. (Figure 6)

7 Predictive Power

Logistic Regression This experiment examined the correlation between proposed features (independent variables) and empathy levels (dependent variable) to identify the most predictive linguistic markers. Model performance was evaluated using average coefficients, p-values, and Cohen's d. Meanwhile, 10-fold cross-validation yielded an average accuracy of 0.68.

Results (Table 4) highlight avg_sociability as the most informative feature, followed by imperative

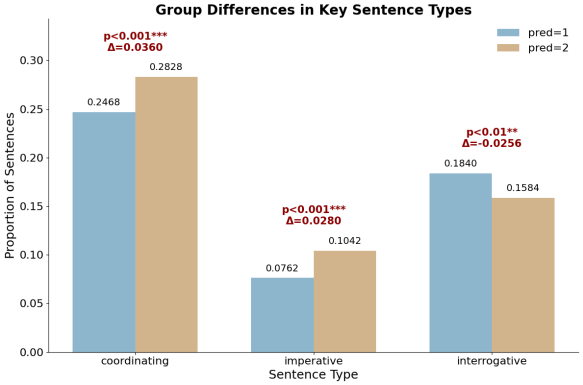


Figure 6: Distribution of sentence structures (Top 3 types) between empathetic and non-empathetic groups

proportion (imperative_prop), NNP, and Embodied score. Notably, the embodied score offers unique supplementary value as it illuminates empathy's linguistic patterns through sensory experience. Moreover, negative coefficients for avg_competence and interrogative_prop support prior findings: empathizers focus on vulnerable targets rather than ability evaluations. Non-empathetic individuals tend to use rhetorical questions, intentionally or unintentionally creating "difference" rather than pursuing emotional alignment.

Table 4: Variables importance ranking (10-Fold Cross-Validation)

Variable	Avg. Coef	Coef Std	Cohen's d	Sig.
avg_sociability	0.386	0.020	0.390	***
imperative_prop	0.253	0.058	0.128	***
NNP	-0.176	0.013	-0.168	***
Embodied	0.172	0.013	0.201	***
avg_competence	-0.120	0.010	0.067	*
avg_trust	0.117	0.017	0.309	***
interrogative_prop	-0.116	0.007	-0.084	**
politeness_score	0.104	0.008	0.183	***
coordinating_prop	0.024	0.015	0.101	**

Sig.: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

8 Machine Learning Models

Building on prior predictive power analysis, we further validated these features through classification tasks. Three models—logistic regression (LR), support vector machine (SVM), and random forest (RF), were used to test feature generalization. The top 3 key features (400D TF-IDF baseline supplemented by handcrafted features) were input to train the classifiers. Evaluation combined 10-fold cross-validation with an independent test set, with F_1 and AUC as core metrics.

Results (Table 5) show that Sociability, as the

most impactful feature, improves LR's F_1 by 4.68% over the baseline and consistently enhances F_1 and AUC across models. Combining all features (the top 3 features) yields the best performance, with SVM achieving the highest F_1 (0.689) and LR leading in AUC (0.718), verifying the synergistic and complementary effects of multi-dimensional features. However, there are differences in model adaptability: RF naturally adapts to heterogeneous handcrafted features, LR relies on feature weight optimization, and SVM is sensitive to feature combinations but has limited utilization of single features. In conclusion, subtle performance optimizations still matter in complex tasks such as empathy classification.

Table 5: Model Performance with TOP 3 Features

Model	Features	F_1 (Mean)	F_1 (Std)	AUC (Mean)	AUC (Std)
LR	Baseline	0.624	0.021	0.676	0.024
	S	0.671	0.022	0.712	0.023
	I	0.636	0.024	0.688	0.027
	NNP	0.625	0.021	0.676	0.024
	All	**0.683**	0.016	**0.718**	0.020
SVM	Baseline	0.614	0.016	0.648	0.027
	S	0.672	0.025	0.710	0.024
	I	0.634	0.026	0.672	0.027
	NNP	0.622	0.016	0.648	0.025
	All	0.689	0.020	0.722	0.024
RF	Baseline	0.597	0.017	0.627	0.023
	S	0.643	0.011	0.654	0.019
	I	0.629	0.016	0.633	0.017
	NNP	0.602	0.010	0.629	0.021
	All	0.673	0.027	0.712	0.030

S: avg_sociability; I: imperative_prop.

9 Conclusion

We adopted a comprehensive computational method to uncover empathy's communication patterns at the group level. Integrating SCM and EAT, we identified signals distinguishing empathetic vs. non-empathetic expression, analyzed usage differences, and validated their effectiveness via supervised learning in the empathy classification task. We summarized two new findings:

- **Empathy and sociability**: Sociability emerges as the strongest predictor of empathy (supported by DP-norm lists and model results). This highlights a key trait of empathetic language: stronger affiliation intentions drive richer empathetic communication.

- **Empathy and perceived embodied presence**: Greater sensory perception intensity correlates with stronger empathetic resonance. Theoretically, this extends the theory of communication's "embodied presence" to the context of social media disasters. Specifically, the empathetic group prefers concrete, vivid sensory details (e.g., "stand with you") to build a sense of co-presence.

These findings enrich empathy's multiple representations and highlight its interplay with cognition, embodiment, and interpersonal dimensions. We hope to pave the way for follow-up research on generalizing to near-synonymous affective expressions (e.g., "mercy," "care," "sympathy") or other implicit ones.

Limitations

First, potential annotation subjectivity exists despite extensive bias-mitigation measures. Second, we focus on textual features. Future work will expand contexts and integrate multimodal data to enhance real-world empathy modeling.

Acknowledgments

We thank members of the LLT (Linguistics and Language Technology) group at PolyU for their valuable suggestions, anonymous reviewers for their helpful feedback that definitely improved this paper, and the highly empathetic Twitter community for renewing our faith in human kindness.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Salvatore Giorgi, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing Pathogenic Empathy in social media. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 448–451, Montreal, Canada. AAAI Press.

Andrea E. Abele and Bogdan Wojciszke. 2014. Communal and agentic content in social cognition: A dual perspective model. *Advances in Experimental Social Psychology*, pages 198–255.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2016. Can we detect speakers' empathy?: A real-life case study. *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000059–000064.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on EMNLP*, pages 4758–4765, Brussels, Belgium. ACL.

Amy Coplan and Peter Goldie. 2011. *Empathy: Philosophical and Psychological Perspectives*. Oxford University Press.

P. Dey and R. Girju. 2022. Enriching deep learning with frame semantics for empathy classification in medical narrative essays. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 207–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Priyanka Dey and Roxana Girju. 2023. Investigating stylistic profiles for the task of empathy classification in medical narrative essays. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 63–74, Washington, D.C. Association for Computational Linguistics.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878–902.

Norbert Francis. 2007. Carol myers-scotton: Multiple voices: An introduction to bilingualism. *Applied Linguistics*, 28(1):155–158.

James Gibson, Nikos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Interspeech*.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *CoRR*, abs/2102.00272.

Ilona Herlin and Laura Visapää. 2016. Dimensions of empathy in relation to language. *Nordic Journal of Linguistics*, 39(2):135–157.

M. L. Hoffman. 1984. Interaction of affect and cognition in empathy. *Emotion, Cognition, and Behavior*, pages 103–131.

Chu-Ren Huang, Huarui Zhang, and Shiwen Yu. 2005. *On Predicting and Verifying a Basic Lexicon: Proposals inspired by Distributional Consistency*, pages 57–69.

Daniel Hutto. 2012. Truly enactive emotion. *Emotion Review*, 4:176–181.

Trevor Kann. 2017. *Measuring Linguistic Empathy: An Experimental Approach to Connecting Linguistic and Social Psychological Notions of Empathy*. Ph.D. thesis, University of California, Los Angeles.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 519–529, Bangkok, Thailand. ACL.

K. Lewin. 2013. *A Dynamic Theory of Personality - Selected Papers*. Read Books Limited.

Gesa Lindemann and David Schünemann. 2020. Presence in digital spaces. a phenomenological concept of presence in mediatized communication. *Human Studies*, 43(4):627–651.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Kevin MacDonald. 1992. Warmth as a developmental construct: An evolutionary analysis. *Child Development*, 63:753–773.

Saif M. Mohammad. 2025. Words of warmth: Trust and sociability norms for over 26k English words. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18830–18850, Vienna, Austria. ACL.

Becky Lynn Omdahl. 1995. *Cognitive Appraisal, Emotion, and Empathy*. Lawrence Erlbaum.

Rosalind W. Picard. 1997. Affective computing.

J.J. Prinz. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Philosophy of Mind. Oxford University Press.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on EMNLP*, pages 5263–5276, Online. ACL.

E. J. Vanman. 2016. The role of empathy in intergroup relations. *Current Opinion in Psychology*, 11:59–63.

Clara M. Wan, Qi Su, Rong Xiang, and Chu-Ren Huang. 2022. Data-driven analytics of covid-19 'infodemic'. *International Journal of Data Science and Analytics*, 15.

A Appendix

A Annotation

Prompt Instructions Classify tweets regarding the 2024-2025 California wildfires into three categories based on empathy expression:

2 (empathy): Tweets show empathy toward the wildfire disaster. Trigger if any of these apply:

- Affective Empathy: Expressions of sadness, sympathy, mourning, comfort, worry, or support (e.g., "Praying for families affected by California wildfires").
- Cognitive Empathy: Rational understanding of impacts (e.g., policy failures, environmental damage) or solution-oriented analysis (e.g., "Better forest management could reduce wildfire risks").
- Behavioral Intent: Calls to action (donations, volunteer work, advocacy) (e.g., "Donate to help wildfire victims").

1 (no empathy): Tweets lack empathy. Trigger:

- Detached/Sarcastic: Cynical, mocking, or critical tones (e.g., "California wildfires? Just nature's population control").
- Trivializing: Entertainment-focused or flippant framing (e.g., "Another year of barbecued California #wildfireseason").
- Indirect/Uncaring: Disaster-related context but no concern (e.g., "CaliforniaWildfires? Idiots in power caused this").

0 (irrelevant): Tweets are objective news updates unrelated to empathy)

Classification Criteria Leverage semantic content, emotional tone, disaster context to determine

- 2 (empathy): Emotional concern, rational understanding of impacts, or proactive support.
- 1 (no empathy): Cynicism, trivialization, or uncaring tones in a disaster context.
- 0 (irrelevant): Purely factual updates (no empathy/antipathy).

Examples align with these rules (e.g., "Praying for victims" = 2; "Wildfires? Just nature's way" = 1; "Fire size: 5,389 acres" = 0).

Experimental Metrics and Results

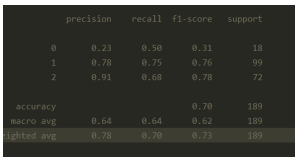


Figure 7: LLM Classification Performance

B Supplementary Materials

B.1 Complete Lexical Lists for Empathy and Non-Empathy

see Table 6 and Table 7.

Rank	Lemma	DP_norm	Rank	Lemma	DP_norm	Rank	Lemma	DP_norm
1	prayer	0.99	30	displaced	0.88	59	thanks	0.79
2	pet	0.98	31	community	0.88	60	neighbor	0.79
3	heart	0.98	32	family	0.87	61	critical	0.79
4	drive	0.98	33	thank	0.87	62	learn	0.79
5	shelter	0.98	34	helping	0.87	63	everyone	0.79
6	foundation	0.98	35	recover	0.87	64	member	0.79
7	impacted	0.96	36	hero	0.86	65	join	0.79
8	donate	0.96	37	donation	0.86	66	care	0.78
9	organization	0.95	38	relief	0.86	67	impact	0.78
10	animal	0.95	39	safe	0.86	68	update	0.78
11	volunteer	0.95	40	charity	0.86	69	devastation	0.77
12	affected	0.94	41	jan	0.85	70	lost	0.77
13	responder	0.94	42	open	0.85	71	benefit	0.76
14	difference	0.93	43	sending	0.85	72	center	0.76
15	rescue	0.93	44	survivor	0.84	73	offer	0.76
16	supporting	0.93	45	stay	0.84	74	recent	0.75
17	food	0.92	46	link	0.84	75	first	0.75
18	pray	0.92	47	devastating	0.84	76	important	0.75
19	together	0.92	48	child	0.83	77	across	0.75
20	providing	0.91	49	school	0.83	78	folk	0.74
21	resource	0.91	50	small	0.82	79	including	0.73
22	raise	0.90	51	help	0.81	80	donated	0.72
23	effort	0.90	52	recovery	0.81	81	stand	0.72
24	assistance	0.90	53	donating	0.81	82	supply	0.71
25	amazing	0.89	54	hope	0.80	83	service	0.71
26	altadena	0.89	55	please	0.80	84	firefighter	0.71
27	support	0.89	56	love	0.80	85	information	0.70
28	team	0.88	57	share	0.80			
29	working	0.88	58	friend	0.80			

Table 6: Distinct Lemmas for Empathy (85 words)

Rank	Lemma	DP_norm	Rank	Lemma	DP_norm	Rank	Lemma	DP_norm
1	resign	1.00	11	money	0.94	21	cagovernor	0.81
2	vote	1.00	12	president	0.93	22	chief	0.80
3	trump	0.98	13	potus	0.91	23	cut	0.79
4	rickcaruso	0.96	14	mayor	0.89	24	tax	0.74
5	bass	0.95	15	elonmusk	0.88	25	government	0.73
6	karenbass	0.95	16	gavinnewsom	0.88	26	something	0.73
7	arson	0.95	17	knew	0.87	27	real	0.73
8	right	0.94	18	newsom	0.86	28	official	0.71
9	gavin	0.94	19	tell	0.85			
10	democrat	0.94	20	policy	0.85			

Table 7: Distinct lemmas for non-empathy (28 words)

B.2 Calculation of Sentence-Level Trust and Sociability Scores

Combined lexicon-based exact matching with GloVe semantic retrieval (cosine threshold = 0.5) for unmatched tokens. (Figure 8).

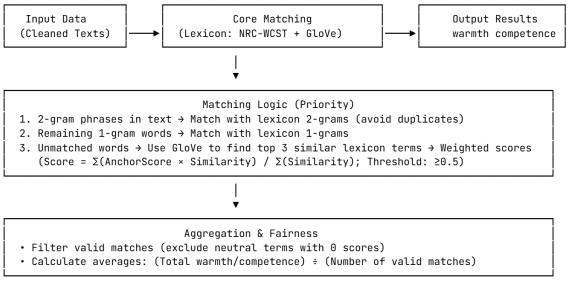


Figure 8: A flowchart showing the process of calculating sentence-level trust and sociability scores

Ancient Tamil Palm Leaf character recognition using Transformers

Aswinkumar I[1], Sangeetha Sivanesan[2], S Jaya Nirmala[2]
[1]Thiagarajar College of Engineering Madurai, Tamil Nadu, India,
[2]National Institute of Technology Tiruchirappalli, Tamil Nadu, India
tceaswin@gmail.com, sangeetha@nitt.edu, sjaya@nitt.edu

Abstract

Optical character recognition (OCR) for ancient Tamil scripts is highly challenging because of the variability, degradation and differences in styles of historical handwritten documents. This research was aimed at developing an effective OCR system for the Vattaeluthu script, which is a historical form of Tamil text, using transformer-based architecture. We created an updated labelled training dataset, which consisted of 15,423 unique glyph samples from 75 different categories. The dataset contains a combined manual annotation from palm-leaf glyphs as well as a significant amount of synthetic data augmented using an AI image augmentation technique. The experimental results demonstrate the potential of deep learning within this context. The CNN reached a classification accuracy of 90.63%, and the Transformer based model, which we optimized, achieved a classification accuracy of 96.18%. We ultimately combined the final models using a Streamlit-based user interface, enabling efficient character recognition, visualizations, and easy reconstruction of output. This research provides a solid foundation for digitally preserving ancient Tamil texts, as well as serving as a benchmark for future heritage optical character recognition research.

Keywords: Vattaeluthu, Tamil OCR, Deep Learning, TrOCR, CNN, Palm-leaf Manuscripts, Heritage Digitization, Glyph Recognition.

1 Introduction

South India's linguistic and cultural legacy is exemplified by the Vattaeluthu script, an ancient style of Tamil writing. It is a vital tool for comprehending historical writings, inscriptions, and classical Tamil literature. Even with its scholarly importance, mechanically recognizing and digitizing Vattaeluthu characters still presents difficulties for document interpretation and historical text preservation. The use of contemporary machine learning methods is restricted by the absence of sizable, annotated datasets. Recent developments in deep learning, particularly Transformers, have opened up new avenues for the recognition of ancient scripts in order to address these problems. These models can handle both stylistic variations and the physical damage often present in ancient manuscripts. A comprehensive OCR pipeline created especially for Vattaeluthu character recognition is shown in this study. The system had created by training CNN and finetuned the TrOCR models separately, and using a sizable synthetic dataset increases its efficacy. This improvement is essential for increasing the model's ability to generalize across various document circumstances and writing styles. By using this approach, we hope to promote a more comprehensive conservation of historical culture by improving the digital preservation and scholarly accessibility to Tamil cultural literature.

2 Literature Survey

The Optical Character Recognition for historical and ancient scripts has been indeed undergone a significant transformation, evolving from foundational classical methods to more sophisticated deep learning solutions, particularly in the context of ancient Tamil inscriptions. Earlier efforts in character recognition for ancient scripts frequently involved classical image processing techniques and machine learning algorithms. These approaches often served as a groundwork for understanding and digitizing historical texts.

Features like bends, loops, and curls from segmented characters, Scale Invariant Feature Transform (SIFT), corner detection and area property were extracted and Support Vector Machines (SVM), Artificial Neural Networks (ANN), and K-Nearest Neighbours (KNN) were were applied by

the researchers in 2019. (Merline Magrina and Santhi, 2019).

Current methods use automation and end-to-end learning processes to try to overcome the constraints of earlier approaches. Enhancement of images is now directly performed using Fully Convolutional Networks (FCNs) (Sivashanth et al., 2024), employing ResNet-50 backbones and DeepLabV3+ architectures as models. Dilated convolutions are used to extract contextual characteristics at various scales.

One notable system is DR-LIFT: Detection, Recognition, and Labeling Text Interpreter Framework, a customized deep learning framework that integrates several models across three primary stages to address inscriptions dating back to the third century BCE. The tasks include alphabet detection, Text detection with arbitrary shapes , Prediction of text shapes. It also adopts context-aware recognition Finally, for labeling and sentence recognition, a Neural Graph Machine (NGM) is adopted (Murugan and Visalakshi, 2024)

The old Vattaeluthu script has visually similar glyphs, and the extreme lack of digitized, annotated datasets for training contemporary machine learning models are the main causes of this difficulty. Additionally, previous efforts frequently relied on very small datasets, which limited the models' capacity to generalize across variations in handwriting, fading, and manuscript deterioration.

Our research addresses these issues in a substantially more thoroughly. We have curated a significantly larger and more diverse dataset by combining manually annotated palm-leaf manuscript samples with synthetically generated data to balance character class distribution. This augmentation process ensures that even the least frequently occurring glyphs are adequately represented in training, improving recognition consistency across the full character set.

Another key differentiator of this work is the introduction of Transformer-based architectures specifically the TrOCR model into the domain of ancient Tamil palm-leaf manuscript recognition. To the best of our knowledge, no prior work has been successfully applied Transformer models for Vattaeluthu OCR.

Using our enhanced dataset to refine the TrOCR model, this study establishes a new standard for historical Tamil OCR.

3 Proposed Work

The proposed work outlines a comprehensive system for the recognition of Tamil Vattaeluthu characters, encompassing meticulous dataset construction, a multi-stage processing pipeline, and the training and evaluation of advanced deep learning models.

3.1 Dataset Description

The foundation of this research relies on a robust and diverse dataset, critical for training high-performing deep learning models. The initial base for our dataset was derived from the Mendeley repository, specifically the "Tamil palmleaf Vatteluttu Dataset"(Sasikaladevi, 2024). This foundational collection originally comprised 7,100 glyph samples covering 71 distinct Vattaeluthu characters.

To ensure high-quality dataset generation and improve the diversity of character samples, a meticulous manual annotation process was carried out. We had been using Label Studio, an open-source data annotation tool, to annotate characters from raw, binarized palm-leaf manuscript images. These binarized inputs were derived from historically significant Tamil texts, including *Naladiyar* (27 original images, resulting in 26 usable ground truth images), *Tholkappiyam* (221 original images, yielding 163 usable samples), and *Thirikadugam* (14 original images, contributing 10 final samples) (Jailingeswari and Gopinathan, 2024).

The annotation process was performed entirely by hand. Each character within the palm-leaf manuscript was manually identified and enclosed using rectangular bounding boxes. This was done within Label Studio's tool as shown in Figure 1. [1]

Each character was cropped from the original binarized manuscript image using the bounding box coordinates. To maintain consistency with the original background conditions, padding was added around each character to simulate the black background used in earlier datasets. This step helped standardize the image dimensions and background contrast across all samples.

The final annotated dataset combined these newly processed samples with previously curated Mendeley data, ultimately producing a total of 2,180 manually labelled character images across 59 distinct ancient Tamil characters. This refined dataset forms a robust foundation for training deep

[1] Label Studio

Dataset	# Characters	# Images
Mendeley dataset (Sasikaladevi, 2024)	71	7,100
Manually labeled instances	59	2,180
AI generated instances	68	6,143
Combined Dataset	**75**	**15,423**

Table 1: Dataset composition from Mendeley repository, manual annotation, and AI-generated augmentation.

learning models in ancient Tamil OCR tasks.

Recognizing the potential for class imbalance and the need for a larger volume of training data, we implemented a sophisticated augmentation strategy using Gemini AI. A custom Python script was developed to interact with the Gemini multimodal generative model.

The script was configured to take an existing glyph image and generate diverse handwritten-style variants. The core prompt provided to Gemini AI instructed it to:
[2]

> "You are generating synthetic training data for a historical Tamil Vatteluttu OCR model. Given a single example glyph image, generate diverse handwritten-style variants that represent the same character. Apply variations in stroke thickness, stroke curvature, ink flow, brush/pen pressure, natural writing slant, and aging effects such as mild fading or bleed. Ensure each generated image contains only one glyph, centered, with no noise, borders, artifacts, or text outside the character. Output each image as a clean black glyph on a transparent or pale off-white background, in a centered square format (e.g., 224×224). Maintain legibility while simulating realistic handwriting diversity."

The process of creating the dataset was to balance the 200 photos per class over all 75 characters. The script used `gemini-2.5-flash` as the model and asked for 5 image variants per API call. This AI-driven augmentation significantly expanded the dataset, bringing its final size to 15,423 images as shown in Table 1. For consistent model input, every image in the final dataset was normalized and reduced to 448×448 pixels.

Finally, a Python script was used to cross-validate the similarity between the real images and the images generated by Gemini AI. This process measured the Cosine Similarity for each class of Tamil Vattaeluthu characters to ensure the AI generated images were a close match to the real ones. The results of this comparison are shown in Table 6.

3.2 Methodology

The proposed OCR pipeline for Tamil Vattaeluthu character recognition follows a systematic multi-stage approach, designed to process raw image inputs into recognizable text outputs. The system architecture facilitates both preprocessing and segmentation of glyphs before feeding them into deep learning classifiers for recognition. As shown in Figure 2, each phase of this pipeline is detailed in the following subsections.

3.2.1 Pre-Processing

The first phase in processing any palm leaf manuscript image is to prepare it for accurate character recognition by enhancing its visual clarity and structural consistency. To begin, the image is converted into shades of grey, reducing it to essential visual data by removing color distractions. This simplification lays the foundation for further refinement. Then, a method called Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied that enhances contrast in limited areas and then it is used to make low-contrast or faded characters more noticeable. This guarantees that even the manuscript's faintest markings are visible. A soft-focus filter is used to softly smooth the image once the contrast has been improved. This keeps each character's edges crisp and undamaged while reducing undesired graininess and background noise. The image is then converted to black-and-white using adaptive thresholding that adjusts to changes in illumination in various manuscript sections. Because uneven exposure and fading are typical in historical manuscripts.

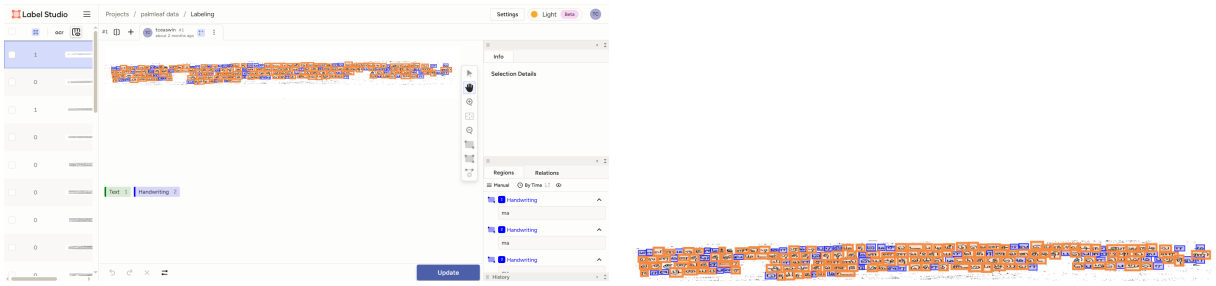
The outcome of this stage is a cleaner, sharper image where the individual characters are well-defined and ready for precise segmentation. [3]

3.2.2 Segmentation

The next stage is to meticulously dissect the manuscript image into its component characters after it has been improved and transformed into a binary format. This is accomplished by identifying each character's outer bounds within the

[2] Gemini app

[3] Open Cv



(a) Manual mapping of letters in Label Studio.

(b) Binarized palm leaf with marked letters.

Figure 1: Label Studio annotation process: (a) bounding box mapping of characters, (b) annotated binarized palm leaf manuscript.

image. Following the identification of possible character shapes, a rigorous filtering procedure is used to guarantee that only significant and properly formed glyphs are retained for additional examination. Every shape that is detected is assessed using a set of real-world criteria:

First, the size of the image is assessed. Essentially, specks or tiny dots that might be noise are eliminated since only those that take up a sizable portion of the visual field are kept.

The ratio of the shape's height to width is evaluated. A shape is removed if it looks too stretched in one direction, because likely a mistake or unwanted mark and not a real character.

The valid segmented glyphs are first scaled and padded to uniform proportions before being given into the classifier. The extracted ROIs are specifically prepared to meet the input dimensions needed by the classification model (e.g., 128×128 for CNN, or 384×384 for TrOCR) by the resize and pad function. Glyphs are sorted line-wise after segmentation, first by grouping according to their Y-axis position and then by sorting along the X-axis inside each line.

3.2.3 Vattaeluthu Script Recognition Models

Feature extraction is the core mechanism by which the deep learning models learn meaningful representations from the pre-processed glyph images.

TrOCR Model To accurately recognize ancient script characters, a highly capable deep learning model known as TrOCR is employed(Li et al., 2021). This model is based on an advanced Transformer architecture pre-trained on a large variety of handwritten text, providing a strong foundation for understanding intricate visual patterns.

Similar to neural machine translation models, TrOCR follows a two-stage encoder–decoder design as shown in Figure 3.

Encoder: In our implementation, the encoder is adapted to work directly with convolution-based feature extraction. This modification aligns with our specific dataset characteristics and the challenges of ancient palm-leaf manuscripts. The encoder processes the input glyph image to extract high-level visual features while preserving spatial relationships essential for accurate character recognition.

Decoder: The extracted feature representations are passed to a Transformer-based language decoder, which interprets these visual patterns and generates the corresponding character sequence. The decoder handles sequential dependencies, making it capable of performing both individual glyph classification and continuous text recognition.

Feature Extraction in the Encoder The encoder processes the grayscale glyph image through multiple convolutional layers to identify edges, curves, and finer script patterns. Batch normalization and activation functions ensure stable learning. The processed feature maps are then transformed into a sequence of feature tokens understandable by the Transformer decoder.

Training Objective Loss Function: Typically, Cross-Entropy Loss is used during training to compare predicted sequences with the ground-truth labels.

The focus here is the model is fine-tuned using individual characters, with the focus on character-level training rather than the conventional full-text transcription. The image of a single glyph is first processed by the model's encoder, which extracts deep and meaningful visual features. These features are then passed through an additional processing layer that helps convert this visual understanding into a specific character prediction.

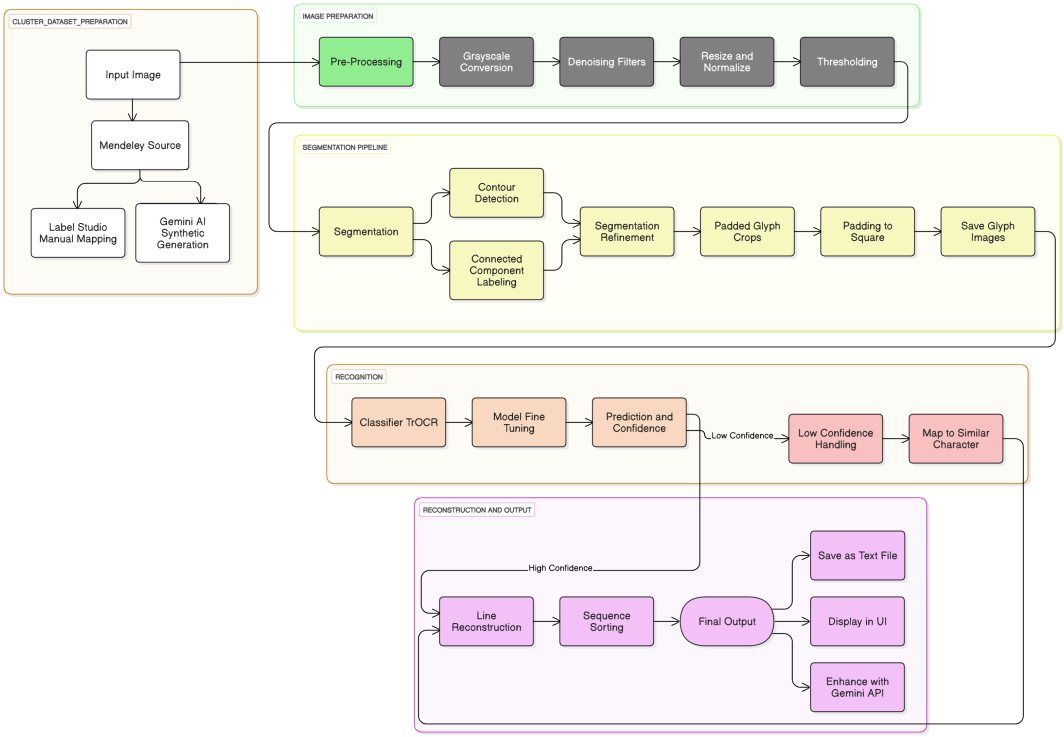


Figure 2: System Architecture Diagram.

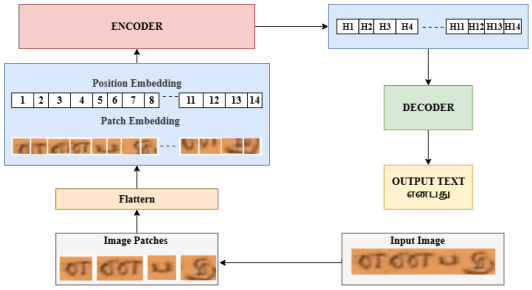


Figure 3: TrOCR architecture diagram.

The model has been adapted to classify characters from a set of 75 distinct symbols, each representing an ancient glyph. By doing this, the system uses the Transformer's ability to detect even the smallest visual features to convert visual patterns into precise character labels.

All glyph pictures are downsized to a square format with equal height and width, each measuring 384 pixels, to provide uniformity and compatibility with the model's training settings. Regardless of the scale or shape of the original image, this resizing aids to the improvement of model's accuracy.

Fine-Tuning TrOCR Model for Ancient Tamil Script Recognition

The Transformer-based OCR model, optimization also used the Adam algorithm with a moderate learning rate suitable for fine-tuning. To avoid overtraining and reduce unnecessary computation, early stopping was enabled. If validation accuracy failed to improve over five consecutive rounds, training would automatically stop, and the best-performing version of the model would be retained. The full training process was limited to a maximum of fifty rounds to strike a balance between thorough learning and computational efficiency as shown in table 8.

This made it possible for the system to begin with a solid grasp of common handwriting patterns and then modify it to fit the particulars of the old script. The training procedure concentrated on individual glyphs, each coupled with its appropriate descriptor, rather than overall sentence-level input for training.

The model's final classification layer was modified to identify 75 different characters used in the script in order to conform to the particular task. To do this, the output stage of the model had

to be resized to fit the target character set.

The entire dataset was meticulously divided into three parts to provide a fair evaluation as shown in table 7. The remaining sample was utilized just to assess the model's accuracy on previously unseen data, while a smaller portion was used to validate the model's progress during training and the rest was used to train the model.

CNN Model A bespoke Convolutional Neural Network (CNN) was trained utilizing an adaptable deep learning architecture in order to compare the effectiveness of Transformer Model with the standard CNN. The model received a constant stream of image samples during training.

The training was carried out using the Adam optimizer algorithm with a carefully chosen low learning rate to ensure stable convergence. The augmentation methods including comprised slight brightness shifts, moderate contrast alterations, and random horizontal flips. These augmentation methods improved the model's capacity to generalize to unseen glyph samples by making it more resistant to changes in illumination.

To balance performance and efficiency, two key strategies were used during training: one to automatically adjust the learning rate when learning rate is reduced, and another to stop training early if no improvement was observed when compared to previous iterations. This will prevent overfitting, even though the training process was permitted to run for up to 100 complete passes through the dataset.

3.3 Results and Discussion

As per the evaluations, the TrOCR-based classifier and CNN-based classifier are both capable of identifying Tamil Vattaeluthu characters. A comparison investigation demonstrates how much better the Transformer-based architecture performs.

On all important criteria, however, the improved TrOCR model performed noticeably better than the CNN. A remarkable 96.18% test accuracy was attained. Specifically, TrOCR's macro-averaged precision, recall, and F1-score were 0.9614, 0.9634, and 0.9616 as shown in Table 2.

Character Error Rate (CER), Word Error Rate (WER), and Levenshtein Distance(LD) were among the other text-based evaluation measures that were calculated to give a more thorough picture of the models' performance. With a CER of 0.0208, the TrOCR model demonstrated an extremely low percentage of wrong characters in comparison to all anticipated characters. The partial nature of glyph-based predictions, where accuracy at the character level affects complete word reconstruction, was reflected in the Word Error Rate (WER), which was assessed at 0.5000. The model's capacity to provide predictions with little textual deviance was further supported by the discovery that the Average Levenshtein Distance between the predicted and ground truth sequences was 0.50 as shown in Table 3. These extra measures highlight the model's proficiency in producing character sequences that are almost accurate in addition to its classification accuracy, which is crucial for developing trustworthy OCR systems for old scripts.[4]

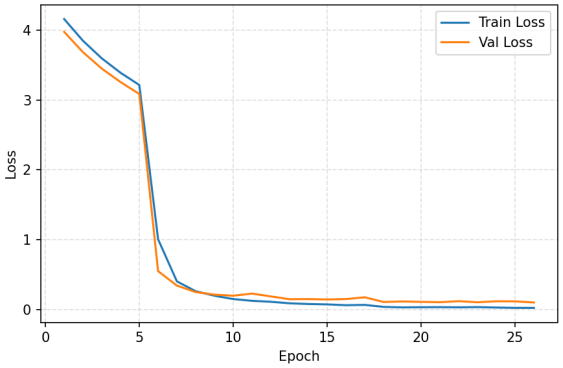


Figure 4: TrOCR fine-tuning: Train Loss vs Validation Loss.

This feature enables TrOCR to more accurately decipher the intricate and frequently deteriorated characteristics of Vattaeluthu glyphs, resulting in more reliable recognition, especially on difficult and deteriorated samples.

An overall accuracy of 90.63% was attained by the CNN classifier. Additionally, precision, recall, and F1-score were used to assess its performance; the macro-averaged results were 0.9086, 0.9058, and 0.9056, respectively. A thorough classification report showed that different character classes performed differently. Some classes performed worse than others, such Glyphs like '൹', which had a precision of 0.41 and recall of 0.38, but several classes had high scores (e.g., Glyphs '◌ൻ': 1.00 precision, 0.95 recall; Glyphs 'ഩ': 1.00 accuracy, 1.00 recall; Glyphs 'ഇ': 1.00 precision, 1.00 recall). This draws attention to certain difficulties the CNN had with particular glyph variations.

[4] Github Link of the code

The final fine-tuned model of TrOCR, have been successfully deployed with a Streamlit-based user interface.

The effect of dataset frequency on recognition accuracy is a significant finding from the assessment procedure. In the manually gathered dataset, a number of glyphs were extremely rare, resulting in an imbalance in the data that can jeopardize the generalization of the model. In order to guarantee a minimum of 200 samples per class, synthetic augmentation was carried out on low-occurrence characters. The training process was greatly improved by this endeavor, particularly for characters with limited representation.

Interestingly, the classification results indicate that both rare and frequently occurring characters yielded very high accuracy, suggesting that the model learned to generalize well even from initially imbalanced data thanks to synthetic augmentation. For example:

Glyphs like 'ீ' and 'ீ', which originally had 0 samples in manual annotation, achieved precision and recall near or equal to 1.00 after augmentation.

Characters with extremely low original frequency like 'ணி' (1), 'ளி' (0), 'தீ' (0), and 'கூ' (4) in manual annotation were also recognized with high precision and recall.

Conversely, high-frequency glyphs, such as 'ா' (300), 'ம' (143), and 'க' (150), also achieved consistently excellent recognition metrics, reinforcing the robustness of the augmentation strategy.

This highlights the efficacy of the data synthesis pipeline driven by Gemini AI and shows the model's ability to function consistently across different character distributions. In addition to improving classification metrics, the balanced dataset made the system more applicable to historical documents with irregular glyph distributions in the actual world.

Model	Accuracy	Precision	Recall	F1 Score
TrOCR	96.18%	0.9614	0.9634	0.9616
CNN	90.63%	0.9086	0.9058	0.9056

Table 2: Model Performance Metrics.

Model	CER	WER	Avg.LD
TrOCR	0.0208	0.5000	0.50
CNN	0.2292	1.0000	5.50

Table 3: Character Error Rate, Word Error Rate, and Average Levenshtein Distance for both models.

4 Conclusion

This research presents a system designed for ancient Tamil Vattaeluthu scripts. It achieves high accuracy by Transformer-based TrOCR. Our dual-model approach compares a custom CNN with a fine-tuned Transformer-based TrOCR architecture. This comparison shows the advantages of Transformer-based models in classifying ancient glyphs. The TrOCR model effectively learns complex patterns and long-range dependencies.

This work offers several contributions. We created a large, carefully curated dataset through manual annotation and innovative AI-driven synthetic data augmentation. We also developed a complete image processing and segmentation pipeline and successfully deployed the recognition models. This system provides a useful resource and sets a new standard for the digital preservation and accessibility of Tamil heritage texts.

Acknowledgments

We sincerely thank all the dataset contributors and the active open-source community for their efforts in improving this field including Mendeley Data.

References

I Jailingeswari and S Gopinathan. 2024. Tamil handwritten palm leaf manuscript dataset (thplmd). *Data in Brief*, 53:110100.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: transformer-based optical character recognition with pre-trained models (2021). *arXiv preprint arXiv:2109.10282*.

M Merline Magrina and M Santhi. 2019. Ancient tamil character recognition from epigraphical inscriptions using image processing techniques. matjournals 4 (2): 40–48.

Balasubramanian Murugan and P Visalakshi. 2024. Ancient tamil inscription recognition using detect, recognize and labelling, interpreter framework of text method. *Heritage Science*, 12(1):1–21.

Sasikaladevi. 2024. Tamil palmleaf vatteluttu dataset.

Suthakar Sivashanth, Eugene Yugarajah Andrew Charles, and Siyamalan Manivannan. 2024. A fully automated approach for enhancing ancient tamil inscriptions using deep learning. In *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1–5. IEEE.

Appendix

Class	Prec.	Rec.	F1	Class	Prec.	Rec.	F1	Class	Prec.	Rec.	F1
மூ	1.00	0.92	0.96	ட	1.00	0.95	0.98	ப	1.00	0.95	0.97
ா	0.98	0.98	0.98	ம	0.97	1.00	0.99	ெ	0.93	0.81	0.86
ய	0.96	0.92	0.94	ல	1.00	1.00	1.00	ன	0.95	0.91	0.93
ற	1.00	1.00	1.00	க	0.91	0.95	0.93	வ	0.91	1.00	0.95
ண	0.95	0.90	0.92	த	0.95	1.00	0.98	ச	1.00	0.93	0.97
ங	1.00	1.00	1.00	ள	0.94	1.00	0.97	தி	1.00	1.00	1.00
வி	1.00	0.90	0.95	றி	1.00	1.00	1.00	லி	0.92	1.00	0.96
னி	1.00	1.00	1.00	யி	1.00	0.90	0.95	ழி	0.95	1.00	0.98
பி	1.00	1.00	1.00	ரி	1.00	1.00	1.00	சி	0.96	1.00	0.98
ணி	1.00	1.00	1.00	மி	0.96	0.96	0.96	கி	0.88	1.00	0.94
டு	0.80	0.94	0.86	கு	1.00	1.00	1.00	ளு	0.52	0.59	0.55
லு	1.00	0.87	0.93	மு‌	0.86	0.92	0.89	ணு	0.94	0.94	0.94
னு	0.95	0.95	0.95	ஞு	1.00	1.00	1.00	பு	0.96	1.00	0.98
று	1.00	1.00	1.00	ருு	1.00	1.00	1.00	சு	1.00	1.00	1.00
து	1.00	1.00	1.00	வு	0.89	0.89	0.89	யு	0.94	1.00	0.97
ழு	1.00	1.00	1.00	டி	1.00	1.00	1.00	ளி	0.96	1.00	0.98
எ	1.00	1.00	1.00	ழு	1.00	1.00	1.00	கீ	1.00	0.93	0.96
றூ	1.00	1.00	1.00	மீ	1.00	1.00	1.00	வீ	1.00	1.00	1.00
நூ	1.00	1.00	1.00	றீ	1.00	1.00	1.00	தீ	1.00	1.00	1.00
கூ	1.00	1.00	1.00	தூ	1.00	1.00	1.00	சூ	0.91	1.00	0.95
யீ	1.00	1.00	1.00	லூ	0.96	1.00	0.98	உ	1.00	1.00	1.00
அ	1.00	1.00	1.00	ழீ	1.00	1.00	1.00	யூ	1.00	1.00	1.00
சீ	1.00	1.00	1.00	ணீ	1.00	0.95	0.97	ஆ	1.00	1.00	1.00
இ	0.94	1.00	0.97	ை	0.88	0.94	0.91	ர	0.93	0.93	0.93
ந	0.97	1.00	0.98	ஒ	0.96	0.96	0.96	ளூ	1.00	1.00	1.00
Accuracy: (0.96), **Macro Avg:** (0.96, 0.96, 0.96), **Weighted Avg:** (0.96, 0.96, 0.96))											

Table 4: Detailed classification report for TROCR model on Tamil character recognition.

Class	Prec.	Rec.	F1	Class	Prec.	Rec.	F1	Class	Prec.	Rec.	F1
மூ	0.83	0.87	0.85	ா	1.00	0.95	0.97	ட	1.00	0.91	0.95
ம	0.82	0.78	0.80	ப	0.95	0.90	0.93	ெ	0.92	0.90	0.91
ய	1.00	0.76	0.86	ல	0.90	0.90	0.90	ன	0.80	0.86	0.83
ற	1.00	0.90	0.95	க	0.85	0.85	0.85	வ	0.66	0.86	0.75
ண	0.93	0.91	0.92	த	0.89	0.92	0.90	ச	0.91	0.75	0.82
ங	0.96	0.98	0.97	ள	0.85	0.85	0.85	தி	0.89	0.97	0.93
வி	0.94	0.88	0.91	றி	0.95	0.97	0.96	லி	0.91	0.94	0.93
னி	0.89	0.97	0.93	யி	0.92	0.80	0.86	ழி	0.97	0.95	0.96
பி	1.00	0.86	0.93	ரி	0.97	0.97	0.97	சி	0.97	0.84	0.90
ணி	1.00	0.90	0.95	மி	0.84	0.92	0.88	கி	0.82	0.77	0.79
டு	0.86	0.79	0.83	கு	0.76	0.89	0.82	ளு	0.41	0.38	0.39
லு	0.88	0.95	0.92	முு	0.82	0.87	0.85	ணு	0.94	0.81	0.87
னு	0.78	0.88	0.83	ஞு	0.88	0.86	0.87	பு	0.98	1.00	0.99
று	0.86	0.90	0.88	ருு	0.96	0.81	0.88	சு	0.95	0.91	0.93
து	0.90	0.96	0.93	வு	0.93	0.98	0.95	யு	0.98	0.96	0.97
ழு	0.84	0.78	0.81	டி	0.94	0.98	0.96	ளி	0.93	0.97	0.95
எ	0.90	1.00	0.95	ழு	0.92	0.92	0.92	கீ	1.00	0.98	0.99
றூ	0.95	0.97	0.96	மீ	0.89	0.91	0.90	வீ	0.89	0.97	0.93
நூ	1.00	0.98	0.99	றீ	0.97	0.92	0.95	தீ	0.95	0.95	0.95
கூ	0.94	1.00	0.97	தூ	0.95	0.97	0.96	சூ	0.90	0.95	0.93
யீ	0.97	0.88	0.92	லூ	0.95	0.95	0.95	உ	1.00	0.97	0.99
அ	0.94	0.94	0.94	ழீ	0.95	0.95	0.95	யூ	1.00	1.00	1.00
சீ	1.00	0.97	0.99	ணீ	1.00	0.97	0.99	ஆ	0.94	1.00	0.97
ளூ	0.52	0.60	0.56	இ	0.85	0.95	0.90	ை	0.90	0.96	0.93
ர	0.98	0.96	0.97	ந	0.92	1.00	0.96	ஒ	1.00	1.00	1.00
Accuracy: (0.91), **Macro Avg:** (0.91, 0.91, 0.91), **Weighted Avg:** (0.91, 0.91, 0.91)											

Table 5: Detailed classification report for CNN model on Tamil character recognition.

Class	Avg Similarity	Class	Avg Similarity	Class	Avg Similarity
மூ	0.86	ா	0.93	ட	0.88
ம	0.92	ப	0.92	ெ	0.92
ய	0.90	ல	0.93	ன	0.92
ற	0.90	க	0.94	வ	0.87
ண	0.92	த	0.92	ச	0.87
ங	0.92	ள	0.88	தி	0.79
வி	0.93	றி	0.79	லி	0.71
னி	0.87	யி	0.90	ழி	0.89
பி	0.90	ரி	0.89	கி	0.75
ணி	0.78	மி	0.89	கி	0.76
டு	0.91	கு	0.88	ளு	0.78
லு	0.74	மு	0.86	ணு	0.76
னு	0.89	ஞு	0.71	பு	0.91
று	0.88	ருபு	0.91	சு	0.71
து	0.89	வு	0.76	யு	0.90
ழு	0.76	டி	0.90	ளி	0.82
எ	0.84	ழ	0.83	கீ	0.88
றூ	0.87	மீ	0.90	வீ	0.80
நூ	0.89	றீ	0.87	தீ	0.89
கூ	0.76	தூ	0.84	சூ	0.75
யீ	0.88	லூ	0.89	உ	0.80
அ	0.78	ழீ	0.89	யூ	0.81
சீ	0.85	ணீ	0.87	ஆ	0.82
ளூ	0.90	இ	0.81	ை	0.90
ர	0.91	நி	0.90	ஒ	0.90
Total Avg Similarity : 0.86					

Table 6: Cosine similarity between Gemini AI generated and real images for each class.

Model	Train	Validation	Test
TrOCR	70%	20%	10%
CNN	80%	0%	20%

Table 7: The split ratio of the dataset.

Hyperparameter	Value/Setting
Learning Rate (LR)	
LR_ENCODER	1e-5
LR_CLASSIFIER	1e-4
Optimizer Settings	Adam optimizer
Encoder params	lr=1e-5
Classifier params	lr=1e-4
Batch Size	16
Loss Function	Cross-Entropy Loss
Early Stopping	PATIENCE = 5
Maximum Epochs	50

Table 8: TrOCR Critical Hyperparameters.

MTikGuard System: A Transformer-Based Multimodal System for Child-Safe Content Moderation on TikTok

Dat Thanh Nguyen[1,4,5], **Nguyen Hung Lam**[3,4,5],
Anh Thi-Hoang Nguyen[1,4,5], **Trong-Hop Do**[2,4,5]

[1]Faculty of Information Science and Engineering, University of Information Technology
[2]Faculty of Software Engineering, University of Information Technology
[3]Faculty of Computer Science, University of Information Technology
[4]University of Information Technology, Ho Chi Minh City, Vietnam
[5]Vietnam National University, Ho Chi Minh City, Vietnam

Emails: {22520224, 22520968}@gm.uit.edu.vn, {anhnth, hopdt}@uit.edu.vn

Corresponding author: Trong-Hop Do (hopdt@uit.edu.vn)

Abstract

With the rapid rise of short-form videos, TikTok has become one of the most influential platforms among children and teenagers, but also a source of harmful content that can affect their perception and behavior. Such content, often subtle or deceptive, challenges traditional moderation methods due to the massive volume and real-time nature of uploads. This paper presents **MTikGuard**, a real-time multimodal harmful content detection system for TikTok, with three key contributions: (1) an *extended* TikHarm dataset expanded to 4,723 labeled videos by adding diverse real-world samples, (2) a multimodal classification framework integrating visual, audio, and textual features to achieve **state-of-the-art** performance with 89.37% accuracy and 89.45% F1-score, and (3) a scalable streaming architecture built on Apache Kafka and Apache Spark for real-time deployment. The results demonstrate the effectiveness of combining dataset expansion, advanced multimodal fusion, and robust deployment for practical large-scale social media content moderation. The dataset is available at https://github.com/ntdat-8324/MTikGuard-System.git.

Disclaimer: This paper contains images and content from social networks that may be considered sensitive, harmful, or related to adult content and suicide.

1 Introduction

Harmful video content refers to videos containing visuals, audio, or messages that may negatively impact viewers' cognition, emotions, and behaviors - particularly those of children. Such content may involve depictions of violence, explicit language, antisocial behavior, hate speech, or dangerous challenges disguised as "fun" activities that children can easily imitate. In the era of short-form video as a dominant mode of entertainment, the spread of harmful content occurs faster than ever before, often beyond the supervision of parents or schools. Children, with limited analytical and self-protective capabilities, are especially vulnerable to these negative role models, often failing to recognize their long-term consequences for personality development and behavior. Consequently, the early detection and timely warning of harmful video content are essential to maintaining a safe online environment for young audiences.

Online platforms, particularly social media, enable harmful videos to proliferate rapidly, making control and prevention increasingly challenging. To address this issue, we build upon and extend the publicly available TikHarm dataset to improve model training effectiveness and develop a system that can curb the spread of harmful trends targeting children.

In this study, we expand the TikHarm dataset by collecting additional videos following the original selection criteria, thereby improving the model's learning capacity and robustness in real-world deployment. Leveraging this extended dataset, we introduce **MTikGuard**, *a real-time harmful content detection system for TikTok*. MTikGuard processes multiple modalities from videos, including visual frames and textual information (extracted via OCR and speech-to-text), enabling richer contextual understanding (see Appendix A.1 for the formal task definition). With a multimodal transformer-based architecture, MTikGuard outperforms single-modal baselines, particularly in detecting harmful trends presented in subtle or disguised forms. Our main contributions are summarized as follows:

- We expand the TikHarm dataset by collecting and annotating an additional 775 videos, resulting in a total of 4,723 labeled samples.

This extension improves model robustness and generalization.

- We propose a transformer-based multimodal architecture for harmful video detection that integrates visual frames, OCR-extracted text, and speech-to-text transcripts. By jointly modeling these modalities, the system captures nuanced cross-modal cues, enabling more accurate identification of harmful content, even when disguised or indirectly presented.

- We introduce the MTikGuard system, a real-time video analysis and detection pipeline for harmful content. Beyond its robust multimodal model, MTikGuard integrates advanced technologies such as Apache Spark, Apache Kafka, Docker, and Apache Airflow, enabling scalable, reliable deployment for detecting harmful content from online sources in production environments.

2 Related Work

Many studies have focused on detecting inappropriate content using only video data. (Singh et al., 2019) introduced the KidsGUARD system, which utilizes an LSTM-based autoencoder to learn video representations from features extracted by the VGG16 CNN network, aiming to detect unsafe content for children. The model achieved a recall of 81% and a precision of 80% on a dataset of 109,835 video clips. (Yousaf and Nawaz, 2022) proposed a method using EfficientNet-B7 for feature extraction and a BiLSTM network for video representation to classify inappropriate content in animated YouTube videos, achieving 95.66% accuracy and an F1-score of 92.67%. (Murthy et al., 2024) conducted a study applying computer vision techniques to detect e-cigarette-related content in TikTok videos, building an object detection model based on YOLOv7 that achieved a recall of 77%, precision of 86.30%, and F1-score of 81.40%. Transformer-based models such as TimeSformer (Bertasius et al., 2021), VideoMAE, and ViViT have also been effectively applied in video classification, enhancing accuracy and efficiency in detecting inappropriate content. Notably, TikGuard (Balat et al., 2024) employs advanced Transformer models like TimeSformer, VideoMAE, and ViViT to classify TikTok content, achieving 86.7% accuracy on the TikHarm dataset, demonstrating the potential of these models in detecting harmful content for children.

Recognizing the limitations of relying solely on video data, recent studies have shifted towards multimodal approaches, combining information from video, audio, and text to improve the detection of inappropriate content. (Phukan et al., 2025) proposed SNIFR, a new framework that leverages Transformers to integrate audio and visual information for detecting harmful content targeting children, outperforming single-modality methods. (Das et al., 2023) introduced HateMM, a multimodal dataset comprising approximately 43 hours of video from BitChute labeled as hate or non-hate, along with frame-level rationale annotations for label decisions. Their findings showed that incorporating all modalities significantly improved overall hate speech detection performance (accuracy = 79.80%, macro F1-score = 79%), with a gain of approximately 5.7% in macro F1-score compared to the best single-modality model.

3 Dataset

3.1 TikHarm Dataset

TikHarm is a dataset developed to support the task of detecting harmful content on the TikTok platform, with a particular focus on protecting children from such content. The TikHarm dataset is available at Link . The dataset consists of 3,948 videos collected from TikTok through unofficial APIs, focusing on popular hashtags and keywords commonly appearing in child-oriented content. Videos were selected based on engagement metrics (views, likes, comments) while ensuring diversity in topics and posting times.

After the collection process, all videos were manually annotated into four content categories: **Safe**, **Adult Content**, **Harmful Content**, and **Suicide**. The annotation was carried out by a team of trained experts, following unified guidelines to minimize subjective bias. The inter-annotator agreement, measured by the Fleiss' Kappa coefficient, reached 81.25%, indicating high consistency and reliability in the classification process.

The dataset is split into three subsets: *training*, *development*, and *testing*. Table 1 presents the number of samples, average duration, and total video duration for each split.

The class distribution of the entire dataset is detailed in Table 2. It can be observed that videos in the **Safe** class tend to have a significantly longer

	Duration (s)			
	Samples	**Min**	**Max**	**Avg**
Train set	2,762	3.88	600	38.71
Dev. set	396	1.95	600	38.77
Test set	790	5.04	600	38.57

Table 1: Distribution of Video Samples and Duration by Data Split in the TikHarm Dataset.

average duration, reflecting the characteristics of educational or wholesome entertainment content. Conversely, the **Suicide** class shows a much shorter average duration, partly due to TikTok's strict moderation policies. With balanced class distribution and reliable annotations, TikHarm serves as a robust foundation for training and evaluating deep learning models in harmful content detection tasks. Notably, it is the first dataset in Vietnam dedicated to TikTok content safety for children, supporting research in video processing and real-time content moderation systems.

	Duration (s)			
	Samples	**Min**	**Max**	**Avg**
Safe	997	16.96	65.36	18.10
Adult Content	977	9.84	36.25	18.10
Harmful Content	990	9.88	35.92	18.10
Suicide	984	4.63	16.96	18.10

Table 2: Distribution of Video Samples and Duration by Label in the TikHarm Dataset.

3.2 Extended Dataset

One of the **main contributions** of this study is the construction and release of an Extended Dataset (available at https://www.kaggle.com/datasets/kusnguyen/extra-dataset) to enhance the coverage and representation of trending harmful content, thereby improving the detection of more subtle and sophisticated forms of harmful material on TikTok. This dataset was collected and annotated using a rigorous procedure derived from the original TikHarm dataset, ensuring consistent quality and labeling criteria, while incorporating new samples that reflect current content trends on the platform.

Dataset Collection. The supplementary data collection began by visiting TikTok's hashtag pages to identify trending hashtags. These hashtags were categorized based on their potential to yield videos falling into the four labels of the dataset. From this filtered list, we randomly collected 1,000 videos

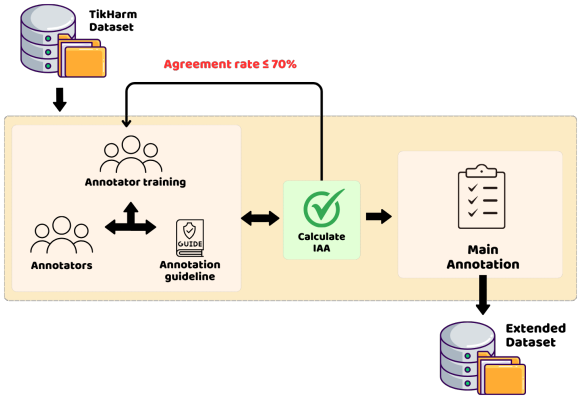


Figure 1: The dataset extending process.

using Selenium in combination with the TikTok-Content-Scraper tool. The selection was guided by engagement metrics (views, likes, comments) and ensured diversity in topics and posting times, consistent with the original TikHarm collection methodology.

Dataset Annotation. Figure 1 illustrates the overall dataset extending process. Before annotation, we refined the labeling guidelines by clearly defining each label and providing detailed illustrative examples. This ensured the extended dataset maintained high quality and alignment with the original TikHarm annotations. Three annotators, thoroughly trained on the updated guidelines, first labeled a test set of 100 samples from the original dataset. Their agreement with the ground truth achieved Cohen's Kappa scores of 0.82, 0.78, and 0.80, reflecting substantial to almost perfect agreement. Inter-annotator analysis revealed very high consensus, with two annotators achieving a Cohen's Kappa of 0.97. The overall Fleiss' Kappa across all three annotators reached 0.86, indicating almost perfect agreement. These results validated the annotators' consistent application of the labeling criteria, enabling each annotator to proceed independently with the remaining extended data.

After annotation and removal of low-quality samples, we obtained 775 qualified videos. Unlike the original TikHarm dataset, which was intentionally balanced across labels, the extended dataset has an imbalanced label distribution (Table 3). This imbalance was intentionally preserved to better reflect the real-world distribution of harmful content on TikTok, where certain categories appear more frequently than others. Such a distribution allows models to learn from "in-the-wild" data, improving their generalization in real-world scenarios. Further

details of the annotation guidelines are presented in Appendix A.2.

	Duration (s)			
	Samples	**Min**	**Max**	**Avg**
Safe	251	3.04	60	25.97
Adult Content	164	4.69	60	18.78
Harmful Content	203	5.06	60	22.61
Suicide	157	6.00	60	19.83

Table 3: Distribution of Video Samples and Duration by Label in the Extended Dataset.

3.3 Final Combined Dataset

The Extended Dataset (Section 3.2) was merged with TikHarm to form the Final Combined Dataset, which integrates both the balanced distribution of the original dataset and the realistic distribution of the extended one. This combination provides a unique benchmark that not only retains comparability with prior studies (by keeping the original TikHarm test set intact) but also challenges models with a more representative and diverse range of harmful content.

The extended samples were split into training and development sets at an 8.5:1.5 ratio, while the TikHarm test set was kept unchanged to ensure fair evaluation against existing methods. The detailed distribution of samples across the three datasets is presented in Table 4.

Dataset	**Train**	**Dev.**	**Test**	**Total**
TikHarm	2,762	396	790	3,948
Extended Dataset	656	119	-	775
Final Combined Dataset	3,418	515	790	4,723

Table 4: Sample distribution across TikHarm, Extended Dataset, and the Final Combined Dataset.

By uniting the curated balance of TikHarm with the real-world skew of the Extended Dataset, the Final Combined Dataset represents one of the most comprehensive and practical resources for harmful content detection on TikTok to date, making it a key contribution of this research.

4 Method

4.1 Data Processing

Data preprocessing is a crucial step to ensure high-quality input for the multimodal classification system, especially in the noisy and diverse context of short-form TikTok videos. We design a processing pipeline that extracts and refines linguistic information from both the audio and the visual frames of each video.

Text Extraction from Audio. The audio signal is first separated from the video using the `librosa` library, with each sample limited to a maximum duration of 60 seconds to optimize processing efficiency. After normalization, low-energy segments are removed based on the Root Mean Square (RMS) criterion. Speech recognition is performed in parallel for Vietnamese and English using the *Whisper large-v3* model (Radford et al., 2022). The hypothesis with the most complete contextual coverage is retained. To reduce hallucinated or generic outputs, we apply a filter to remove short or repetitive sentences, as well as phrases without meaningful content (e.g., "thank you", "hello", "bye"). In addition, a specialized spam filter, implemented via regular expressions, is used to remove advertising text, engagement calls, or meaningless repetitions common in social-media content.

Text Extraction from Frames. In parallel with the audio processing, text appearing in the video frames is extracted using *EASYOCR*[1], which supports Vietnamese and English. To balance recognition accuracy and computational cost, two representative frames are selected at 30% and 70% of the total video duration. Each frame is resized so that its longer side does not exceed 640 pixels. The OCR results are then processed through the same spam and language filters as the audio transcripts. Only Vietnamese or English text that is non-duplicated and passes the noise filter is retained.

Synthesis and Standardization of Text Input. The retained transcripts from both modalities (audio and OCR) are concatenated into a single string following the template: `Audio: ... | OCR: ...`. A lightweight language detector (Joulin et al., 2016) ensures that only Vietnamese or English content is preserved, while all other languages and non-linguistic text are removed. This process produces a compact, semantically rich textual input that reduces noise and improves the reliability of downstream classification.

4.2 Multimodal Harmful Content Detection Architecture

The proposed architecture (Figure 2) leverages the strengths of two pretrained encoders: *TimeSFormer*

[1]https://github.com/JaidedAI/EasyOCR

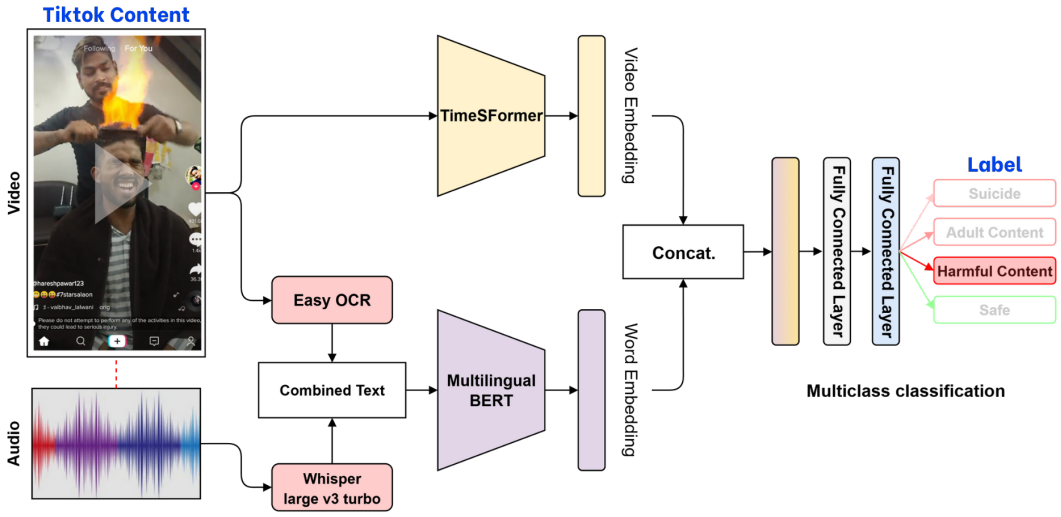


Figure 2: Proposed multimodal harmful content detection architecture.

for video representation and a transformer-based language model for text representation.

Video Encoder. We adopt *TimeSFormer*, an advanced transformer-based model that applies the Divided Space–Time Attention mechanism to separately model spatial and temporal dependencies. This design has been shown to improve classification efficiency and achieve state-of-the-art performance on large-scale benchmarks such as Kinetics-400 and Kinetics-600.

Text Encoder. Since the content of the videos primarily includes Vietnamese and English, we use *Multilingual DistilBERT* (Sanh et al., 2020) to effectively capture multilingual context. We also experiment with *ViSoBERT* (Nguyen et al., 2023), a language model optimized for Vietnamese social-media text. Each encoder output is passed through a separate projection block consisting of a linear layer, a ReLU activation, and dropout to reduce overfitting.

Feature Fusion and Classification. The projected features from the video and text encoders are concatenated and passed through a fusion network composed of multiple linear layers, Layer Normalization, ReLU activations, and dropout. The resulting joint multimodal representation is then fed into a final linear layer to perform four-way classification into *Safe*, *Adult Content*, *Harmful Content*, and *Suicide*. This late-fusion strategy preserves the strengths of each individual modality while enabling the system to exploit cross-modal cues, improving its generalization ability in multimodal harmful content detection. In addition to the late-fusion strategy, we also design an architecture with attention-based fusion. Specifically, similar to prior studies on multimodal learning, after concatenating the two modalities, the combined features are passed through Multihead Self-Attention layers, which allow the model to capture global relationships between textual and visual features, enrich the joint representation, and focus on the most relevant information.

5 Experiments

5.1 Experimental Configurations

We conducted a series of offline experiments to evaluate the effectiveness of the proposed multimodal architecture for harmful video content detection. The evaluation was performed on both the original *TikHarm* dataset and an *expanded dataset*, where the latter was constructed by augmenting *TikHarm* with additional crawled samples in a train–validation ratio of 8.5:1.5.

For the video encoder, we adopted the pretrained *TimeSFormer* model, while for the text encoder, we experimented with *Multilingual DistilBERT* and *ViSoBERT* to determine the most suitable language model for integration into the actual system.

To ensure fairness across experiments, we fixed the hyperparameters for all runs: training was conducted for 6 epochs with a batch size of 8 for both

training and evaluation. The initial learning rate was set to 1e-4, combined with a warmup ratio and weight decay, and gradients were accumulated every two steps. Losses on the training and validation sets were computed every 100 steps, and performance was evaluated using accuracy, macro-precision, macro-recall, and macro-F1. The macro formulation was chosen to mitigate class imbalance by giving equal weight to each class, including minority labels, thereby reducing bias toward dominant classes - a common challenge in harmful content classification tasks.

We retained up to three checkpoints per training run and selected the best model based on the highest macro-F1 score on the validation set. The final reported results on the test set were obtained using the checkpoint with the best macro-F1 score. The implementation was carried out in *PyTorch* and executed on the Kaggle platform with an Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100 GPU (16 GB, CUDA 11.4).

5.2 Main Results

As shown in Table 5, on the original TikHarm dataset, the results demonstrate the clear advantage of multimodal integration with attention-based fusion. On the validation set, ViSoBERT with attention-based fusion achieved the highest performance, surpassing its concat-based late-fusion variant. This indicates that when linguistic and visual features are deliberately integrated at the attention layer, the model can better exploit cross-modal interactions to reinforce classification decisions. On the TikHarm test set, Multilingual BERT achieved the highest macro-F1 score of 89.45 (Acc 89.37), followed by its attention-based variant. The inversion of rankings between the validation and test sets suggests that the cross-lingual generalization capability of Multilingual BERT is advantageous when evaluating on unseen data. In contrast, TimeSFormer-only, relying solely on visual features, performed substantially worse (F1 86.62 on validation and 86.68 on test), reinforcing the critical role of textual information in decoding the semantics of short-form videos.

On the Final Combined dataset (our extended version), validation performance dropped slightly to around 80 F1, with Multilingual BERT achieving the highest score (F1 80.77; Acc 80.78), followed by ViSoBERT (F1 80.41; Acc 80.58). This decline is expected, as the dataset was deliberately expanded to validate model robustness under real-world conditions with label imbalance and ambiguous cases. On the test set, however, models maintained strong performance close to 88% F1. Multilingual BERT with attention-based fusion achieved the best result (F1 88.44; Acc 88.48), slightly outperforming its concat late-fusion counterpart (F1 88.31; Acc 88.35) and both ViSoBERT configurations. This finding suggests that when data becomes more diverse and noisy, attention-based fusion enables the model to better capture cross-modal interactions, thereby improving the detection and classification of harmful video content in real-world scenarios.

Notably, since both the original and extended datasets naturally include noisy audio, blurred frames, and code-switched language, the strong performance across all settings indicates that MTik-Guard remains robust under such imperfect conditions. This highlights the system's ability to generalize well even in noisy, real-world TikTok scenarios, with qualitative examples provided in Appendix A.4.

5.3 Ablation Study

To better understand the contribution of each modality, we conducted an ablation study where the model was trained and evaluated using only one modality at a time: **OCR**, **ASR**, or **Video**. The results on both TikHarm and the Final Combined Dataset are presented in Table 6.

The results indicate that the **Video** modality plays the most critical role, achieving nearly 89% accuracy and F1-score on the TikHarm dataset, and maintaining high performance on the Final Combined Dataset. **ASR** contributes moderately, with F1 around 47–50%, showing that speech transcripts provide useful but less stable information due to noise, informal speech, or lack of context. **OCR** performs weakest, below 37% accuracy and around 30% F1, largely because textual content in TikTok videos is often sparse, noisy, or unrelated (e.g., ads or captions). Overall, although the Video modality dominates, combining it with ASR and OCR in a multimodal fusion setup significantly improves performance, particularly for subtle or disguised harmful content that is difficult to detect with visuals alone.

5.4 Discussion

The confusion matrix (Figure 3) reveals that the system achieved an overall accuracy of **89.37%** on the test set. The *Safe* category was most accurately pre-

Dataset	Language Model	Validation set		Test set	
		Acc.	F1	Acc.	F1
TikHarm	TimeSFormer-only	86.66	86.62	86.71	86.68
	Multilingual BERT	88.64	88.56	**89.37**	**89.45**
	Multilingual BERT [⧫]	88.13	88.14	88.35	88.38
	ViSoBERT	89.14	89.19	87.22	87.18
	ViSoBERT [⧫]	**90.15**	**90.17**	87.97	87.87
Final Combined	Multilingual BERT	**80.78**	**80.77**	88.35	88.31
	Multilingual BERT [⧫]	80.58	80.32	**88.48**	**88.44**
	ViSoBERT	80.58	80.41	87.47	87.39
	ViSoBERT [⧫]	80.39	80.22	88.10	88.03

Table 5: Performance Comparison of Multimodal Models on TikHarm and Final Combined Datasets. The table shows classification results for various language models integrated with TimeSFormer, evaluated on the Validation and Test sets. The notation [⧫] denotes models employing attention-based fusion to integrate visual and textual features. Bold values represent the best result, while underlined values signify the second-best performance within each dataset group. The TimeSFormer-only row refers to the baseline architecture utilizing only the visual encoder without any language model, with results adopted directly from (Balat et al., 2024).

Dataset	Modality			Validation set		Test set	
	Video	OCR	ASR	Acc.	F1	Acc.	F1
TikHarm	×			**88.64**	**88.64**	88.99	88.92
		×		36.11	31.33	35.44	30.27
			×	46.72	47.05	49.11	48.88
	×	×	×	**88.64**	88.56	**89.37**	**89.45**
Final Combined	×			80.58	80.44	87.97	87.90
		×		33.40	27.77	36.2	30.52
			×	44.85	44.72	50.51	50.26
	×	×	×	**80.78**	**80.77**	**88.35**	**88.31**

Table 6: Ablation Study on Modality Effectiveness Using the Best-Performing Architecture. This table shows the contribution of each modality: Video, OCR and ASR, using the best model configuration on TikHarm Dataset (last row) as the baseline. The × symbol indicates the modality was retained; removing it (empty cell) demonstrates its individual impact on performance (Acc. and F1).

dicted (183 correct classifications) with minimal mislabeling, indicating high reliability in detecting non-harmful content. *Suicide* also performed strongly (183 correct predictions) but showed some confusion with *Adult Content* and *Harmful Content*. While *Adult Content* achieved 172 correct identifications, the *Harmful Content* category (168 correct predictions) exhibited the **highest misclassification rate into *Suicide***, clearly suggesting overlapping expression patterns in sensitive themes.

We adopted a multi-class approach due to the observed content overlap and the implicit hierarchy among harmful labels (e.g., Suicide is more severe than Harmful). Therefore, the authors of the TikHarm dataset and we prioritized the maximum risk in the labeling process, assigning a single, definitive label to each piece of content. This choice simplifies the decision-making process into a clear, automated action (BLOCK/ALLOW), avoiding the complexity of multi-label output. To manage cases where the model has low confidence (is "confused"), we will utilize Label Smoothing during training and set a confidence threshold to flag ambiguous cases for human review, thus effectively managing the overlap without altering the core label structure.

Besides the intrinsic issue of label overlap, data imbalance is also a challenging factor, particularly affecting the minority class, *Suicide*. We observed higher misclassification between *Suicide* and *Harmful/Adult content*. Nevertheless, we retained this imbalance to reflect real-world TikTok distributions, where suicide-related videos are rarer and often subtle. This improves generalization in deployment, though future work may adopt reweighting strategies or focal loss to further boost minority-

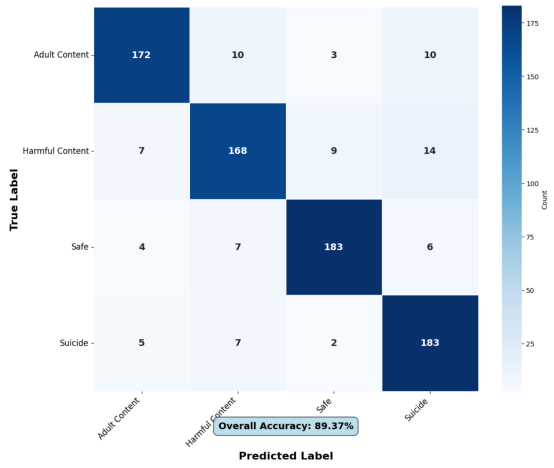


Figure 3: Confusion matrix of the best-performing model on the test set.

class recall.

TikTok videos often contain noisy audio, blurred frames, and code-switched language, which naturally introduce errors in OCR and ASR. Despite these challenges, MTikGuard demonstrates strong robustness thanks to multimodal fusion, where errors in one modality can be compensated by signals from another. For instance, background noise may reduce ASR quality, but OCR-extracted harmful text can still support correct classification, and vice versa. Ablation results (Table 6) confirm this effect: unimodal OCR or ASR alone performs poorly, while multimodal fusion consistently improves accuracy by about 2–3%. Moreover, Multilingual DistilBERT effectively handles code-switching (e.g., Vietnamese speech with English captions). Together, these findings highlight the system's reliability in noisy, real-world TikTok scenarios.

Overall, the model demonstrated robust discriminative ability across all categories despite the diverse and unpredictable nature of TikTok short videos.

6 MTikGuard System

Figure 4 illustrates the overall architecture of **MTikGuard**, a modular and extensible system designed for real-time multimodal content moderation on TikTok. The design ensures adaptability to various deployment environments and scalability to handle high-throughput streaming data. The system is structured into three primary functional layers, presented from bottom to top as follows:

The foundation layer focuses on dataset preparation and model training. The system leverages the TikHarm dataset, augmented with an extended dataset to enhance coverage and robustness. For each video, visual features are extracted using the TimeSFormer video transformer, while textual features are derived from transcribed audio via Whisper and scene text via EasyOCR. Text representations are encoded using either Multilingual DistilBERT or ViSoBERT, as described in previous sections. The resulting multimodal architecture is trained and evaluated using macro-F1 as the primary selection metric. Once the optimal model is obtained, it is serialized and integrated into the real-time censorship pipeline.

The middle layer implements the automated streaming pipeline, orchestrated with Apache Airflow[2] and containerized using Docker[3] for portability and reproducibility. The pipeline includes:

- A Selenium-based crawler[4] together with the TikTok Content Scraper module retrieves trending TikTok videos based on hashtags, extracts metadata and video IDs, and publishes them to an Apache Kafka[5] topic .
- Kafka consumers download the videos, extract text from audio using Whispe and from frames using EasyOCR, and then classify the content using the trained multimodal model. The classification results, along with metadata, are stored in MongoDB Atlas [6] for further querying and analysis.
- Airflow coordinates the execution of producers and consumers, monitors pipeline health, and manages periodic reporting, ensuring stability and scalability.

The top layer emphasizes the modular nature of MTikGuard, allowing flexible deployment in both cloud-based and on-premise environments. Each component - such as Kafka consumers, inference services, or data storage - can be scaled independently. The architecture also facilitates integration with additional modalities or extension to other short-video platforms in future work.

7 Conclusion

We introduced **MTikGuard**, a real-time multimodal detection system for harmful TikTok con-

[2]https://airflow.apache.org
[3]https://www.docker.com
[4]https://www.selenium.dev
[5]https://kafka.apache.org
[6]https://www.mongodb.com

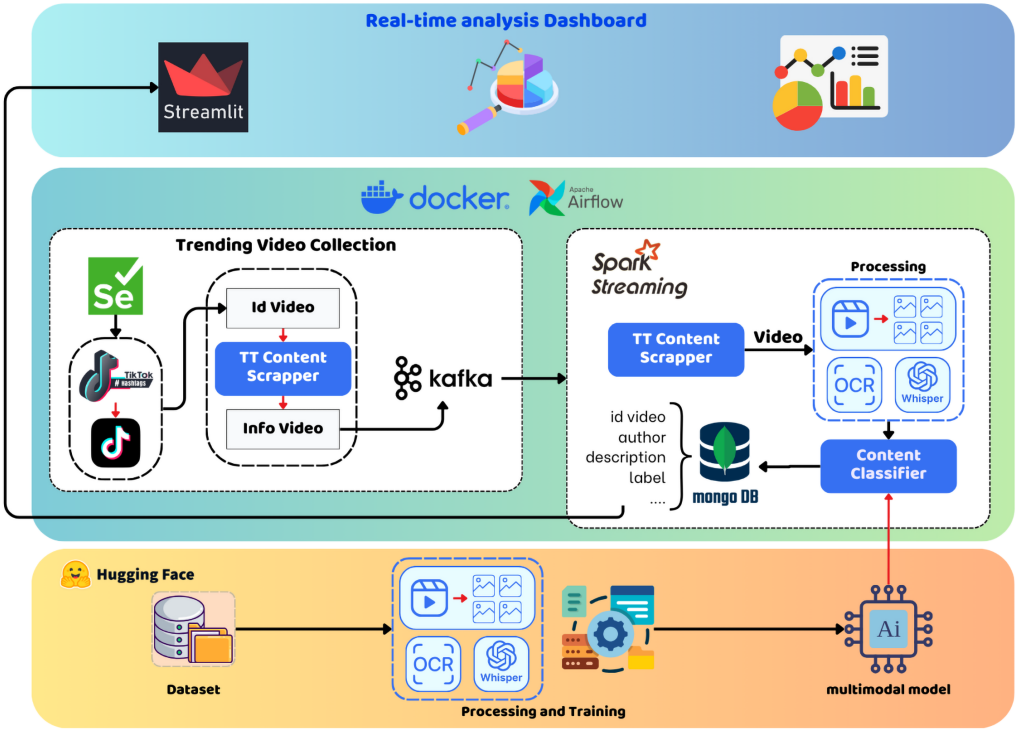


Figure 4: Overview of the MTikGuard system.

tent that jointly analyzes visual, audio, and textual modalities. By extending the TikHarm dataset to 4,723 annotated videos and leveraging transformer-based encoders (TimeSFormer and DistilBERT) within a scalable streaming pipeline, MTikGuard achieves **state-of-the-art performance** with 89.37% accuracy and 89.45% macro-F1. These results highlight the effectiveness of combining dataset expansion, multimodal fusion, and practical deployment infrastructure in safeguarding young users from harmful online content. In addition, our pipeline is inherently multilingual: Whisper and EasyOCR support dozens of languages, while Multilingual DistilBERT enables robust cross-lingual text encoding. Thus, extending to other regions would require minimal adjustments, primarily re-training with a small annotated subset to adapt to local linguistic nuances.

While the current system adopts a late-fusion strategy, it does not yet explore more advanced mechanisms to dynamically weight each modality's contribution. In future work, we plan to enhance audio processing to better handle non-speech cues and incorporate attention-driven multimodal fusion to improve robustness and interpretability in complex real-world scenarios.

Acknowledgments

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D4-2025-14.

References

Mazen Balat, Mahmoud Gabr, Hend Bakr, and Ahmed B. Zaky. 2024. Tikguard: A deep learning transformer-based solution for detecting unsuitable tiktok content for kids. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 337–340.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *Preprint*, arXiv:2102.05095.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. *Preprint*, arXiv:2305.03915.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *Preprint*, arXiv:1607.01759.

Dhiraj Murthy, Rachel Ouellette, Tanvi Anand, Srijith Radhakrishnan, Nikhil Mohan, Juhan Lee, and Grace Kong. 2024. Using computer vision to detect e-cigarette content in tiktok videos. *Nicotine and Tobacco Research*, 26:S36–S42.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *Preprint*, arXiv:2310.11166.

Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Girish, Swarup Ranjan Behera, Abu Osama Siddiqui, Sarthak Jain, Priyabrata Mallick, Jaya Sai Kiran Patibandla, Pailla Balakrishna Reddy, Arun Balaji Buduru, and Rajesh Sharma. 2025. Snifr : Boosting fine-grained child harmful content detection through audio-visual alignment with cascaded cross-transformer. *Preprint*, arXiv:2506.03378.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. Kidsguard: fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 2104–2111, New York, NY, USA. Association for Computing Machinery.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298.

A Appendix

A.1 Task Definition

The harmful content detection task in TikTok videos is formulated as a supervised **multimodal classification** problem. Given a video V, the system processes three modalities: (1) *visual frames* F_v, (2) *audio transcripts* T_a (from speech-to-text), and (3) *scene text* T_o (from OCR on frames). After noise removal and normalization, each modality is encoded into feature vectors: x_v for visual data and x_t for text (combining T_a and T_o). These are fused using a late fusion strategy:

$$x_m = \text{Fusion}(x_v, x_t) \tag{1}$$

The classifier f_θ maps x_m to one of four labels:

$$y \in \{\text{Safe},\ \text{Adult Content},\ \text{Harmful Content},\ \text{Suicide}\} \tag{2}$$

The goal is to minimize cross-entropy loss on the labeled dataset $\mathcal{D}$ while maximizing the **macro-F1 score** to ensure balanced performance across classes.

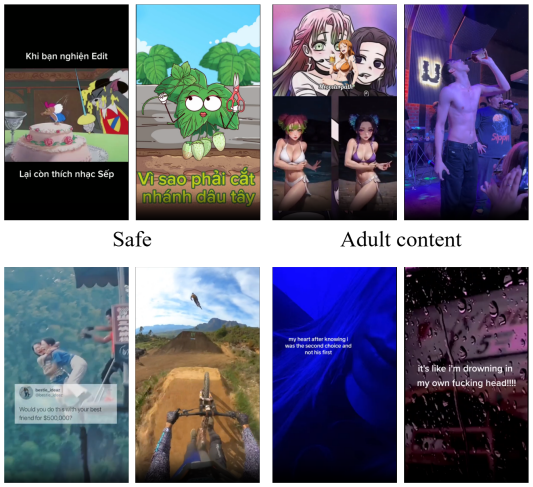


Figure 5: An example of the Harmful Content Detection task.

A.2 Annotation Guidelines

A.2.1 Annotation Tool

To support this complex annotation task, we developed a dedicated web-based annotation tool (available at `https://github.com/Kusssssss/tiktok-labeling-tool.git`). The system displays a list of videos and associated metadata (user name and video ID), a video player for observing content directly, and a label selection panel. The

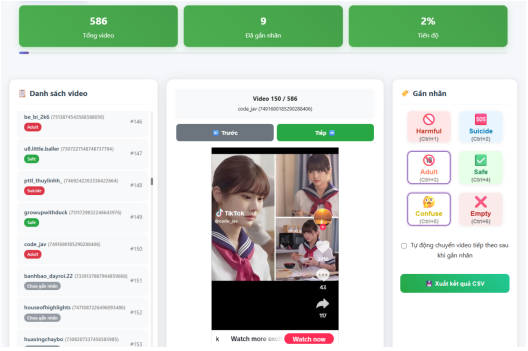


Figure 6: The web application interface for annotators.

interface also provides progress tracking, together with two auxiliary options, *confuse* and *empty*, to handle ambiguous cases and ensure labeling quality and consistency (Figure 6).

A.2.2 Annotation Guidelines

Each annotator must strictly follow the label definitions derived from the TikHarm dataset. Only one label can be assigned to each video; multiple labels are considered invalid. Samples that are unclear are first labeled based on the annotator's best judgment and later cross-validated with others.

- **Safe:** Content that is appropriate and safe for children, without violence, sexuality, strong language, or dangerous behaviour. *(e.g., educational programs, family-friendly vlogs, cartoons, DIY activities).*
- **Adult Content:** Content containing sexual elements, nudity, or language not suitable for children, including sexually suggestive material. *(e.g., provocative dancing, lingerie try-ons, sexual jokes or innuendo).*
- **Harmful Content:** Content depicting violence, dangerous acts, or harmful behaviour that may negatively influence children. *(e.g., fights, bullying, dangerous challenges involving weapons or chemicals).*
- **Suicide:** Content referring to suicidal behaviour, intentions, or methods that may negatively impact viewers, especially minors. *(e.g., self-harm attempts, suicide tutorials, confessions of suicidal thoughts).*

A.3 Error Analysis Details

We observed two main error types. The first is misclassification, where the predicted label differs from the ground truth (Figure 7). For example,

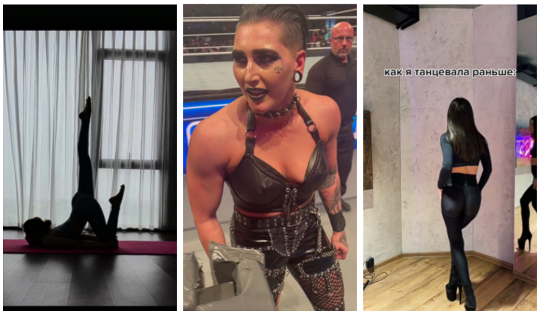


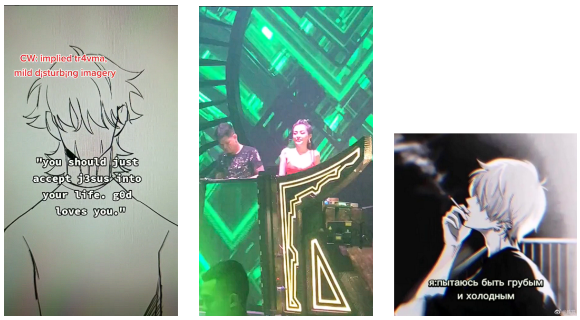
Figure 7: Examples of misclassification cases where predicted labels differ from ground truth. From left to right, Safe as Harmful Content, Harmful Content as Adult Content and Adult Content as Harmful Content.

Figure 8: Examples of overlapping-label cases where videos could plausibly belong to multiple categories. From left to right, Harmful Content also Adult Content, Suicide also Adult Content and Adult Content also Harmful Content.

Safe videos predicted as *Harmful Content* due to poses resembling unsafe contexts, *Harmful Content* misclassified as *Adult Content* because of clothing or makeup cues, and *Adult Content* predicted as *Harmful Content* due to aggressive thematic elements. In some cases, these predictions may still be justifiable under stricter moderation perspectives. The second type is overlapping-label cases, where videos plausibly belong to more than one harmful category (Figure 8), such as content mixing aggressive and sexual elements, suicide-related context with adult cues, or mature visuals with harmful implications. These findings suggest that future work should explore multi-label classification and improved multimodal fusion to address ambiguous cases.

A.4 Noisy Cases

As shown in Figure 9, the case (a) combines multiple types of noise: teencode in the opening frames, numeric overlays in later segments, loud rock-style background audio, and flickering visuals. The case (b) suffers from strong lighting distortions



(a) (b) (c)

Figure 9: Examples of noisy cases. From left to right: (a) stylized and numeric text with blurred, flickering frames, (b) nightclub environment with light and audio noise, (c) code-switched content with unrelated audio.

and audio that lacks linguistic content, while (c) mixes cross-lingual subtitles with English audio unrelated to the annotated label. Despite these imperfections, MTikGuard successfully classified all cases into their correct categories - (a) *Suicide*, (b) *Adult Content*, (c) *Suicide*, confirming that the multimodal fusion design provides robustness even under noisy and ambiguous conditions typical of TikTok videos.

Personalized Graph-Based Retrieval for Large Language Models

Steven Au[1], Cameron J. Dimacali[1], Ojasmitha Pedirappagari[1], Namyong Park, Franck Dernoncourt[2], Yu Wang[3], Nikos Kanakaris[4]*, Hanieh Deilamsalehy[2] Ryan A. Rossi[2], Nesreen K. Ahmed[5]

[1]University of California Santa Cruz, [2]Adobe Research,
[3]University of Oregon, [4]Amazon Web Services, [5]Cisco AI Research

https://pgraphrag-benchmark.github.io

Abstract

As large language models (LLMs) continue to evolve, their ability to deliver personalized, context-aware responses holds significant promise for enhancing user experiences. However, most existing personalization approaches rely solely on user history, limiting their effectiveness in cold-start and sparse-data scenarios. We introduce Personalized Graph-based Retrieval-Augmented Generation (PGraphRAG), a framework that enhances personalization by leveraging user-centric knowledge graphs. By integrating structured user information into the retrieval process and augmenting prompts with graph-based context, PGraphRAG improves both relevance and generation quality. We also present the Personalized Graph-based Benchmark for Text Generation, designed to evaluate personalized generation in real-world settings where user history is minimal. Experimental results show that PGraphRAG consistently outperforms state-of-the-art methods across diverse tasks, achieving average ROUGE-1 gains of 14.8% on long-text and 4.6% on short-text generation—highlighting the unique advantages of graph-based retrieval for personalization.

1 Introduction

The rapid advancement of large language models (LLMs) has enabled a wide range of NLP applications, including conversational agents, content generation, and code synthesis. Models like GPT-4 (OpenAI, 2024) now power virtual assistants capable of answering complex queries and engaging in multi-turn dialogue (Brown et al., 2020). As these models continue to evolve, their ability to generate personalized, context-aware responses offers new opportunities to enhance user experiences (Salemi et al., 2024b; Huang et al., 2022). Personalization enables LLMs to adapt outputs to individual preferences and goals, resulting in richer, more relevant interactions (Zhang et al., 2024). While personalization has been studied in areas such as information retrieval and recommender systems (Xue et al., 2009; Naumov et al., 2019), its integration into LLMs for generation tasks remains relatively underexplored.

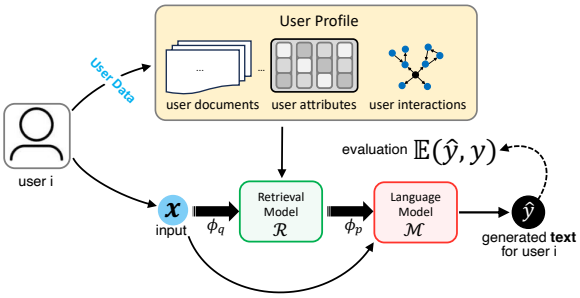


Figure 1: Overview of the proposed PGraphRAG framework. We construct user-centric graphs from user profile and interaction data, then retrieve structured, user-relevant information from the graph. This context is used to condition the language model's generation, producing personalized outputs for user i.

One of the key challenges in advancing personalized LLMs is the lack of benchmarks that adequately capture the complexities of personalization tasks. Popular natural language processing (NLP) benchmarks (e.g., (Wang et al., 2019b), (Wang et al., 2019a), (Gehrmann et al., 2021)) primarily focus on general language understanding and generation, with limited emphasis on personalization. As a result, researchers and practitioners lack standardized datasets and evaluation metrics for developing and assessing models designed for personalized text generation. Recently, efforts such as LaMP (Salemi et al., 2024b) and LongLaMP (Kumar et al., 2024) have begun addressing this gap. LaMP evaluates personalization for tasks like email subject and news headline generation, while LongLaMP extends this to long-text tasks such as email and abstract generation. However, both benchmarks rely exclusively on user his-

*The work does not relate to the author's position at Amazon.

tory to model personalization. Here, user history typically refers to a set of previously written texts by the same user—such as past reviews, messages, or profile-specific documents—which are used as context to condition the generation.

Challenges with Cold-Start Users. While leveraging user history is valuable for capturing individual style and preferences, it presents a cold-start challenge: many users have little or no prior data. In fact, as shown in Figure 2, over 99.99% of users in the Amazon Reviews dataset have fewer than three interactions. Benchmarks like LaMP and LongLaMP filter out these users by imposing a minimum user profile size threshold to ensure sufficient data for personalization. As a result, they exclude the vast majority of users, making their evaluations less representative of real-world deployment. This design choice leads to model failures when prompts lack sufficient context, often resulting in generic outputs.

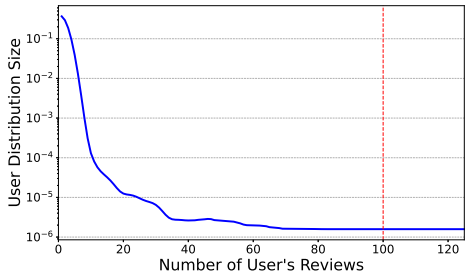


Figure 2: Distribution of user profile sizes in the Amazon user-product dataset. The vast majority of users have only a few reviews, highlighting the prevalence of sparse profiles. The red vertical line indicates the minimum profile size threshold used in prior benchmarks such as LaMP and LongLaMP.

Proposed Approach. To address these challenges, we propose *Personalized Graph-based Retrieval-Augmented Generation* (PGraphRAG), a novel framework that enhances personalized text generation by leveraging user-centric knowledge graphs. These structured graphs represent user information — such as interests, preferences, and prior interactions — in an interconnected graph structure. During inference, PGraphRAG retrieves semantically relevant context from both the user's own profile and neighboring profiles extracted from the graph, and augments the prompt with this information to guide generation. This graph-based approach enables the model to produce contextually appropriate and personalized outputs, even when user history is sparse or unavailable (see Figure 1).

Formally, the target task of PGraphRAG is personalized text generation conditioned on user-specific context retrieved from a structured knowledge graph. Given a user query (e.g., a product title or review prompt), the system retrieves relevant entries from the graph-based profile and generates an output tailored to the user's preferences. This setup generalizes personalization beyond pure user text history, enabling context-rich generation even in sparse or cold-start settings.

Proposed Benchmark. To evaluate our approach, we introduce the *Personalized Graph-based Benchmark for Text Generation*, a novel evaluation benchmark designed to fine-tune and assess LLMs on twelve personalized text generation tasks, including long- and short-form generation as well as classification. This benchmark addresses the limitations of existing personalized LLM benchmarks by providing datasets that specifically target personalization capabilities in real-world settings where user history is sparse. In addition, it enables a more comprehensive assessment of a model's ability to personalize outputs based on structured user information.

Our benchmark supports evaluation in sparse-profile settings, and PGraphRAG is designed to retrieve semantically relevant context not only from the user's own profile but also from neighboring profiles extracted from the graph — enabling effective personalization even when the user has only a single input (e.g., one review in their profile). Empirically, PGraphRAG significantly outperforms LaMP in these low-profile scenarios, demonstrating the advantages of graph-based reasoning over strict reliance on user history.

Our contributions are summarized as follows:

1. **Benchmark.** We introduce the *Personalized Graph-based Benchmark for Text Generation*, consisting of 12 tasks spanning long-form generation, summarization, and classification. To support further research, we release the benchmark publicly. [1]
2. **Method.** We propose *PGraphRAG*, a retrieval-augmented generation framework that addresses the cold-start problem by augmenting generation with structured, user-specific information from a knowledge graph.
3. **Effectiveness.** We show that PGraphRAG achieves state-of-the-art performance across all tasks in our benchmark, demonstrating the value of graph-based reasoning for personal-

[1] https://pgraphrag-benchmark.github.io

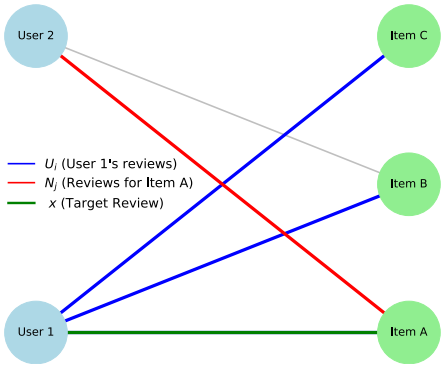


Figure 3: Example of a bipartite user-centric graph $G = (U, V, E)$ showing users, items, and interaction edges (e.g., reviews).

ized text generation.

2 Personalized Graph-based Benchmark for LLMs

We introduce the *Personalized Graph-Based Benchmark* to evaluate LLMs on their ability to generate personalized outputs across twelve tasks, spanning long-form generation, short-form generation, and ordinal classification. The benchmark is constructed from real-world datasets across multiple domains.

2.1 Personalized Text Generation: Problem Definition

Each benchmark instance includes: (1) an input sequence x to the LLM, (2) a target output y the model is expected to generate, and (3) a user profile P_i derived from a structured user-centric graph. Given an input-output pair (x, y) associated with user i, the goal is to generate a personalized output $\hat{y}$ that aligns with the semantics and style of y, conditioned on the user profile P_i.

We assume user context is represented using a bipartite user-centric graph that captures user-item interactions (see Figure 3 for an illustration). The profile P_i is constructed from this graph and includes both interactions authored by the user and related signals from similar items or neighboring users. The full construction of P_i is detailed in Section 3.

Formally, the personalized generation task is defined as:

$$\hat{y} = \arg\max_{y'} \Pr(y' \mid x, P_i) \tag{1}$$

where x is the input query, y is the target output, and P_i denotes the profile of user i derived from a user-item interaction graph. The model generates an output $\hat{y}$ that maximizes the likelihood of personalized text conditioned on the input and user profile. This formulation enables generalization beyond user history by leveraging structured, graph-derived context.

In practice, our framework retrieves a personalized context $\mathcal{R}(P_i) \subseteq P_i$ from the graph to condition generation, yielding the operational objective:

$$\hat{y} = \arg\max_{y'} \Pr(y' \mid x, \mathcal{R}(P_i)) \tag{2}$$

where $\mathcal{R}(P_i)$ represents the retrieved subset of user- and item-level interactions used as context during generation.

Finally, statistics for all benchmark tasks and their associated graphs are summarized in Table 1 and Table 2. Additional dataset split details are provided in the appendix.

2.2 Task Definitions

Task 1: User Product Review Generation. Personalized review text generation has progressed from incorporating user-specific context to utilizing LLMs for producing fluent and contextually relevant reviews and titles (Ni and McAuley, 2018). This task aims to generate a product review i_{text} for a target user, conditioned on their own review title i_{title} and a set of additional reviews P_i from their user profile. We construct this dataset from the Amazon Reviews 2023 corpus (Hou et al., 2024), spanning multiple product categories and used to define a bipartite user-item graph.

Task 2: Hotel Experience Generation. Hotel reviews often contain rich narratives reflecting personal experiences, making personalization essential to capturing individual preferences and expectations (Kanouchi et al., 2020). This task focuses on generating a personalized hotel experience story i_{text}, using the target user's review summary i_{title} and contextual reviews P_i. We use the Hotel Reviews dataset, a subset of Datafiniti's Business Database (Datafiniti, 2017), to construct a user-hotel bipartite graph.

Task 3: Stylized Feedback Generation. Writing style — influenced by grammar, punctuation, and expression — is deeply personal and often shaped by geographic and cultural factors (Alhafni et al., 2024). This task involves generating personalized

Task	Type	Avg. Input Length	Avg. Output Length	Avg. Profile Size	# Classes
User-Product Review Generation	Long Text Generation	3.754 ± 2.71	47.90 ± 19.28	1.05 ± 0.31	-
Hotel Experiences Generation	Long Text Generation	4.29 ± 2.57	76.26 ± 22.39	1.14 ± 0.61	-
Stylized Feedback Generation	Long Text Generation	3.35 ± 2.02	51.80 ± 20.07	1.09 ± 0.47	-
Multilingual Product Review Generation	Long Text Generation	2.9 ± 2.40	34.52 ± 12.55	1.08 ± 0.33	-
User-Product Review Title Generation	Short Text Generation	30.34 ± 37.95	7.02 ± 1.14	1.05 ± 0.31	-
Hotel Experiences Summary Generation	Short Text Generation	90.40 ± 99.17	7.64 ± 0.92	1.14 ± 0.61	-
Stylized Feedback Title Generation	Short Text Generation	37.42 ± 38.17	7.16 ± 1.11	1.09 ± 0.47	-
Multilingual Product Review Title Generation	Short Text Generation	22.17 ± 20.15	7.15 ± 1.09	1.08 ± 0.33	-
User-Product Review Ratings	Ordinal Classification	34.10 ± 38.66	-	1.05 ± 0.31	5
Hotel Experiences Ratings	Ordinal Classification	94.69 ± 99.62	-	1.14 ± 0.61	5
Stylized Feedback Ratings	Ordinal Classification	40.77 ± 38.69	-	1.09 ± 0.47	5
Multilingual Product Ratings	Ordinal Classification	25.15 ± 20.75	-	1.08 ± 0.33	5

Table 1: Data statistics for the PGraphRAG Benchmark across the four datasets. For each task, we report the average input and output lengths (in words), measured on the test set using BM25-based retrieval with GPT. The average profile size indicates the number of reviews per user used for personalization.

Dataset	Users	Items	Edges/Reviews	Average Degree
User-Product Review Graph	184,771	51,376	198,668	1.68
Hotel Experiences Graph	15,587	2,975	19,698	2.12
Stylized Feedback Graph	58,087	600	71,041	2.42
Multilingual Product Review Graph	112,993	55,930	131,075	1.55

Table 2: Graph statistics for the datasets used in the personalized tasks. Each row reports the number of users, items, and edges (i.e., reviews), as well as the average degree of the resulting user-centric bipartite graph. The four graphs correspond to: User-Product, Multilingual Product, Stylized Feedback, and Hotel Experiences.

product feedback i_{text}, based on the user's feedback title i_{title} and additional feedback samples P_i from their profile. We utilize the Grammar and Online Product dataset, a subset of the Datafiniti Business corpus (Datafiniti, 2018), which reflects stylistic variation across multiple platforms and domains.

Task 4: Multi-lingual Review Generation. Personalization in multilingual review generation presents unique challenges due to differences in linguistic structures, cultural norms, and stylistic conventions (Cortes et al., 2024). This task focuses on generating product reviews i_{text} in Brazilian Portuguese, using the target user's review title i_{title} and additional reviews P_i from their profile. We construct this dataset using B2W-Reviews (Real et al., 2019), sourced from Brazil's largest e-commerce platform.

Task 5: User Product Review Title Generation. Short text generation for personalized review titles is particularly challenging, requiring the model to summarize sentiment and reflect user-specific phrasing preferences. This task generates a review title i_{title} for a given user, using their review text i_{text} and additional profile reviews P_i, without relying on parametric user embeddings (Xu et al., 2023). The dataset is derived from Amazon Reviews (Hou et al., 2024).

Task 6: Hotel Experience Summary Generation. Helping users write summaries of hotel experiences requires distilling detailed narratives into concise summaries that reflect individual preferences (Kamath et al., 2024). This task generates a hotel experience summary i_{title} based on the user's full experience text i_{text} and additional hotel reviews P_i. We use the Hotel Reviews dataset from the Datafiniti Business Database (Datafiniti, 2017).

Task 7: Stylized Feedback Title Generation. Stylized feedback summarization aims to capture individual voice and tone in generating short-form feedback. This task benchmarks stylized opinion generation across domains such as music, groceries, and household items (Iso et al., 2024). The model generates the target user's feedback title i_{title} based on their full feedback text i_{text} and additional feedback P_i from similar users. The dataset is built from the Datafiniti Products dataset (Datafiniti, 2018).

Task 8: Multi-lingual Review Title Generation. Multilingual short-text personalization adds further complexity, particularly in Brazilian Portuguese, where style and syntax vary significantly across users (Scalercio et al., 2024). This task gen-

erates a personalized review title i_{title} using the user's full review text i_{text} and contextual examples P_i from their graph neighborhood. Data: B2W-Reviews (Real et al., 2019).

Task 9: User Product Review Ratings. Predicting personalized product ratings involves understanding sentiment, user bias, and historical feedback. This task formulates rating prediction as an ordinal classification problem, where the model predicts $i_{\text{rating}} \in \{1, 2, 3, 4, 5\}$ based on the user's review text i_{text}, title i_{title}, and additional profile context P_i. The dataset is constructed from Amazon Reviews (Hou et al., 2024).

Task 10: Hotel Experience Ratings. Hotel ratings often reflect nuanced factors such as location, cleanliness, and service. This task models hotel experience rating i_{rating} prediction as a classification problem based on the user's review story i_{text}, summary i_{title}, and surrounding review context P_i. Data: Datafiniti Hotel Reviews (Datafiniti, 2017).

Task 11: Stylized Feedback Ratings. Cross-domain sentiment prediction explores how writing quality and sentiment expression vary across platforms (Yu et al., 2021). This task assigns a numerical feedback rating i_{rating} to a stylized user review using the input review text i_{text}, review title i_{title}, and personalized context P_i. The dataset is taken from the Datafiniti Product Database on Grammar and Online Product Reviews (Datafiniti, 2018).

Task 12: Multi-lingual Product Ratings. While sentence-level sentiment classification in Portuguese has seen success (de Araujo et al., 2024), this task extends to full review-level sentiment modeling in a multilingual setting. The model predicts a Portuguese user-product rating i_{rating} using both the review text i_{text}, the title i_{title}, and additional user-item interactions P_i. We construct this dataset using B2W-Reviews (Real et al., 2019).

3 The PGraphRAG Framework

Personalizing LLMs in real-world settings requires addressing two key challenges: (1) user profiles are often sparse or unavailable, and (2) incorporating additional user-related context must remain relevant, efficient, and scalable. To tackle these issues, PGraphRAG leverages structured user-centric knowledge graphs for context construction, and combines this with retrieval-augmented prompting. This design enables the model to generalize beyond parametric user embeddings or history-based filtering by dynamically retrieving relevant signals from graph-based user profiles that extend beyond the user's direct history.

Here, we present *PGraphRAG*, our proposed framework for personalizing large language models (LLMs) through graph-based retrieval augmentation. PGraphRAG enhances generation by conditioning a shared LLM on structured, user-specific context extracted from a user-centric knowledge graph. This enables tailored and context-aware outputs, especially in sparse or cold-start scenarios.

PGraphRAG leverages a bipartite user-centric graph $G = (U, V, E)$ to incorporate contextual signals beyond direct user history. We represent user context as a bipartite graph, where U is the set of user nodes, V the set of item nodes, and E the set of interaction edges (see Figure 3 for an illustration). An edge $(i, j) \in E$ corresponds to an interaction between user i and item j, such as a review that includes metadata like text, title, and rating. The user profile P_i consists of the set of reviews written by user i, along with reviews for the same items j written by other users $k \neq i$. For a given user $i \in U$, we define the profile P_i as the union of:

- the set of interactions authored by user i: $\{(i, j) \in E\}$,
- the set of interactions for the same items j written by other users $k \neq i$: $\{(k, j) \in E \mid (i, j) \in E\}$.

$$P_i = \{(i, j) \in E\} \cup \{(k, j) \in E \mid (i, j) \in E\} \quad \forall j \in V,\ k \in U,\ k \neq i \tag{3}$$

Due to context window limitations and efficiency considerations, we apply retrieval augmentation to select only the most relevant entries from P_i for conditioning the model. Given an input sample (x, y) for user i, the PGraphRAG workflow proceeds in three steps: a query function, a graph-based retrieval module, and a prompt construction function, as illustrated in Figure 1:

1. **Query Function** (ϕ_q): The query function transforms the input x into a query q for retrieval.
2. **Graph-Based Retrieval** ($\mathcal{R}$): The retrieval function $\mathcal{R}(q, G, k)$ takes as input the query q, the bipartite graph G, and a threshold k.

It first constructs the user profile P_i from G as defined above, and then retrieves the top-k most relevant entries from the user profile P_i with respect to q.

3. **Prompt Construction** (ϕ_p): The prompt construction assembles a personalized prompt for user i by combining the input x with the retrieved entries.

The final input to the LLM is a personalized, context-augmented prompt $\tilde{x}$ defined as:

$$\tilde{x} = \phi_p(x, \mathcal{R}(\phi_q(x), G, k)) \tag{4}$$

The pair $(\tilde{x}, y)$ is then used for inference or fine-tuning. This modular pipeline enables efficient, graph-aware personalization across diverse tasks and user sparsity levels.

Modularity and Extensibility. While we define P_i as a hybrid of user-authored and neighbor-authored interactions, PGraphRAG is modular by design. The underlying graph can be leveraged in alternative ways depending on the application: for example, practitioners may define P_i using only user-specific data, only neighbor interactions, or other graph-based traversal strategies (e.g., multi-hop reasoning or community-based filtering). Each component of the framework—query formulation, retrieval logic, and prompt construction—can be adapted independently, making PGraphRAG extensible to a wide range of personalized retrieval scenarios. In addition, the retrieval module supports plug-and-play compatibility with a variety of retrievers, such as BM25, or Contriever, allowing flexibility in balancing speed, semantic relevance, and computational cost.

4 Experiments

Setup. We evaluate our methods using two LLM backbones. The first is the LLaMA 3.1 8B Instruct model (Touvron et al., 2023), implemented with the Huggingface `transformers` library and configured to generate up to 512 tokens. The second is the GPT-4o-mini model (OpenAI, 2024), accessed via the Azure OpenAI Service (Services, 2023) using the `AzureOpenAI` interface, with a decoding temperature of 0.4. All experiments are conducted on an NVIDIA A100 GPU with 80GB of memory.

Dataset Splits and Graph Construction We construct bipartite user-entity graphs and split users into training, development, and test sets while preserving connectivity. Full details on data construction, neighbor filtering, and stratification are provided in Appendix A.

Graph Construction. We construct a bipartite user-entity graph from the selected user profiles in the validation and test splits. Each user node is connected to entity nodes (e.g., products, hotels, feedback targets) based on authored content, with edges representing user interactions such as reviews, summaries, or ratings. This graph supports two retrieval configurations: (1) *user-only*, which retrieves content authored solely by the target user (i.e., from their personal profile), and (2) *user+neighbor*, which additionally includes content from neighboring users who have interacted with the shared target entity. In both modes, the retrieved content defines the personalized context passed to the language model.

Ranking and Retrieval. The query used for retrieval varies by task type: for *Long Text Generation*, we use the review title; for *Short Text Generation*, the review text; and for *Ordinal Classification*, a combination of title and text. We apply two retrieval models—BM25 (Robertson and Zaragoza, 2009) and Contriever (Lei et al., 2023) to select the top-k most relevant entries from either the user-only or user+neighbor profiles. To enforce consistency between users with high activity and cold-start users, we cap retrieval at k, even if more candidate entries are available (see Table 7 and Figure 2). All textual inputs are tokenized using NLTK's `word_tokenize`. We use the default settings for both retrieval models; for Contriever, mean pooling is applied over token embeddings.

LLM Prompt Generation. Once the top-k entries are retrieved, we construct a *template-based prompt* that includes both the user's query (e.g., a request for a full review, a title, or a rating) and the contextual information from the graph. This prompt is passed to the LLM for generation. An illustration of task-specific prompt formatting is shown in Figure 4.

Baseline Methods. We compare PGraphRAG against both non-personalized and personalized baselines. (1) *No-Retrieval* constructs the prompt without any retrieval augmentation; the LLM generates the output solely from the query. (2) *Random-Retrieval* augments the prompt with content randomly sampled from all user profiles, introducing unrelated context. (3) *LaMP* (Salemi et al., 2024b) is a personalized baseline that augments the prompt

using content from the target user's own history (e.g., previously written reviews).

Evaluation. We evaluate each method by providing task-specific inputs and comparing generated outputs against reference labels. For generation tasks (long and short text), we report ROUGE-1, ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) scores. For rating prediction tasks, we measure mean absolute error (MAE) and root mean squared error (RMSE).

4.1 Baseline Comparison

We compare PGraphRAG against baselines on the three task types in our benchmark — long-text generation, short-text generation, and rating prediction.

Long Text Generation. Tables 3 and 16 show that PGraphRAG consistently outperforms all baseline methods—including No-Retrieval, Random-Retrieval, and LaMP—across ROUGE-1, ROUGE-L, and METEOR metrics. The largest performance gains are observed in Task Hotel Experience Generation, where PGraphRAG achieves +32.1% in ROUGE-1, +21.7% in ROUGE-L, and +25.7% in METEOR over the LaMP baseline using the LLaMA-3.1-8B-Instruct model. These improvements highlight the benefits of incorporating structured, graph-based context beyond user history.

Short Text Generation. Tables 4 and 17 show that PGraphRAG outperforms the baselines in most cases. In Task User Product Review Title Generation, PGraphRAG achieves consistent gains over LaMP in the LLaMA-3.1-8B-Instruct model: ROUGE-1 (+5.6%), ROUGE-L (+5.9%), and METEOR (+6.8%). These improvements, while smaller than those in long-form tasks, reflect the limited headroom for personalization in very short text generation tasks such as review title. Because the target texts are extremely brief, minor lexical differences can significantly affect overlap-based metrics, and there are fewer opportunities for retrieved context to meaningfully influence generation.

Ordinal Classification. Tables 8 and 18 show that PGraphRAG yields modest improvements over LaMP in rating prediction tasks. It outperforms LaMP in 1 out of 4 tasks with LLaMA-3.1-8B-Instruct and in 2 out of 4 tasks with GPT. The largest gains are observed on the Multilingual Product Ratings task, with improvements in MAE (+1.75%) and RMSE (+1.12%) for LLaMA-3.1-8B-Instruct , and MAE (+2.16%) and RMSE (+3.17%) for GPT. These gains, while small, suggest that user profiles can aid numerical prediction when meaningful variability exists across user preferences. In domains like hotel experiences or digital products, where user expectations tend to be homogeneous, graph-based personalization may offer limited additional signal.

4.2 Ablation Studies

We conduct ablation experiments to assess the impact of different retrieval configurations on PGraphRAG's performance. Specifically, we vary the retrieval depth (i.e., top-k), the retrieval scope (user-only vs. user+neighbors), and the retriever model (BM25 vs. Contriever). Full results and analysis are provided in Appendix A.

5 Conclusion

We presented PGraphRAG, a framework that enhances personalized text generation by integrating user-centric knowledge graphs into retrieval-augmented generation. Unlike prior methods that rely solely on user history, PGraphRAG enriches generation with structured user profiles, enabling adaptive personalization even in sparse data settings. Our experiments show that graph-based retrieval significantly improves performance across diverse tasks, outperforming state-of-the-art baselines. Beyond improved metrics, PGraphRAG introduces a scalable design that generalizes user preferences and adapts to new users through structural retrieval. This work lays a foundation for future personalized LLM systems, particularly in applications requiring robustness to data sparsity, cold starts, and context adaptation.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative*

Long Text Generation	Metric	PGraphRAG	LaMP	No-Retrieval	Random-Retrieval
LLaMA-3.1-8B-Instruct					
Task 1: User-Product Review Generation	ROUGE-1	**0.178**	0.173	0.172	0.124
	ROUGE-L	**0.129**	0.129	0.123	0.094
	METEOR	0.151	0.138	**0.154**	0.099
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.199	0.231	0.216
	ROUGE-L	**0.157**	0.129	0.145	0.132
	METEOR	**0.191**	0.152	0.153	0.152
Task 3: Stylized Feedback Generation	ROUGE-1	**0.217**	0.186	0.190	0.184
	ROUGE-L	**0.158**	0.134	0.131	0.108
	METEOR	**0.178**	0.177	0.167	0.122
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.188**	0.176	0.174	0.146
	ROUGE-L	**0.147**	0.141	0.136	0.116
	METEOR	**0.145**	0.125	0.131	0.109
GPT-4o-mini					
Task 1: User-Product Review Generation	ROUGE-1	**0.189**	0.171	0.169	0.159
	ROUGE-L	**0.130**	0.117	0.116	0.114
	METEOR	**0.196**	0.176	0.177	0.153
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.221	0.223	0.234
	ROUGE-L	**0.152**	0.135	0.135	0.139
	METEOR	**0.206**	0.164	0.166	0.181
Task 3: Stylized Feedback Generation	ROUGE-1	**0.211**	0.185	0.187	0.177
	ROUGE-L	**0.140**	0.123	0.123	0.121
	METEOR	**0.202**	0.183	0.189	0.165
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.194**	0.168	0.170	0.175
	ROUGE-L	**0.144**	0.125	0.128	0.133
	METEOR	**0.171**	0.154	0.152	0.149

Table 3: Zero-shot performance on the test set for the Long Text Generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. For each model, the best retriever configuration was selected based on validation performance.

Short Text Generation	Metric	PGraphRAG	LaMP	No-Retrieval	Random-Retrieval
LLaMA-3.1-8B-Instruct					
Task 5: User Product Review Title Generation	ROUGE-1	**0.131**	0.124	0.121	0.103
	ROUGE-L	**0.125**	0.118	0.115	0.098
	METEOR	**0.125**	0.117	0.112	0.096
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.127**	0.126	0.122	0.118
	ROUGE-L	**0.118**	0.117	0.114	0.110
	METEOR	0.102	**0.106**	0.101	0.093
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.149**	0.140	0.136	0.133
	ROUGE-L	**0.142**	0.134	0.131	0.123
	METEOR	**0.142**	0.136	0.129	0.121
Task 8: Multi-lingual Review Title Generation	ROUGE-1	0.124	0.121	**0.125**	0.120
	ROUGE-L	0.116	**0.122**	0.117	0.110
	METEOR	**0.108**	0.094	0.092	0.103
GPT-4o-mini					
Task 5: User Product Review Title Generation	ROUGE-1	**0.115**	0.108	0.113	0.102
	ROUGE-L	**0.112**	0.105	0.110	0.099
	METEOR	**0.099**	0.091	0.093	0.085
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.116**	0.108	0.114	0.112
	ROUGE-L	**0.111**	0.104	0.109	0.107
	METEOR	**0.081**	0.075	0.079	0.076
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.122**	0.113	0.114	0.115
	ROUGE-L	**0.118**	0.109	0.110	0.111
	METEOR	**0.104**	0.096	0.097	0.093
Task 8: Multi-lingual Review Title Generation	ROUGE-1	0.111	0.115	**0.118**	0.108
	ROUGE-L	0.105	0.107	**0.110**	0.102
	METEOR	0.083	0.088	**0.089**	0.078

Table 4: Zero-shot performance on the test set for the Short Text Generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. For each model, the best retriever configuration was selected based on validation performance.

AI Systems (PERSONALIZE 2024), pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Reinald Kim Amplayo, Jihyeok Kim, Sua Sung, and Seung-won Hwang. 2018. Cold-start aware user and product attention for sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2535–2544, Melbourne, Australia. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation.

Eduardo G. Cortes, Ana Luiza Vianna, Mikaela Martins, Sandro Rigo, and Rafael Kunst. 2024. LLMs and translation: different approaches to localization between Brazilian Portuguese and European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 45–55, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Datafiniti. 2017. Hotel reviews, version 5. Retrieved September 15, 2024 from `https://www.kaggle.com/datasets/datafiniti/hotel-reviews/data`.

Datafiniti. 2018. Grammar and online product reviews, version 1. Retrieved September 15, 2024 from `https://www.kaggle.com/datasets/datafiniti/grammar-and-online-product-reviews`.

Gladson de Araujo, Tiago de Melo, and Carlos Maurício S. Figueiredo. 2024. Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? a preliminary study. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 13–21, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Xiaolei Huang, Lucie Flek, Franck Dernoncourt, Charles Welch, Silvio Amir, Ramit Sawhney, and Diyi Yang. 2022. Usernlp'22: 2022 international workshop on user-centered natural language processing. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 1176–1177, New York, NY, USA. Association for Computing Machinery.

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2014. Enriching cold start personalized language model using social network information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 611–617, Baltimore, Maryland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, and Yoshi Suhara. 2024. Noisy pairing and partial supervision for stylized opinion summarization. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 13–23, Tokyo, Japan. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Srinivas Ramesh Kamath, Fahime Same, and Saad Mahamood. 2024. Generating hotel highlights from unstructured text using LLMs. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 280–288, Tokyo, Japan. Association for Computational Linguistics.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation.

Shin Kanouchi, Masato Neishi, Yuta Hayashibe, Hiroki Ouchi, and Naoaki Okazaki. 2020. You may like this hotel because ...: Identifying evidence for explainable recommendations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 890–899, Suzhou, China. Association for Computational Linguistics.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ziqing Liu, Enwei Peng, Shixing Yan, Guozheng Li, and Tianyong Hao. 2018. T-know: a knowledge graph-based question answering and infor-mation retrieval system for traditional Chinese medicine. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 15–19, Santa Fe, New Mexico. Association for Computational Linguistics.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized recommendation via prompting large language models.

Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Karen, Zeeshan Ahmed, Dinesh Manocha, and Xuedong Zhang. 2024. DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5132–5139, Torino, Italia. ELRA and ICCL.

Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep learning recommendation model for personalization and recommendation systems. *CoRR*, abs/1906.00091.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01: an open product reviews corpus. In *STIL-Symposium in Information and Human Language Technology*.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Ahmmad O. M. Saleh, Gokhan Tur, and Yücel Saygín. 2024. Sg-rag: Multi-hop question answering with large language models through knowledge graphs. In *International Conference on Natural Language and Speech Processing*.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Hong Kong, China. Association for Computational Linguistics.

Arthur Scalercio, Maria Finatto, and Aline Paes. 2024. Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15076–15091, Bangkok, Thailand. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Azure AI Services. 2023. Openai (gpt-4o-mini-20240718) [large language model]. https://learn.microsoft.com/en-us/azure/ai-services/openai.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Hongyan Xu, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Pre-trained personalized review summarization with effective salience estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10743–10754, Toronto, Canada. Association for Computational Linguistics.

Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. 2009. User language model for collaborative personalized search. *ACM Trans. Inf. Syst.*, 27(2).

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777, Online. Association for Computational Linguistics.

Hongyu Zang and Xiaojun Wan. 2017. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 168–177, Santiago de Compostela, Spain. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of large language models: A survey.

A Appendix

A.1 Data Construction and Splitting

To construct the user–item interaction graph, we represent users and domain-specific entities (e.g., products, hotels, feedback targets) as nodes, with edges corresponding to user-generated content (e.g., reviews, summaries, ratings). To support graph-based personalization, we require that each selected user has at least one interaction with an entity that is also associated with another user — i.e., a shared neighbor in the bipartite graph. If a randomly selected user interaction does not meet this criterion, we instead sample a different interaction from the same profile. Users without any neighbor-compatible interactions remain in the dataset but are excluded from gold-label selection, since sampling is performed at the edge level rather than over full profiles. This filtering ensures that the graph remains connected and supports comparative evaluation and cold-start scenarios, where even users with minimal history share contextually linked entities with others.

After identifying each user's valid neighbor-linked interaction(s), we divide users into training, development, and test sets while preserving graph connectivity across splits. To ensure that personalization signals remain intact, we apply two levels of neighbor preservation:

1. **Global Neighbor Preservation:** Entities with multiple associated users are grouped so that at least one other user in the same split has interacted with the same entity.

2. **Local Neighbor Preservation:** Once a user is assigned to a split, any other users who interacted with the same entity are also placed in that split to maintain graph connectivity.

We further stratify each split based on user profile size to match the original distribution of user activity while preserving both global and local connectivity. This joint control over profile stratification and neighbor assignment ensures that the resulting graphs in each split maintain realistic interaction patterns and structural properties. Graph statistics are shown in Table 2, task-level data statistics in Table 1, and dataset splits in Table 5.

A.2 Performance Gains

Table 6 shows the relative percent gains of PGraphRAG compared to LaMP across Tasks 1–7.

Dataset	Train Size	Validation Size	Test Size
User-Product Review	20,000	2,500	2,500
Multilingual Product Review	20,000	2,500	2,500
Stylized Feedback	20,000	2,500	2,500
Hotel Experiences	9,000	2,500	2,500

Table 5: Dataset split sizes across training, validation, and test sets for the four domains.

Notably, Task 8 (Multi-lingual Review Title Generation) shows reduced gains, which we attribute to cultural differences in review conventions—for example, the frequent use of the generic phrase Muito bom" (Very good") in Brazilian Portuguese titles. In long-text generation with GPT-4o-mini, PGraphRAG achieves improvements of approximately 15% in ROUGE-1, 13% in ROUGE-L, and 15% in METEOR. Similar trends are seen with LLaMA-3.1-8B, with improvements of 15%, 11%, and 13% respectively. In short-text generation, GPT shows improvements of 5% across all metrics, while LLaMA gains range from 2–6%.

In addition, Table 7 shows the review density per product, where sparsity is balanced from the original graph for both product and user nodes.

A.3 Prompt and Output Examples

Figure 4 shows the prompt template across task types. Below, we show the output for Task 2 comparing PGraphRAG and LaMP against the gold label. PGraphRAG captures specific contextual cues from the graph (e.g., correct location and hotel amenities), while LaMP's output often relies on generic context from the target user's own reviews, which leads to generating irrelevant content (e.g., wrong hotel location).

Gold Review

Title: Amazing stay! **Review:** Had a great stay, our room was very clean and very peaceful. It is in the heart of DT Seattle. We took the train to the hotel and it was pretty easy. We enjoyed walking all over downtown Seattle—no need to rent a car. Major shopping and amazing restaurants right around the corner! We will stay here again!

Model	Metric	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
	ROUGE-1	10.53	18.96	14.05	15.48	6.48	7.41	7.96	-3.48
GPT-4o-mini	ROUGE-L	11.11	12.59	13.82	15.20	6.67	6.73	8.26	-1.87
	METEOR	11.36	25.61	10.38	11.04	8.79	8.00	8.33	-5.68
	ROUGE-1	2.89	32.16	16.67	6.82	5.65	0.79	6.43	2.48
LLaMA-3.1-8B-Instruct	ROUGE-L	0.00	21.71	17.91	4.26	5.93	0.85	5.97	-4.92
	METEOR	9.42	25.66	0.56	16.00	6.84	-3.77	4.41	14.89

Table 6: Relative percentage gains of PGraphRAG over LaMP across Tasks 1–8 using *GPT-4o-mini* and *LLaMA-3.1-8B-Instruct*.

Reviews (n)	**Exact Count (Pct.)**	**Cumulative Count (Pct.)**
1	25,530 (49.69%)	25,530 (49.69%)
2	9,488 (18.47%)	35,018 (68.16%)
3	4,784 (9.31%)	39,802 (77.47%)
4	2,639 (5.14%)	42,441 (82.61%)
5	1,836 (3.57%)	44,277 (86.18%)

Table 7: Distribution of review counts per product in the Amazon Review dataset (Tasks 1, 5, and 9). Most products have five or fewer reviews.

LaMP

Title: Amazing stay! **Review:** I recently had an amazing stay at the Newark Garden Hotel. Despite the circumstances that brought my family together, the hotel provided a comforting and welcoming environment. The rooms were immaculate, and the staff went above and beyond to ensure we felt at home. The hot breakfast each morning was a delightful touch that made our mornings brighter. I highly recommend this hotel for anyone looking for a pleasant and accommodating stay!

PGraphRAG

Title: Amazing stay! **Review:** I had an amazing stay at the Grand Hyatt Seattle! From the moment I checked in, the staff was incredibly welcoming and attentive. The room was spacious, clean, and beautifully appointed, with a stunning view of the city. The bathroom was a highlight, featuring a luxurious soaking tub and a large walk-in shower. The location was perfect, just a short walk to Pike Place Market and other attractions. I also enjoyed the fantastic amenities, including the well-equipped gym and the delicious food options available at the hotel. Overall, it was a wonderful experience, and I can't wait to return!

A.4 PGraphRAG Ablation Details

To assess the contributions of user-specific and neighbor-derived context in our retrieval framework, we conduct an ablation study comparing three variants of PGraphRAG:

- **PGraphRAG**: The full method, which retrieves context from both the target user's profile and neighboring users who share entities (e.g., items or experiences).
- **PGraphRAG-N**: A neighbor-only variant that excludes the target user's own interactions and relies solely on neighboring users for context.
- **PGraphRAG-U**: A user-only variant that restricts retrieval to the target user's own history, ignoring all neighbor signals.

Table 9 shows the results for long-text generation (Tasks 1–4) using GPT-4o-mini and LLaMA-3.1-8B. Both PGraphRAG and PGraphRAG-N consistently outperform PGraphRAG-U across datasets, highlighting the value of graph-based retrieval. Notably, PGraphRAG-N performs on par with or slightly below the full PGraphRAG method, suggesting that neighboring-user context alone is often sufficient for high-quality personalization — especially in low-profile or cold-start scenarios where the target user's history is sparse.

Results for short-text generation tasks (Tasks 5–8) are shown in Table 10. Similar patterns hold, with PGraphRAG and PGraphRAG-N outperforming PGraphRAG-U across most tasks. One exception is Task Hotel Experience Summary Generation, where PGraphRAG-U slightly outperforms all graph-based variants, possibly due to limited variation in the data or a mismatch between neighbor context and task-specific semantics.

A.5 Impact of the Retrieved Items k

To understand how the size of the retrieved context affects performance, we conduct an ablation study varying the number of retrieved entries $k \in 1, 2, 4$. Table 11 reports results for long-text generation (Tasks 1–4), using *GPT-4o-mini* and *LLaMA-3.1-8B-Instruct*. Corresponding results for short-text generation (Tasks 5–8) appear in Table 12.

Ordinal Classfication	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 9: User Product Review Ratings	MAE ↓	0.3400	**0.3132**	0.3212	0.3272
	RMSE ↓	0.7668	**0.7230**	0.7313	0.7616
Task 10: Hotel Experience Ratings	MAE ↓	0.3688	0.3492	**0.3340**	0.3804
	RMSE ↓	0.6771	0.6527	**0.6372**	0.6971
Task 11: Stylized Feedback Ratings	MAE ↓	0.3476	**0.3268**	0.3256	0.3704
	RMSE ↓	0.7247	**0.6803**	0.6806	0.7849
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4928**	0.5016	0.5084	0.5096
	RMSE ↓	**0.8367**	0.8462	0.8628	0.8542
GPT-4o-mini					
Task 9: User Product Review Ratings	MAE ↓	0.3832	0.3480	**0.3448**	0.4188
	RMSE ↓	0.7392	**0.7065**	**0.7065**	0.8082
Task 10: Hotel Experience Ratings	MAE ↓	**0.3284**	0.3336	0.3336	0.3524
	RMSE ↓	**0.6083**	0.6197	0.6197	0.6384
Task 11: Stylized Feedback Ratings	MAE ↓	0.3476	**0.3448**	0.3416	0.4080
	RMSE ↓	0.6738	**0.6669**	0.6711	0.7370
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4348**	0.4444	0.4564	0.4700
	RMSE ↓	**0.7367**	0.7608	0.7718	0.8112

Table 8: Performance comparison on rating prediction tasks (Tasks 9-12) using *GPT-4o-mini* and *LLaMA-3.1-8B*.

Long Text Generation	Metric	PGraphRAG	PGraphRAG-N	PGraphRAG-U
LLaMA-3.1-8B-Instruct				
Task 1: User-Product Review Generation	ROUGE-1	0.173	**0.177**	0.168
	ROUGE-L	0.124	**0.127**	0.125
	METEOR	0.150	**0.154**	0.134
Task 2: Hotel Experiences Generation	ROUGE-1	0.263	**0.272**	0.197
	ROUGE-L	0.156	**0.162**	0.128
	METEOR	0.191	**0.195**	0.121
Task 3: Stylized Feedback Generation	ROUGE-1	**0.226**	0.222	0.181
	ROUGE-L	**0.171**	0.165	0.134
	METEOR	**0.192**	0.186	0.147
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.174**	0.172	0.174
	ROUGE-L	0.139	0.137	**0.141**
	METEOR	**0.133**	0.126	0.125
GPT-4o-mini				
Task 1: User-Product Review Generation	ROUGE-1	**0.186**	0.185	0.169
	ROUGE-L	**0.126**	0.125	0.114
	METEOR	**0.187**	0.185	0.170
Task 2: Hotel Experiences Generation	ROUGE-1	0.265	**0.268**	0.217
	ROUGE-L	0.152	**0.153**	0.132
	METEOR	0.206	**0.209**	0.161
Task 3: Stylized Feedback Generation	ROUGE-1	**0.205**	0.204	0.178
	ROUGE-L	**0.139**	0.138	0.121
	METEOR	**0.203**	0.198	0.178
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.191**	0.190	0.164
	ROUGE-L	**0.142**	0.140	0.123
	METEOR	**0.173**	0.169	0.155

Table 9: Ablation study results for long text generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. PGraphRAG-N represents Neighbors-only context retrieval and PGraphRAG-U represents User-only context retrieval.

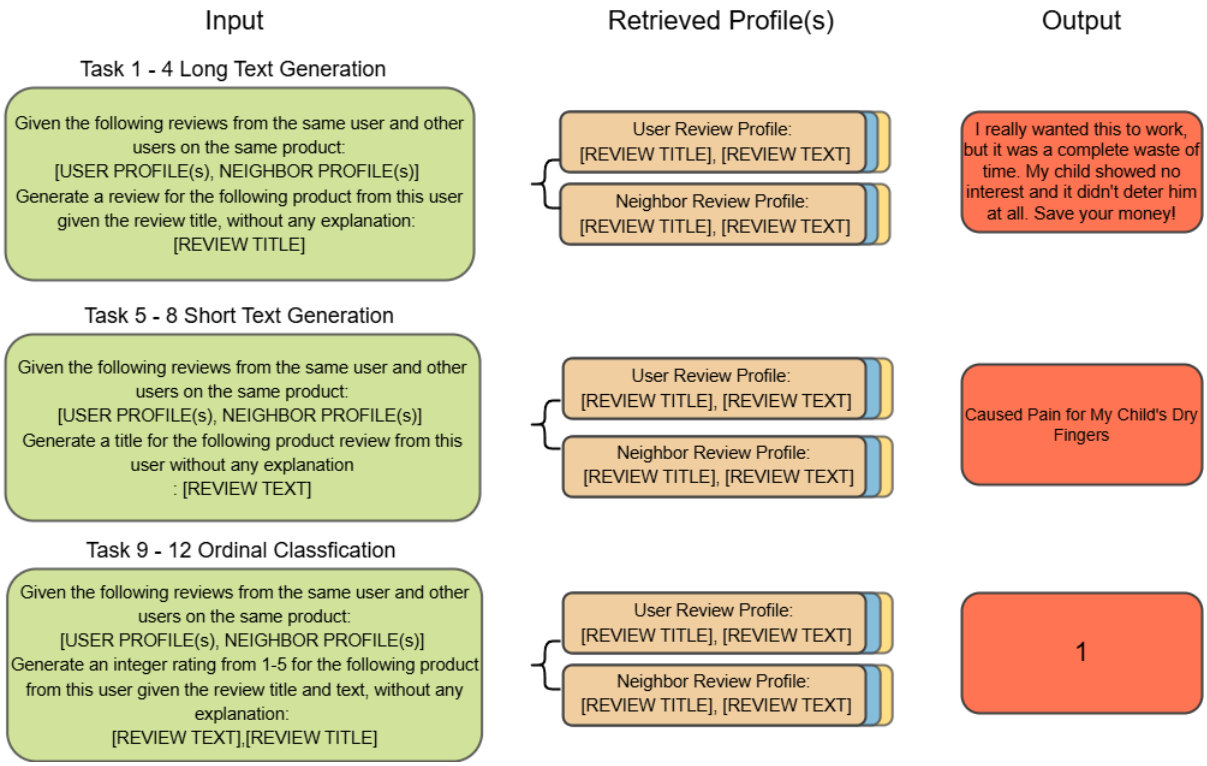


Figure 4: Prompt configurations used for each task type. Teletype placeholders (e.g., `{{title}}`) are replaced with task-specific input and retrieved context at inference time.

Short Text Generation	**Metric**	**PGraphRAG**	**PGraphRAG-N**	**PGraphRAG-U**
LLaMA-3.1-8B-Instruct				
Task 5: User Product Review Title Generation	ROUGE-1	0.125	**0.129**	0.115
	ROUGE-L	0.119	**0.123**	0.109
	METEOR	0.117	**0.120**	0.111
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.121	**0.124**	0.119
	ROUGE-L	0.113	**0.115**	0.111
	METEOR	0.099	0.103	**0.105**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.132	**0.135**	0.128
	ROUGE-L	0.128	**0.130**	0.124
	METEOR	0.129	**0.132**	0.124
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.131**	0.131	0.124
	ROUGE-L	**0.123**	0.122	0.114
	METEOR	**0.118**	0.110	0.098
GPT-4o-mini				
Task 5: User Product Review Title Generation	ROUGE-1	0.111	**0.116**	0.112
	ROUGE-L	0.106	**0.111**	0.108
	METEOR	0.097	**0.099**	0.095
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.118	**0.119**	0.109
	ROUGE-L	0.112	**0.113**	0.104
	METEOR	**0.085**	**0.085**	0.077
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.109**	0.107	0.108
	ROUGE-L	**0.107**	0.105	0.104
	METEOR	**0.096**	0.094	0.091
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.108	0.109	**0.116**
	ROUGE-L	0.104	0.104	**0.109**
	METEOR	0.082	0.089	**0.091**

Table 10: Ablation study results for short text generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. PGraphRAG-N represents Neighbors-only context retrieval and PGraphRAG-U represents User-only context retrieval.

Overall, increasing k generally leads to improved generation performance across tasks and models. This trend highlights the value of larger retrieved contexts, which provide richer signals about user preferences and item semantics. The gains are especially evident when moving from $k = 1$ to $k = 2$, though marginal returns diminish between $k = 2$ and $k = 4$ in some cases.

That said, the benefit of higher k values is constrained by data sparsity. Many user profiles contain fewer than four qualifying interactions—especially in cold-start settings. In such cases, the retriever returns all available entries, even if they are fewer than the specified k. As a result, the effective retrieved context size varies across users, especially in the low-profile regime. This behavior reflects the practical limitations of personalization at scale and underscores the importance of designing retrieval-aware systems that can operate under sparse supervision.

Long Text Generation	**Metric**	$k = 1$	$k = 2$	$k = 4$
LLaMA-3.1-8B-Instruct				
Task 1: User-Product Review Generation	ROUGE-1	0.160	0.169	**0.173**
	ROUGE-L	0.121	**0.125**	0.124
	METEOR	0.125	0.138	**0.150**
Task 2: Hotel Experiences Generation	ROUGE-1	0.230	0.251	**0.263**
	ROUGE-L	0.141	0.151	**0.156**
	METEOR	0.152	0.174	**0.191**
Task 3: Stylized Feedback Generation	ROUGE-1	0.200	0.214	**0.226**
	ROUGE-L	0.158	0.165	**0.171**
	METEOR	0.154	0.171	**0.192**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.163	0.169	**0.174**
	ROUGE-L	0.134	0.137	**0.139**
	METEOR	0.113	0.122	**0.133**
GPT-4o-mini				
Task 1: User-Product Review Generation	ROUGE-1	0.176	0.184	**0.186**
	ROUGE-L	0.121	0.125	**0.126**
	METEOR	0.168	0.180	**0.187**
Task 2: Hotel Experiences Generation	ROUGE-1	0.250	0.260	**0.265**
	ROUGE-L	0.146	0.150	**0.152**
	METEOR	0.188	0.198	**0.206**
Task 3: Stylized Feedback Generation	ROUGE-1	0.196	0.200	**0.205**
	ROUGE-L	0.136	0.136	**0.139**
	METEOR	0.186	0.192	**0.203**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.163	0.169	**0.174**
	ROUGE-L	0.134	0.137	**0.139**
	METEOR	0.113	0.122	**0.133**

Table 11: Ablation study results showing the impact of varying k (number of retrieved neighbors) on PGraphRAG's performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on long-text generation tasks (Tasks 1 - 4).

A.6 Impact of Retriever Method $\mathcal{R}$

We evaluate how the choice of retriever affects the performance of PGraphRAG by comparing two retrieval backends: BM25, a sparse keyword-based retriever, and Contriever, a dense unsupervised retriever based on sentence embeddings.

Short Text Generation	**Metric**	$k = 1$	$k = 2$	$k = 4$
LLaMA-3.1-8B-Instruct				
Task 5: User Product Review Title Generation	ROUGE-1	**0.128**	0.123	0.125
	ROUGE-L	**0.121**	0.118	0.119
	METEOR	**0.123**	0.118	0.117
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.122**	0.121	0.121
	ROUGE-L	0.112	**0.114**	0.113
	METEOR	**0.104**	0.102	0.099
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.129	**0.132**	**0.132**
	ROUGE-L	0.124	0.126	**0.128**
	METEOR	0.129	**0.130**	0.129
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.129	0.126	**0.131**
	ROUGE-L	0.120	0.119	**0.123**
	METEOR	0.117	0.116	**0.118**
GPT-4o-mini				
Task 5: User Product Review Title Generation	ROUGE-1	**0.111**	0.110	**0.111**
	ROUGE-L	**0.106**	0.105	**0.106**
	METEOR	0.093	0.094	**0.097**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.114	0.114	**0.118**
	ROUGE-L	0.109	0.109	**0.112**
	METEOR	0.082	0.082	**0.085**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.100	0.103	**0.109**
	ROUGE-L	0.098	0.101	**0.107**
	METEOR	0.087	0.090	**0.096**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.104	0.104	**0.108**
	ROUGE-L	0.098	0.098	**0.104**
	METEOR	0.077	0.078	**0.082**

Table 12: Ablation study results showing the impact of varying k (number of retrieved neighbors) on PGraphRAG's performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on short-text generation tasks (Tasks 5-8).

Table 13 reports results for long-text generation (Tasks 1–4), and Table 14 provides results for short-text generation (Tasks 5–8). Across both GPT-4o-mini and LLaMA-3.1-8B-Instruct models, we observe that PGraphRAG performs consistently well regardless of the retrieval method. The differences between BM25 and Contriever are minor, and no retriever dominates across all datasets or metrics.

These findings indicate that PGraphRAG is robust to the choice of retriever and does not rely on fine-tuned or heavily engineered retrieval strategies. While BM25 sometimes yields slightly higher scores, the overall parity suggests that our graph-based retrieval and prompting framework can effectively integrate contextual signals from either sparse or dense retrieval methods.

A.7 Impact of Ranked Retrieval

Table 15 evaluates the role of ranking in PGraphRAG by comparing the following retrieval variants:

Long Text Generation	Metric	Contriever	BM25
LLaMA-3.1-8B-Instruct			
Task 1: User-Product Review Generation	ROUGE-1	0.172	**0.173**
	ROUGE-L	0.122	**0.124**
	METEOR	**0.153**	0.150
Task 2: Hotel Experiences Generation	ROUGE-1	0.262	**0.263**
	ROUGE-L	0.155	**0.156**
	METEOR	0.190	**0.191**
Task 3: Stylized Feedback Generation	ROUGE-1	0.195	**0.226**
	ROUGE-L	0.138	**0.171**
	METEOR	0.180	**0.192**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.172	**0.174**
	ROUGE-L	0.134	**0.139**
	METEOR	**0.135**	0.133
GPT-4o-mini			
Task 1: User-Product Review Generation	ROUGE-1	0.182	**0.186**
	ROUGE-L	0.122	**0.126**
	METEOR	0.184	**0.187**
Task 2: Hotel Experiences Generation	ROUGE-1	0.264	**0.265**
	ROUGE-L	**0.152**	**0.152**
	METEOR	**0.207**	0.206
Task 3: Stylized Feedback Generation	ROUGE-1	0.194	**0.205**
	ROUGE-L	0.128	**0.139**
	METEOR	0.201	**0.203**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.190	**0.191**
	ROUGE-L	0.141	**0.142**
	METEOR	**0.174**	0.173

Table 13: Ablation study results showing the effect of retriever choice on PGraphRAG performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on the long-text generation task (Tasks 1-4).

Short Text Generation	Metric	Contriever	BM25
LLaMA-3.1-8B-Instruct			
Task 5: User Product Review Title Generation	ROUGE-1	0.122	**0.125**
	ROUGE-L	0.116	**0.119**
	METEOR	0.115	**0.117**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.117	**0.121**
	ROUGE-L	0.110	**0.113**
	METEOR	0.095	**0.099**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.125	**0.132**
	ROUGE-L	0.121	**0.128**
	METEOR	0.122	**0.129**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.126	**0.131**
	ROUGE-L	0.118	**0.123**
	METEOR	0.112	**0.118**
GPT-4o-mini			
Task 5: User Product Review Title Generation	ROUGE-1	**0.113**	0.111
	ROUGE-L	**0.108**	0.106
	METEOR	**0.097**	**0.097**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.113	**0.118**
	ROUGE-L	0.107	**0.112**
	METEOR	0.080	**0.085**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.108	**0.109**
	ROUGE-L	0.106	**0.107**
	METEOR	0.094	**0.096**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.108**	**0.108**
	ROUGE-L	0.103	**0.104**
	METEOR	**0.082**	**0.082**

Table 14: Ablation study results showing the effect of retriever choice on PGraphRAG performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on the short-text generation task (Tasks 5-8).

1. PGraphRAG*: retrieves $k = 4$ randomly sampled entries from the profile without ranking.
2. PGraphRAG**: retrieves and includes all available context within the model's input limit (i.e., $k \to \infty$).

As expected, PGraphRAG** performs best due to its access to a larger and more diverse context. However, our focus is on the impact of removing ranking while keeping k fixed.

Removing ranking (PGraphRAG $\to$ PGraphRAG*) leads to a drop in ROUGE-1 of 2.29% for long-text generation and 3.18% for short-text tasks. The effect is also visible in user-only retrieval (PGraphRAG-U $\to$ PGraphRAG-U*), with decreases of 0.92% and 1.98% for long- and short-text tasks, respectively. These consistent declines underscore the importance of ranking in identifying relevant context.

While PGraphRAG** demonstrates the upper bound of performance, its scalability is limited due to cost and context length constraints. In contrast, ranked retrieval with a fixed k (as in PGraphRAG) offers a strong balance between performance and efficiency, making it more suitable for real-world deployment.

A.8 Evaluating Different GPT Variants

To compare the performance of different GPT variants, we evaluate PGraphRAG using a fixed retrieval configuration (BM25, $k = 4$) across two OpenAI models: GPT-4o-mini and GPT-o1. Among these, GPT-4o-mini demonstrated the best trade-off between accuracy, cost, and consistency on long-text generation tasks.

A.9 Impact of Length Constraints in GPT Model

In short-text generation tasks, controlling output length is essential to balance informativeness and conciseness. We evaluate the effect of fixed output constraints of 3, 5, and 10 words. Empirically, a 5-word constraint offers the best trade-off across evaluation metrics, yielding higher-quality outputs with minimal verbosity. We therefore adopt 5-word

Task	Metric	PGraphRAG	PGraphRAG*	PGraphRAG**	PGraphRAG-U	PGraphRAG-U*	PGraphRAG-U**
Long Text Generation							
Task 1: User-Product Review Generation	ROUGE-1	0.189	0.186	**0.191**	0.171	0.169	0.170
	ROUGE-L	**0.130**	0.125	**0.130**	0.117	0.114	0.117
	METEOR	0.196	0.188	**0.205**	0.176	0.173	0.180
Task 2: Hotel Experiences Generation	ROUGE-1	0.263	0.266	**0.267**	0.221	0.223	0.225
	ROUGE-L	0.152	0.152	**0.153**	0.135	0.134	0.135
	METEOR	0.206	0.209	**0.216**	0.164	0.168	0.171
Task 3: Stylized Feedback Generation	ROUGE-1	**0.211**	0.200	0.210	0.185	0.180	0.186
	ROUGE-L	**0.140**	0.133	0.136	0.123	0.122	0.123
	METEOR	0.202	0.206	**0.225**	0.183	0.184	0.189
Task 4: Multilingual Product Review Generation	ROUGE-1	0.194	0.188	**0.196**	0.168	0.167	0.171
	ROUGE-L	**0.144**	0.138	0.141	0.125	0.125	0.128
	METEOR	0.171	0.176	**0.188**	0.154	0.155	0.155
Short Text Generation							
Task 5: User Product Review Title Generation	ROUGE-1	0.115	0.114	**0.119**	0.108	0.108	0.111
	ROUGE-L	0.112	0.109	**0.114**	0.105	0.102	0.105
	METEOR	0.099	0.121	**0.128**	0.091	0.116	0.119
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.116	0.117	**0.121**	0.108	**0.121**	0.119
	ROUGE-L	0.111	0.107	**0.112**	0.104	0.111	0.110
	METEOR	0.081	0.104	**0.109**	0.075	**0.109**	0.107
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.122**	0.111	0.120	0.113	0.115	0.114
	ROUGE-L	**0.118**	0.105	0.114	0.109	0.109	0.108
	METEOR	0.104	0.117	**0.126**	0.096	0.124	0.123
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.111	0.108	0.112	**0.115**	0.110	0.110
	ROUGE-L	0.105	0.100	0.104	**0.107**	0.103	0.101
	METEOR	0.083	0.101	0.105	0.088	**0.108**	0.107

Table 15: Zero-shot test set results for text generation using *GPT-4o-mini*. **PGraphRAG*** denotes retrieval of $k = 4$ randomly selected entries without ranking, while **PGraphRAG**** represents unbounded retrieval up to the model's context limit ($k \to \infty$).

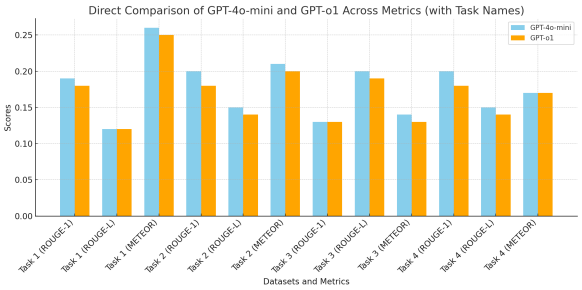


Figure 5: Comparison of *GPT-4o-mini* and *GPT-o1-preview* on the test set across Tasks 1–4 using BM25 retriever with $k = 4$.

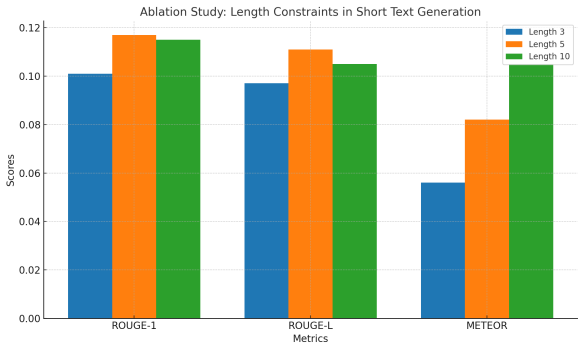


Figure 6: Effect of different output length constraints (3, 5, and 10 words) on short-text generation performance using PGraphRAG, measured on the validation set.

outputs as the default setting for all short-text generation experiments.

A.10 Validation Results

We conduct extensive validation experiments across all representative tasks, evaluating all combinations of language models, retrieval strategies, and top-k settings. The goal is to identify the most effective configuration for each task prior to test-time evaluation.

Results are reported in Tables 16, 17, and 18, corresponding to long-text generation, short-text generation, and ordinal classification tasks, respectively.

For each task, we select the best-performing configuration based on validation performance. These selected settings are then used in the test set evaluation. Notably, trends observed in the validation phase remain consistent in the test results, reinforcing the robustness of our setup.

B Related Work

Personalization in NLP

Personalization in natural language processing (NLP) focuses on tailoring responses to user-specific preferences, behaviors, and contexts, improving user experience and task performance. Early work in personalized generation relied on neural encoder-decoder models and incorporated attributes such as sentiment (Zang and Wan, 2017),

Long Text Generation	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 1: User-Product Review Generation	ROUGE-1	**0.173**	0.168	0.172	0.126
	ROUGE-L	0.124	**0.125**	0.121	0.095
	METEOR	0.150	0.134	**0.152**	0.101
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.197	0.224	0.211
	ROUGE-L	**0.156**	0.128	0.141	0.130
	METEOR	**0.191**	0.121	0.148	0.147
Task 3: Stylized Feedback Generation	ROUGE-1	**0.226**	0.181	0.177	0.142
	ROUGE-L	**0.171**	0.134	0.125	0.104
	METEOR	**0.192**	0.147	0.168	0.119
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.174**	0.174	0.173	0.146
	ROUGE-L	0.139	**0.141**	0.134	0.117
	METEOR	**0.133**	0.125	0.130	0.110
GPT-4o-mini					
Task 1: User-Product Review Generation	ROUGE-1	**0.186**	0.169	0.168	0.157
	ROUGE-L	**0.126**	0.114	0.113	0.112
	METEOR	**0.187**	0.170	0.173	0.148
Task 2: Hotel Experiences Generation	ROUGE-1	**0.265**	0.217	0.222	0.233
	ROUGE-L	**0.152**	0.132	0.133	0.138
	METEOR	**0.206**	0.161	0.164	0.164
Task 3: Stylized Feedback Generation	ROUGE-1	**0.205**	0.178	0.177	0.168
	ROUGE-L	**0.139**	0.121	0.119	0.117
	METEOR	**0.203**	0.178	0.184	0.160
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.191**	0.164	0.167	0.171
	ROUGE-L	**0.142**	0.123	0.125	0.131
	METEOR	**0.173**	0.155	0.153	0.150

Table 16: Zero-shot Validation set results for long text generation using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on Tasks 1-4.

stylistic cues (Dong et al., 2017), and demographic metadata (Huang et al., 2014). To address data sparsity, approaches such as warm-start attention (Amplayo et al., 2018) and user embeddings were developed.

Recent efforts have expanded personalization using retrieval-augmented generation (RAG) strategies. Methods like in-context prompting (Lyu et al., 2024), retrieval-enhanced summarization (Richardson et al., 2023), and optimization via reinforcement learning or distillation (Salemi et al., 2024a) have improved output fluency and relevance. Benchmarking frameworks such as LaMP (Salemi et al., 2024b) and LongLaMP (Kumar et al., 2024) have standardized evaluation of personalized tasks (e.g., email writing, abstract generation). Meanwhile, retrieval-enhanced generation pipelines (Kim et al., 2020) improve long-form text by incorporating relevant user history.

However, most prior work assumes dense, high-coverage user history, limiting effectiveness in cold-start or sparse-profile scenarios. Few approaches leverage structured representations (e.g., knowledge graphs) to generalize beyond individual user traces. This gap highlights a need for models that can retrieve personalized yet diverse context using structured user-item relationships.

Knowledge Graphs and Retrieval-Augmented Generation (RAG)

Knowledge graphs (KGs) provide structured, relational context useful in a variety of NLP tasks such as question answering, entity linking, and reasoning (Liu et al., 2018; Schneider et al., 2022). By leveraging graph traversal and multi-hop paths, KGs enable precise contextualization in tasks that require reasoning over entity relationships (Salnikov et al., 2023). Recent techniques such as data synthesis and subgraph construction have improved KG scalability and coverage (Agarwal et al., 2021).

In parallel, retrieval-augmented generation (RAG) frameworks enhance LLMs by incorporating external memory or document retrieval into the generation process (Izacard and Grave, 2020).

Short Text Generation	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 5: User Product Review Title Generation	ROUGE-1	**0.125**	0.114	0.111	0.101
	ROUGE-L	**0.119**	0.108	0.105	0.095
	METEOR	**0.117**	0.111	0.104	0.094
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.121**	0.119	0.115	0.115
	ROUGE-L	**0.113**	0.111	0.108	0.107
	METEOR	**0.105**	**0.105**	0.100	0.094
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.132**	0.128	0.127	0.108
	ROUGE-L	**0.128**	0.124	0.122	0.104
	METEOR	**0.129**	0.124	0.118	0.103
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.132**	0.128	0.108	0.127
	ROUGE-L	**0.128**	0.124	0.104	0.122
	METEOR	**0.129**	0.124	0.103	0.118
GPT-4o-mini					
Task 5: User Product Review Title Generation	ROUGE-1	**0.114**	0.106	0.109	0.107
	ROUGE-L	**0.107**	0.100	0.103	0.102
	METEOR	**0.119**	0.115	0.116	0.109
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.115**	**0.115**	0.114	0.112
	ROUGE-L	0.105	**0.106**	**0.106**	0.103
	METEOR	0.105	**0.106**	**0.106**	0.099
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.105**	0.101	**0.105**	0.098
	ROUGE-L	**0.102**	0.097	0.101	0.093
	METEOR	**0.118**	0.111	0.118	0.105
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.108**	0.106	**0.108**	0.103
	ROUGE-L	0.099	0.098	**0.099**	0.095
	METEOR	0.101	0.102	**0.103**	0.095

Table 17: Zero-shot Validation set results for short text generation using *LLaMA-3.1-8B* and *GPT-4o-mini* on Tasks 5-8.

When integrated with KGs, RAG enables structured multi-hop reasoning (Saleh et al., 2024), rare entity recognition (Mathur et al., 2024), and hallucination reduction in generative outputs (Kang et al., 2023; Chen et al., 2023).

Despite these gains, scaling KGs in real-world systems (e.g., personalized recommendation) remains challenging (Ji et al., 2022). Graph construction, update, and refinement require sophisticated methods to ensure correctness and completeness (Paulheim, 2017). Moreover, traditional RAG pipelines using dense vector retrieval may struggle to integrate symbolic signals from structured graphs or handle noisy or misaligned data sources (Gao et al., 2024).

Toward Structured Personalization via Graph-Augmented RAG

The intersection of personalization, knowledge graphs, and RAG presents a promising research direction. Recent surveys (Zhang et al., 2024) emphasize the importance of personalization in LLMs but call for approaches that generalize across users with limited history and incorporate structured context. Our work addresses this by using user-centric bipartite graphs to retrieve not only user-authored content but also related interactions from similar users, enabling robust personalization under sparse conditions.

Unlike conventional user-history-based personalization, graph-augmented RAG offers a principled way to incorporate both individual and community signals—supporting generalization, diversity, and data efficiency at inference time.

C Limitations

While PGraphRAG demonstrates strong performance across personalized generation tasks, there are several considerations that present opportunities for future enhancement.

Scalability considerations. Although personalization approaches can raise scalability concerns, PGraphRAG is designed for efficient large-scale deployment. It constructs a unified, sparse user-item bipartite graph offline — i.e., graph construction is a one-time cost, similar to those used in scalable

Ordinal Classfication	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 9: User Product Review Ratings	MAE ↓	0.3272	0.3220	**0.3200**	0.3516
	RMSE ↓	0.7531	**0.7280**	0.7294	0.7972
Task 10: Hotel Experience Ratings	MAE ↓	0.3868	0.3685	**0.3614**	0.4008
	RMSE ↓	0.6989	0.6750	**0.6643**	0.7178
Task 11: Stylized Feedback Ratings	MAE ↓	**0.3356**	0.3368	0.3372	0.3812
	RMSE ↓	0.6856	0.6859	**0.6826**	0.7759
Task 12: Multi-lingual Product Ratings	MAE ↓	0.5228	**0.5216**	0.5282	0.5392
	RMSE ↓	0.8483	**0.8395**	0.8519	0.8704
GPT-4o-mini					
Task 9: User Product Review Ratings	MAE ↓	0.3652	0.3508	**0.3484**	0.4176
	RMSE ↓	0.7125	0.6943	**0.6925**	0.7792
Task 10: Hotel Experience Ratings	MAE ↓	**0.3308**	0.3472	0.3528	0.3640
	RMSE ↓	**0.6056**	0.6394	0.6475	0.6627
Task 11: Stylized Feedback Ratings	MAE ↓	**0.3340**	0.3364	0.3356	0.3972
	RMSE ↓	0.6515	0.6545	**0.6484**	0.7158
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4568**	0.4832	0.4908	0.4820
	RMSE ↓	**0.7414**	0.7808	0.7897	0.7917

Table 18: Performance comparison on rating prediction tasks (Tasks 9-12) using *GPT-4o-mini* and *LLaMA-3.1-8B-Instruct* on the validation set. Results are reported using MAE and RMSE metrics across retrieval methods.

recommender systems. As shown in Table 2, the graph is inherently sparse, enabling efficient storage and indexing. At inference time, rather than retrieving over the entire corpus as in traditional RAG settings, PGraphRAG scopes retrieval to a localized subgraph centered on the input user. This subgraph includes both the user's own interactions and those of neighboring users who share items. Standard retrievers (e.g., BM25 or Contriever) are then applied over this constrained set, significantly reducing search overhead while retaining personalized context. This design keeps runtime and memory usage low and supports scalable deployment across large user bases. In future work, we plan to explore compression techniques and real-time profile updates to further enhance scalability in dynamic environments.

Graph completeness and data sparsity. While the quality of retrieval can be influenced by the completeness of the user-centric graph, PGraphRAG is explicitly designed to operate under sparse and noisy conditions. Our benchmark includes users with minimal interaction history, yet results show strong performance across tasks compared to baseline methods. This robustness arises from PGraphRAG's graph-based retrieval strategy, which leverages neighboring nodes to provide relevant contextual signals even when direct user data is limited. Nonetheless, integrating implicit signals (e.g., click rate or engagement time) and developing more resilient retrieval methods for incomplete graphs remains a promising direction for future work.

Generalization vs. user adaptation. A core challenge lies in developing training strategies that balance individual personalization with generalization across user populations. While our approach augments prompts with structured context, future work may explore personalized fine-tuning or adapter layers to enhance this tradeoff further.

Static user profiles. Currently, user profiles are treated as static during evaluation. In real-world scenarios, preferences evolve over time. Extending the framework to model temporal dynamics and support profile updates is a promising direction for improving long-term personalization.

Improving Classical Language Machine Translation using Supervised Fine-Tuning with Philological Commentary

Yuzuki Tsukagoshi
The University of Tokyo / Tokyo, Japan
yuzuki@l.u-tokyo.ac.jp

Ikki Ohmukai
The University of Tokyo / Tokyo, Japan
i2k@l.u-tokyo.ac.jp

Abstract

This paper presents a novel approach to improving ancient language translation by integrating scholarly philological commentary into language model training. Using the Ṛgveda with authoritative English translations and annotations from Jamison & Brereton, we employ supervised fine-tuning on five models: GPT-4.1 nano, Gemini 2.5 flash, Llama 3.2 3B, Llama 3.1 8B, and the Sanskrit-specialized Gemma 2 Mitra. Our methodology compares standard fine-tuning against commentary-enhanced training, where models receive both Sanskrit texts and philological commentary as input. Evaluation using BLEU scores demonstrates consistent improvements across four larger models when incorporating scholarly commentary, with particularly strong gains for culturally-specific and morphologically complex translations.

1 Introduction

Sanskrit, the ancient liturgical language of Brahmanism and Hinduism and one of the world's oldest documented languages, presents formidable challenges for machine translation due to its complex morphological system, rich literary heritage, and profound cultural context. Vedic Sanskrit, the earliest attested form preserved in the Vedas, poses additional difficulties through its archaic vocabulary, distinctive accent system, and highly ritualistic contexts that require deep linguistic, cultural, and religious understanding for accurate interpretation.

While recent advances in large language models have demonstrated remarkable capabilities for machine translation across numerous language pairs, their application to low-resource ancient languages remains limited. Existing approaches typically rely on parallel corpora alone, overlooking the wealth of scholarly expertise encoded in philological commentaries that provide essential linguistic, cultural, and religious context.

This paper introduces a methodology that leverages scholarly philological commentary to enhance neural machine translation of Vedic Sanskrit. Our approach recognizes that successful translation of ancient sacred texts requires not merely linguistic competence, but also deep understanding of historical, cultural, and ritualistic contexts –knowledge traditionally preserved in centuries of scholarly commentary.

We make the following contributions: (1) We demonstrate that integrating philological commentary significantly improves Vedic Sanskrit translation quality across multiple large language models; (2) We provide comprehensive evaluation across five diverse architectures, including both general-purpose and Sanskrit-specialized models; (3) We establish a replicable methodology for incorporating scholarly expertise into neural translation of ancient languages.

2 Vedas and Philological Commentary

The Vedas are the oldest sacred texts in the Indian subcontinent, composed in Vedic Sanskrit between about 1500 and 500 BCE. The Ṛgveda (ṚV), the oldest of the four Vedas, is a collection of hymns dedicated to various deities. All hymns in the Ṛgveda comprise complex verses which are often difficult to translate, making it difficult even to identify the correct meaning of individual words. The constraints of metrical structure and the use of archaic phonology, morphonology, and syntax further complicate the translation process.

Vedic texts include not only verses, but also explanatory prose. Saṃhitās, the collections of verses in the Vedas including the Ṛgveda, are highly ritualistic and sophisticated texts, where each hymn is composed with a specific purpose and context. Brāhmaṇas, Āraṇyakas, and Upaniṣads are prose texts that provide detailed explanations of the Saṃhitās about their meanings, their ritual-

istic significance, and their philosophical contexts.

The translations of Vedic texts have been published for centuries, some of which have commentaries that provide essential philological and linguistic insights. These commentaries are crucial for understanding the complex meanings of the hymns. Even one Saṃhitā, the Ṛgveda, has been translated by multiple scholars (Grassmann, 1876; Geldner, 1951; Renou, 1955; Elizarenkova, 1999; Jamison and Brereton, 2014; Witzel et al., 2007, 2013; Dōyama and Gotō, 2022). These academic translations themselves are significant works of philology and linguistics. In addition, the commentaries present the basis of the translators' interpretation. For example, ṚV 8.5.22 [1] is translated and added to the commentary by Jamison and Brereton (2014) as below:

Translation:

> When did the son of Tugra, abandoned in the sea, do reverence to you, o men, so that your chariot would fly with its birds?"

Commentary:

> The subjunctive pátāt seems to be used in an unusual past prospective sense in this mythological context. This may be an English problem, however. Since the verb of the main clause is injunc. vidhat, this context is not necessarily preterital, but "timeless," and the subjunctive can therefore be expressing pure future modality. The fact that the next verse is also mythological and contains an undoubted present tense form daśasyathaḥ shows that mythological tense is fluid here. Re remarks (ad vs. 23) that the indifference between present and preterite underlines the reflection of the current human situation in the legendary material.

When human experts read closely the original Vedic text, they refer to these commentaries to understand the meaning of the hymns and the thoughts of the translators. Due to the limited number of syllables and the need to maintain a regular rhythm, the word order in Vedic verses is often complex and many elements are omitted, making interpretation difficult. Therefore commentaries represent the invaluable insights of earlier scholars.

[1]The source text is: *kadā vāṁ taugriyó vidhat samudré jahitó narā yád vāṁ rátho víbhiṣ pátāt*.

3 Related Works

Recent research in neural machine translation has increasingly recognized the value of incorporating linguistic annotations and contextual information to improve translation quality, particularly for morphologically rich and low-resource languages. This section reviews key developments in annotation-enhanced machine translation, with particular attention to approaches relevant to our work on ancient language processing.

3.1 Morphological and Syntactic Annotations

Early work by Sennrich and Haddow (2016) demonstrated that incorporating explicit linguistic features such as POS tags and morphological information as input features to neural machine translation systems yields improvements in translation quality. Their experiments on English-German and English-Romanian showed lower perplexity and higher BLEU scores compared to word-only baselines, establishing that explicit linguistic annotations provide complementary information to end-to-end neural approaches.

García-Martínez et al. (2016) introduced Factored Neural Machine Translation (FNMT), which generates multiple outputs for each word including lemmas and morphological attributes. This approach proved particularly effective for morphologically rich languages, reducing vocabulary size and handling unknown words more effectively. Similarly, Dalvi et al. (2017) showed that injecting target-language morphological information into the decoder through joint training improved translation accuracy by 0.2-0.6 BLEU points for German and Czech.

For classical languages, Rosenthal (2023) demonstrated the particular value of morphological annotations in low-resource scenarios. Working with Latin-English translation, they achieved a BLEU score of 22.4 by encoding Latin morphology through stem-morpheme splitting, exceeding Google Translate's performance by over 4 BLEU points. This work is especially relevant to our Sanskrit study, as both Latin and Sanskrit are highly inflected classical languages with complex morphological systems.

3.2 Semantic Annotations

The integration of semantic role labeling (SRL) into neural MT has shown promise for preserving meaning relationships. Marcheggiani et al. (2018) were among the first to incorporate predicate-argument structures using Graph Convolutional Networks, achieving improvements from 23.3 to 24.5 BLEU on English-German translation. Their approach encoded PropBank semantic roles as graphs, demonstrating better preservation of "who-did-what-to-whom" relationships.

Rapp (2022) extended this work by annotating Europarl data with semantic roles across multiple language pairs (English to French, German, Greek, and Spanish), showing consistent but modest BLEU improvements of 0.2-0.5 points. While these gains appear small, they were consistent across different runs and language pairs, suggesting that semantic role annotations help capture meaning nuances that purely sequence-based models might miss.

In low-resource settings, semantic annotations have shown larger relative impacts. Wu et al. (2021) reported an average +1.18 BLEU improvement when injecting predicate-argument labels for Chinese, Mongolian, Uyghur, and Tibetan translations, while Nguyen et al. (2020) leveraged Abstract Meaning Representation (AMR) graphs for English-Vietnamese translation in small-data scenarios.

3.3 Cultural and Contextual Annotations

Recent work has explored incorporating cultural and historical knowledge into translation systems. Conia et al. (2024) introduced KG-MT, which retrieves entries from multilingual knowledge graphs to handle culturally nuanced entity references. Their approach achieved dramatic improvements—129% relative BLEU improvement over strong multilingual models and 62% over GPT-4 on culturally sensitive segments—demonstrating the importance of background knowledge for accurate translation.

This line of research is particularly relevant to our work on Vedic Sanskrit, where cultural, ritualistic, and historical contexts are crucial for accurate interpretation. The annotations provided by scholarly works like Jamison & Brereton's Rig Veda translation contain precisely this type of contextual information that has proven valuable in recent MT research.

3.4 Discourse and Pragmatic Annotations

Document-level translation has benefited from discourse structure annotations. Tan et al. (2022) incorporated Rhetorical Structure Theory (RST) annotations to improve coherence and consistency across sentences, achieving +0.9 BLEU and +1.1 METEOR improvements over strong document-level baselines. Style and register control through annotations has also proven effective, with **?** demonstrating that politeness tags can significantly improve translation appropriateness in German output.

Our work builds on this foundation by applying commentary-enhanced training specifically to Vedic Sanskrit, where philological expertise is essential for accurate translation. Unlike previous work that primarily used automatically generated annotations, we leverage expert scholarly commentary that provides deep cultural and linguistic insights unavailable to automatic annotation tools.

4 Methodology

4.1 Dataset

Our dataset consists of the Ṛgveda, utilizing the comprehensive English translation by Jamison & Brereton (Jamison and Brereton, 2014) and their detailed scholarly commentaries (Jamison and Brereton, 2015). The translation represents the most authoritative modern complete English translation of the Ṛgveda, while the commentary provides essential philological and linguistic insights that are being progressively published on their website[2].

For the source texts of the Vedic Sanskrit, we employ the electronic text available through TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) (Martínez García and Gippert, 1995), which is based on Aufrecht's critical edition (Aufrecht, 1877). The transliteration of the text is converted following the IAST transliteration standard (Royal Asiatic Society of Great Britain and Ireland, 1896). This choice of romanization scheme is motivated by consistency with the English translation and commentary, where Sanskrit words and phrases are consistently referenced using this same IAST transliteration.

For the translation and commentary sources, the English translations are from the published Jamison & Brereton book, providing complete coverage of all 1,028 hymns and over 10,000 verses of

[2] http://rigvedacommentary.alc.ucla.edu/.

the Ṛgveda. The philological commentary is obtained from the authors' website.

The commentary encompasses multiple types of philological information essential for accurate translation: (1) Linguistic explanations including morphological analysis, etymological derivations, and syntactic parsing of complex constructions; (2) Lexical annotations providing semantic clarification of rare or polysemous terms, often with cross-references to other Vedic texts; (3) Cultural and ritualistic explanations elucidating the religious, social, and historical contexts necessary for proper interpretation; (4) Verse-level commentary addressing overall meaning, poetic structure, and intertextual relationships; (5) Background information on mythological references, deity characteristics, and ceremonial practices mentioned in the hymns.

Some hymns contain only commentary for the entire hymn and no verse-level commentary. Although hymn-level commentary provides valuable context for interpreting entire hymns, we restrict our dataset to verse-level commentary only. This decision is motivated by the need to manage input context length for language models, ensuring that each training sample remains within feasible token limits. By focusing on verse-level commentary, we maximize the amount of philological information available for each verse while maintaining compatibility with model context constraints.

This multi-layered commentary structure provides the rich contextual information that distinguishes our approach from previous work relying solely on automatically generated linguistic features. The scholarly commentary represents decades of expert philological analysis, offering insights unavailable to computational annotation tools.

Our final dataset comprises 6,754 total samples split into training (5,282 samples, 78.2%), validation (743 samples, 11.0%), and test (729 samples, 10.8%) sets. Figure 1 provides a box plot visualization of text lengths, highlighting the substantial differences in length and variability between the three text types. The Sanskrit source texts are relatively uniform in length, averaging 117.7 characters (median 129.0, std 27.8) in the training set, with similar statistics across validation and test splits. English translations are consistently longer than their Sanskrit counterparts, averaging 192.6 characters (median 199.0, std 54.6), reflecting the approximately 1.6× expansion typical when translating from Sanskrit to English due to morphological differences and explanatory additions required for clarity.

The philological commentaries show substantially greater length and variability, averaging 1,114.0 characters (median 735.0, std 1,189.4) in the training set, with some commentaries extending beyond 11,000 characters. This high variability reflects the scholarly practice of providing detailed analysis for particularly complex or significant verses while offering more concise notes for straightforward passages. The distribution is right-skewed, indicating that while most commentaries are relatively brief, a significant subset provides extensive philological analysis that substantially exceeds the length of the source texts themselves.

To quantify the lexical overlap between reference and commentary texts, we computed the n-gram Jaccard coefficients (for n = 1, 2, 3, 4). The mean Jaccard coefficients are 0.082, 0.018, 0.0075, 0.0038 for n = 1, 2, 3, 4 respectively. Figure 2 shows the box plots of the n-gram Jaccard coefficients. These low overlap values confirm that the contamination risk from commentary to reference translations is minimal, indicating that the commentary provides largely complementary information rather than redundant content.

We additionally conducted a content characterization of the commentary by randomly sampling 1,000 commentaries and assigning one or more labels from five categories –linguistic, cultural, religious, philosophical, and others– allowing multi-label assignments per commentary The resulting distribution (not mutually exclusive) shows a strong predominance of linguistic content (73.4%), followed by religious (32.6%), philosophical (15.0%), cultural (12.6%), and others (4.0%). At the entry level, 54.4% of commentaries received a single label, while 45.6% were multi-labeled, indicating that nearly half of the commentary have integrated information from multiple aspects.

4.2 Model Selection

We evaluate our approach across five diverse large language models to assess generalizability across different architectures, parameter scales, and training paradigms: **GPT-4.1 nano**[3], **Gemini 2.5**

[3] https://platform.openai.com/docs/models/gpt-4.1-nano

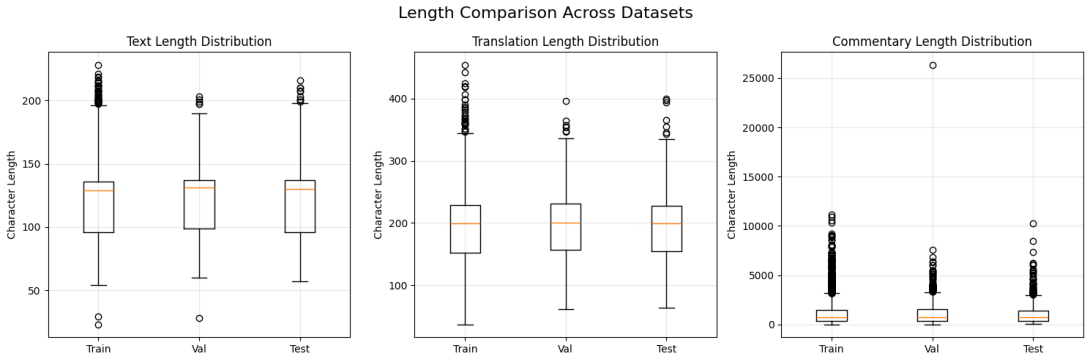


Figure 1: Box plots showing the distribution statistics for text lengths across the three text types. Commentary texts exhibit significantly higher variance and longer tails compared to source texts and translations.

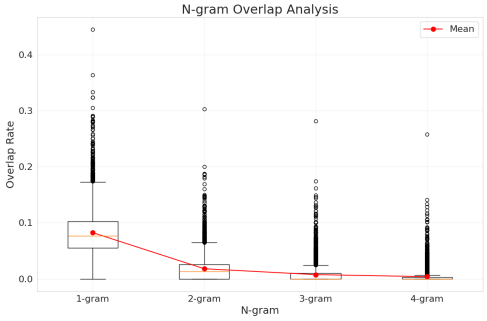


Figure 2: Box plots of n-gram Jaccard coefficients between reference and commentary texts.

flash[4], **Gemma 2 Mitra**[5], **Llama 3.1 8B**[6], and **Llama 3.2 3B**[7]. Gemma 2 Mitra is a specialized model for Sanskrit, developed under the MITRA project and built on Google's Gemma 2 foundation. It is trained on 7B tokens from diverse Buddhist texts, including Sanskrit, Tibetan, English, and Pāli.

4.3 Training Methodology

We formulate the translation enhancement task as conditional text generation, comparing two training conditions:

Standard SFT: Models are fine-tuned using Vedic texts as input to generate English translations.

Commentary-Enhanced SFT: Models are fine-tuned using both Vedic texts and scholarly commentaries in English as input to generate English translations.

Commentary-only SFT: Models are fine-tuned using only the commentaries as input to generate English translations.

For all models, we employed supervised fine-tuning with hyperparameter selection to balance learning efficiency with preservation of pre-trained capabilities. Both proprietary models (GPT-4.1 nano, Gemini 2.5 flash) and open-source models (Llama, Gemma/Mitra) were fine-tuned using platform-appropriate optimization techniques. Detailed training configurations, hyperparameters, and input formatting specifications are provided in Appendix A.

4.4 Evaluation Metrics

Translation quality was assessed using BLEU and COMET scores. The BLEU socores were calculated with SacreBLEU for standardized and reproducible evaluation. We computed BLEU-1 through BLEU-4 scores to capture both local phrase-level and global sentence-level translation quality. COMET scores were calculated using the pre-trained wmt22-comet-da model.

5 Results

5.1 Main Results: BLEU and COMET Scores

Table 1 reports BLEU and COMET scores for all models under three experimental configurations. Across most settings, the integration of philological commentary leads to consistent improvements in both BLEU and COMET, indicating gains in lexical accuracy and semantic alignment. GPT-4.1 nano achieves the highest relative increase in COMET (+0.03), followed by Gemma 2 Mitra (+0.04), while Gemini 2.5 flash maintains overall

[4] https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash

[5] https://huggingface.co/buddhist-nlp/gemma-2-mitra-it

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Table 1: BLEU and COMET scores across pretrained models and dataset configurations (src = source text, com = commentary).

Train Dataset	Test Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	COMET
GPT-4.1 nano						
src+com	src+com	44.0	17.0	7.9	4.2	0.61
src	src	26.2	8.7	3.5	1.6	0.58
src	src+com	26.4	8.7	3.5	1.7	0.59
com	src	20.8	4.7	1.2	0.39	0.52
Gemini 2.5 flash						
src+com	src+com	64.1	35.6	22.0	14.4	0.60
src	src	60.0	31.2	18.2	11.3	0.67
src	src+com	61.4	35.8	21.7	13.6	0.69
com	src	50.9	22.6	11.0	5.1	0.62
Gemma 2 Mitra						
src+com	src+com	49.8	21.4	11.2	6.5	0.63
src	src	42.9	17.5	8.8	4.9	0.59
src	src+com	28.5	9.2	4.2	2.3	0.52
com	src	32.3	10.6	4.2	1.8	0.56
Llama 3.1 8B						
src+com	src+com	48.9	21.8	12.0	7.3	0.62
src	src	47.5	19.4	9.7	5.6	0.63
src	src+com	40.5	17.5	9.2	5.4	0.61
com	src	32.8	8.5	2.8	1.2	0.56
Llama 3.2 3B						
src+com	src+com	41.1	14.5	6.7	3.5	0.59
src	src	44.1	16.0	7.5	3.9	0.60
src	src+com	36.0	12.3	5.5	2.8	0.57
com	src	26.9	4.6	1.0	0.27	0.52

strong performance with COMET values ranging from 0.60 to 0.69. In contrast, Llama 3.2 3B shows a marginal decrease in both BLEU and COMET, suggesting limited benefit from commentary augmentation at smaller model scales.

Table 2 presents the improvements in BLEU-4 scores for each model when using Commentary-Enhanced SFT. GPT-4.1 nano shows the most significant improvement of +155.0%, while Gemini 2.5 flash follows with +45.0%. Large open-source models (Gemma 2 Mitra, Llama 3.1 8B) also demonstrate substantial gains of +30.5% and +31.1% respectively. Llama 3.2 3B, however, shows a decrease of -11.5% in BLEU-4 score with commentary integration, confirming the same trend observed above.

Table 2: Commentary-Enhanced SFT improvement in BLEU-4 scores

Model	Improvement (%)
GPT-4.1 nano	+155.0
Gemini 2.5 flash	+27.4
Gemma 2 Mitra	+30.5
Llama 3.1 8B	+31.1
Llama 3.2 3B	-11.5

5.2 Commentary-Enhanced SFT improvement

We also examine whether the length of the philological commentary (word count) relates to sentence-level BLEU improvement. As visualized in Figure 3, we observe no consistent positive or negative relationship across models: both short and long commentaries can yield gains. This sug-

gests that commentary length per se is not the primary driver of the observed improvements.

5.3 Qualitative Analysis

Our results confirm that philological commentary provides valuable contextual information that significantly enhances Vedic Sanskrit translation quality across diverse model architectures. The consistent improvements observed across four large models –ranging from lightweight 8B parameter models to sophisticated proprietary systems– suggest that the benefits of scholarly commentary integration are robust and generalizable.

Linguistic Insights The complex morphological system in Vedic Sanskrit presents considerable challenges for translation. Especially the language in the Ṛgveda is highly inflected with a slightly different morphological structure compared to later Vedic languages. Philological commentaries frequently provide detailed linguistic explanations: phonological changes, morphological parsing, identifying stems, and inflection patterns. Semantic explanations of ambiguous terms are also provided, enabling models to generate contextually appropriate translations rather than defaulting to the most common or literal meanings.

Cultural Context Most significantly, philological commentary supplies cultural knowledge essential for accurate translation that is unavailable to models trained solely on linguistic data. As shown in Appendix B, the top-5 BLEU improvement examples demonstrate that the linguistic and cultural context in the commentary enables models to better understand and translate Vedic hymns. Vedic hymns assume deep familiarity with Indo-Aryan religious practices, mythological narratives, and social structures. Commentary explicates these assumed contexts, enabling models to generate translations that capture not merely linguistic content but cultural significance.

Model Performance Variations The substantial performance differences between models reveal varying sensitivities to commentary integration. GPT-4.1 nano shows a remarkable improvement, which suggests that smaller, efficient models can particularly benefit from contextual information, potentially compensating for limited parameter capacity through enhanced input quality. By contrast, the performance degradation in Llama 3.2 3B indicates a threshold effect where models below a certain capacity may struggle to effectively utilize complex commentary information.

BLEU Score Interpretation The consistent improvements in BLEU scores across successful models suggest that commentary integration particularly enhances phrase-level and sentence-level coherence rather than merely improving word-level accuracy. This pattern aligns with the nature of philological commentary, which provides contextual and structural insights that facilitate more coherent target language generation.

Commentary Length and BLEU Improvements The scatter plots in Figure 3 indicate little to no relationship between commentary word count and sentence-level BLEU gains. In practice, both very short and very long commentaries can be beneficial, and performance does not systematically increase with length. We hypothesize that the key factor is information quality –e.g., explicit morphological analyses, disambiguation cues, and culturally specific background– rather than sheer volume. For smaller models, excessively long inputs may even strain context capacity, further weakening any length-gain association.

Implications for Ancient Language Translation Our findings have broader implications for computational approaches to ancient languages. The successful integration of philological commentary into neural machine translation suggests that traditional philological methods remain highly relevant in the age of AI. This commentary enhanced methodology could be extended to other underresourced ancient languages where scholarly traditions provide rich contextual frameworks.

Dataset Limitations Our approach, however, faces several methodological limitations. Notably, there exist authoritative German translations such as Geldner's complete work(Geldner, 1951) and the ongoing series(Witzel et al., 2007, 2013; Dōyama and Gotō, 2022), which represent significant contributions to Vedic translation studies. Comparative analysis with these major translations is essential for a comprehensive evaluation of translation quality.

Scalability Limitations Another important limitation concerns scalability. The integration of extensive philological commentary substantially increases input length and complexity, which can challenge both model context windows and com-

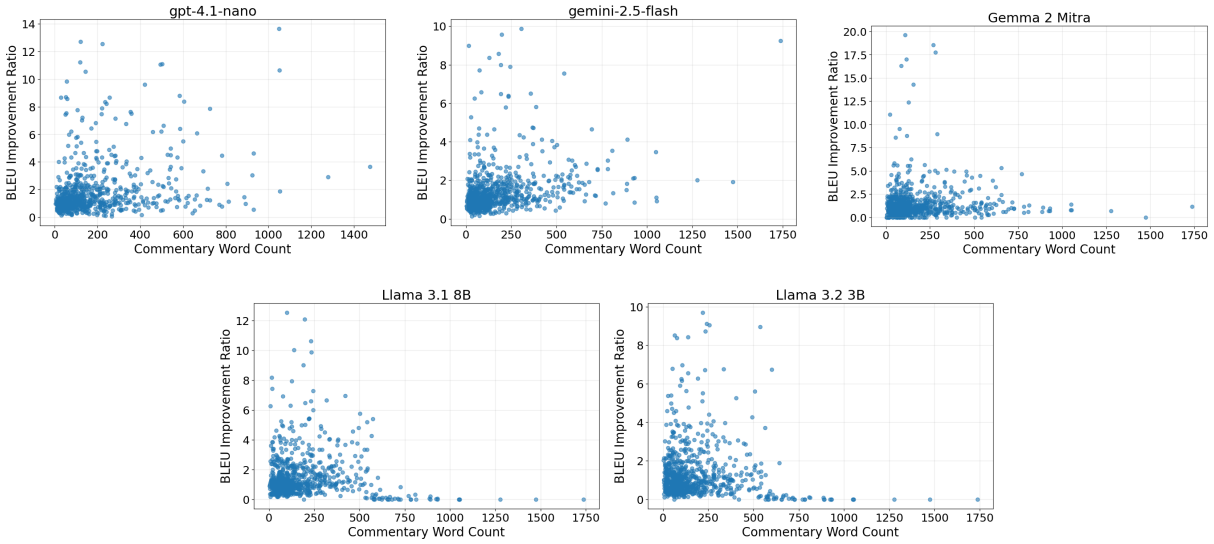


Figure 3: Scatter plots of sentence-level BLEU improvement versus commentary length (word count), grouped by model. Each point represents a verse.

putational resources. As commentary texts often exceed the length of source verses by an order of magnitude, scaling this approach to larger corpora or more comprehensive commentarial traditions may require advanced context management strategies, such as hierarchical encoding or retrieval-augmented generation.

6 Conclusion

This study demonstrates that incorporating scholarly philological commentary significantly improves Vedic Sanskrit translation accuracy across multiple large language model architectures. Our comprehensive evaluation across five diverse models confirms the robustness and generalizability of commentary-enhanced training, with consistent BLEU score improvements observed although this sensitivity varies with model size.

The methodology presented here establishes a replicable framework for leveraging scholarly expertise in neural machine translation of ancient languages. By demonstrating that traditional philological knowledge can effectively enhance modern AI capabilities, this work opens new avenues for computational approaches to ancient language processing and highlights the continued relevance of scholarly commentary in the digital age.

References

Theodor Aufrecht. 1877. *Die Hymnen des Rigveda*. Bonn: Adolph Marcus.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Eijirō Dōyama and Toshifumi Gotō. 2022. *Rig-Veda: das heilige Wissen: sechster und siebter Liederkreis*, 1. aufl edition. Verl. der Weltreligionen, Berlin.

Tat'jana Jakovrevna Elizarenkova. 1999. *Rigveda*. Number 3 in Literaturnye pamjatniki. Nauka.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Karl F. Geldner. 1951. *Der Rig-Veda : aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen*. Number v. 33-36 in Harvard oriental series. Harvard University Press , Oxford University Press , Otto Harrassowitz, Cambridge, Mass. , London , Leipzig.

Hermann Grassmann. 1876. *Rig-Veda : übersetzt und mit kritischen und erläuternden anmerkungen versehen von Hermann Grassmann*. Brockhaus, Leipzig.

Stephanie W. Jamison and Joel P. Brereton. 2014. *The Rigveda: The earliest religious poetry of India*. Oxford University Press, New York.

Stephanie W. Jamison and Joel P. Brereton. 2015. Rigveda translation: Commentary. Center for Digital Humanities, University of California, Los Angeles.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

Francisco Javier Martínez García and Jost Gippert. 1995. Thesaurus indogermanischer text- und sprachmaterialien.

Long H. B. Nguyen, Viet Pham, and Dien Dinh. 2020. Integrating amr to neural machine translation using graph attention networks. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 158–162.

Reinhard Rapp. 2022. Using semantic role labeling to improve neural machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3079–3083, Marseille, France. European Language Resources Association.

Louis Renou. 1955. *Études védiques et pāṇinéennes*. Number sér. in-8o ; fasc. 1-2, 4, 6, 9-10, 12, 14, 16-18, 20, 22-23, 26-27, 30 in Publications de l'Institut de civilisation indienne. E. de Boccard, Paris.

Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language.

Royal Asiatic Society of Great Britain and Ireland. 1896. *Transliteration report*. London : The Society.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022. Towards Discourse-Aware Document-Level Neural Machine Translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4383–4389. International Joint Conferences on Artificial Intelligence Organization.

Michael Witzel, Toshifumi Gotō, Eijirō Dōyama, and Mislav Ježić. 2007. *Rig-Veda : das heilige Wissen Erster und zweiter Liederkreis*. Verlag der Weltreligionen, Frankfurt am Main.

Michael Witzel, Toshifumi Gotō, and Salvatore Scarlata. 2013. *Rig-Veda : das heilige Wissen Dritter bis fünfter Liederkreis*. Verlag der Weltreligionen, Frankfurt am Main.

Nier Wu, Hongxu Hou, Haoran Li, Xin Chang, and Xiaoning Jia. 2021. Semantic Perception-Oriented Low-Resource Neural Machine Translation. In *Machine Translation*, pages 51–62, Singapore. Springer Singapore.

A Training Configuration Details

This section provides comprehensive details on the training configurations, hyperparameters, and input formatting used in our experiments.

A.1 Input Format Specification

The input format for Commentary-Enhanced SFT follows this structure:

```
# Input:
[Original Vedic Sanskrit text]
# Commentary:
[Scholarly philological commentary]
# Translation:
[Target English translation]
```

For Standard SFT, the format is simplified to:

```
# Input:
[Original Vedic Sanskrit text]
# Translation:
[Target English translation]
```

The input format is for training, and the input in inference does not contain the "[Target English translation]" line.

A.2 Proprietary Model Hyperparameters

GPT-4.1 nano Fine-tuning Configuration

- Training epochs: 3
- Batch size: 10
- Learning rate multiplier: 0.1
- Validation split: 10%
- Platform: OpenAI Fine-tuning API

Gemini 2.5 flash Fine-tuning Configuration

- Training epochs: 2
- Learning rate multiplier: 5
- Adapter size: 4
- Platform: Google Vertex AI
- Features: Dynamic batch sizing, automatic optimization

A.3 Open-Source Model Hyperparameters

For Llama and Gemma models, we employed Low-Rank Adaptation (LoRA) for computational efficiency:

LoRA Configuration

- Rank (r): 16
- Alpha: 16
- Dropout: 0.0
- Target modules: All attention and feed-forward layers

Training Parameters

- Learning rate: 3e-4 with linear decay
- Batch size: 8 with gradient accumulation
- Training epochs: 40 with early stopping
- Optimizer: AdamW with 8-bit precision
- Weight decay: 0.01
- Warmup steps: 100
- Gradient clipping: 1.0

B Top-5 Examples with the Largest BLEU-4 Improvements

This appendix presents the five test-set instances with the largest *absolute* gains in BLEU-4 when philological commentary is provided to the model GPT-4.1 nano. For each example we show: the verse number (book, hymn, and verse numbers), the gold/reference translation, the baseline translation *without* commentary, and the translation *with* commentary, together with BLEU-4 before/after.

1. **BLEU-4 gain:** 51.0270 (2.6316 → 53.6585)
   Verse number: ṚV 6.2.7

   Reference translation (gold)
   For now you are to be reverently invoked among the clans, as our dear guest, to be protected like a son who brings delight to the home, (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

   Translation without commentary (BLEU-4: 2.6316)
   For here you are, o widely spread (flame), the fire to be kindled, dear guest to us, delighting like an old man in his fortress, like a son who is to be protected.

   Translation with commentary (BLEU-4: 53.6585)
   For you are eager for battle, dear guest; you are to be protected like a son who brings delight (to the home), (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

2. **BLEU-4 gain:** 33.0282 (1.7544 → 34.7826)
   Verse number: ṚV 9.73.1

   Reference translation (gold)
   In the jaw (of the pressing stones) while the droplet was blowing its blast, they sounded in unison. The ties of lineage have joined together in the womb of truth. The lord [=Soma] has made his three heads to be seized. The boats of the trusty one [=Soma] have brought the good (ritual) performer across to the other shore.

   Translation without commentary (BLEU-4: 1.7544)
   Streams of soma roar as they cross the filter together, the womb of truth together with the navel of truth, the three heads of the lord of the night (being) joined together. The boat of truth, well made, has been set in motion. The trusty boat has been set in motion.

   Translation with commentary (BLEU-4: 34.7826)
   The streams of soma roar together as they cross the filter, and the ties of lineage have joined together in the womb of truth. The demon made his three heads to be seized. The boat of truth, well made, has been set forth.

3. **BLEU-4 gain:** 26.4069 (19.0476 → 45.4545)
   Verse number: ṚV 8.75.11

   Reference translation (gold)
   Surely you will toil for wealth for us, for our quest for cattle, Agni? Make a wide (way) for us, you wide-maker.

   Translation without commentary (BLEU-4: 19.0476)
   Make for us a dwelling place, Agni, for our quest for cattle, for wealth. Make us strong, strong.

   Translation with commentary (BLEU-4: 45.4545)
   O Agni, you will toil for wealth for us, for our quest for cattle. Make it broad and broad for us.

4. **BLEU-4 gain:** 25.2519 (3.3195 → 28.5714)
   Verse number: ṚV 9.65.8

   Reference translation (gold)
   Him whose tawny hue, dripping with honey, they impel (like) a fallow bay with the stones, the drop for Indra to drink,

   Translation without commentary (BLEU-4: 3.3195)
   The tawny color of which they impel the honeyed (soma) with the stones, the tawny (color) they impel for Indra to drink, – Commentary: The first hemistich is somewhat tricky. The relative clause yāsyā (whose) must refer to soma, with the relative clause hanging off (the first hemistich) (see v. 7). In pāda a, the relative yāsyā qualifies the color (vārṇam), which serves as the object of impel (hinvanti) in b. "They impel his color" is a slightly odd locution, and it becomes odder in b, with the accusative hārim. The stem hār- is a color term and could qualify vārṇam ("tawny color"), but it also is regularly applied directly to soma (e.g., in vss. 12 and 25 of this hymn) and also identifies Indra's horse(s). Ge (note 8b) cleverly suggests that hārī- is to be read twice, with the second reading an unmarked simile referring to a horse of the appropriate color. This provides a more appropriate object for impel ("they impel (like) a fallow bay (horse)"), while connecting the putative horse with the "tawny color" (vārṇam … hārim) that is the first object of the verb.

   Translation with commentary (BLEU-4: 28.5714)
   The tawny color of which they impel (like) a fallow bay (horse) with stones, for Indra to drink.

5. **BLEU-4 gain:** 24.2857 (2.3810 → 26.6667)
   Verse number: ṚV 8.21.13

Reference translation (gold)
You are without rival, but by the same token, without friend, Indra, by birth and from of old. Only in battle do you seek friendship.

Translation without commentary (BLEU-4: 2.3810)
Nephew, rival, you are not, Indra. You have been from of old, and you seek to be friend.

Translation with commentary (BLEU-4: 26.6667)
You are without rival, Indra, by the same token, without friend, but from of old you have been there. You will fight if you want to fight.

Lore Coherent Encounter Generation for Dungeons and Dragons: LLM Fine-tuning and Benchmark

Aravinth Sivaganeshan, Nisansa de Silva, Akila Peiris
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{sivaganeshan.22, NisansaDdS, akila.21}@cse.mrt.ac.lk

Abstract

Swathes of tasks that were erstwhile handled by other deep learning models are being taken over by Large Language Models (LLMs). While they may demonstrate reasonable results in the zero-shot configuration for most domains, in the contexts of more niche or esoteric domains, instruction tuning them on the domain at hand has shown to be effective and sometimes necessary. Dungeons & Dragons (D&D) is the most commercially successful Tabletop Role Playing Game (TTRPG) with its own unique lore mostly set in the fantasy domain. The players are confronted with fantastical *monsters* in mathematically balanced *encounters* overcoming which, contribute to the calculation of player progress. Even with the plethora of information available for D&D (or perhaps rather due to its abundance), the generation of an encounter that is coherent with the lore is a time-consuming and difficult process as there are no support tools available for the selection of coherent monsters. Recognizing this gap, we instruction-tuned a Mistral-based LLM that can function as an assistant on this matter, using instructions generated from publicly available D&D datasets. Next, we conducted a number of prompt engineering experiments on the trained LLM, such that the output from the LLM would be a list of coherent monsters when a candidate monster is given in the input. The generated outputs were examined for the coherency of lore, theme, and environment. It was observed that the outputs were partially or fully coherent with the lore in about 66.0% of the 241 candidate monsters tested.

1 Introduction

Language models trained on very large datasets are shown to have high capabilities (Brown et al., 2020; Chowdhery et al., 2022). However, these models are trained on datasets generated by humans with various goals (Rafailov et al., 2023). Therefore, the performance of LLMs might not be desirable for specific downstream tasks or specific domains. While pre-training LLMs entirely on a specific domain data is costly and resource intensive (Cottier et al., 2024), techniques such as instruction fine-tuning are quite useful in this regard.

Dungeons and Dragons (D&D) is a very popular open-ended, Table-Top, Role-Playing Game (TTRPG). It is commercially available since 1974 (Gygax and Arneson, 1974) and it is currently in its 5th edition (Crawford et al., 2014b). D&D has a set of predefined rules and there are several settings in D&D (Peiris and de Silva, 2022, 2023; Squire, 2007; Weerasundara and de Silva, 2024). Each setting has lore that describes the historical and current status of the game world[1].

In a D&D game, combat encounter is one of the most important component, through which players attain progress (Crawford et al., 2014a). In a combat encounter, players are pitted against domain specific entities called *monsters*. These monsters may be considered as either *bosses* or *minions* and are partially defined by their numerical statistics as shown in Figure 1 comparing two boss monsters (Mind Flayer[2] and Red Dragon[3]) and two minions (Intellect Devourer[4] and Kobold[5]).

The generated encounters need to align with the lore to preserve immersion and verisimilitude (Stern, 2002). As a generic example, a party going through a forest being attacked by a lion is an encounter that aligns with the lore. On the

[1]The word *World* is used as an encompassing term that may mean anything from a small region of land (Perkins et al., 2015) to a planet (Crawford et al., 2019) to a universe (Lee et al., 2019) or a multiverse (Arman et al., 2023) depending on the specific lore.

[2]https://www.dndbeyond.com/monsters/5195125-mind-flayer

[3]https://www.dndbeyond.com/monsters/5194875-adult-red-dragon

[4]https://www.dndbeyond.com/monsters/5195088-intellect-devourer

[5]https://www.dndbeyond.com/monsters/16939-kobold

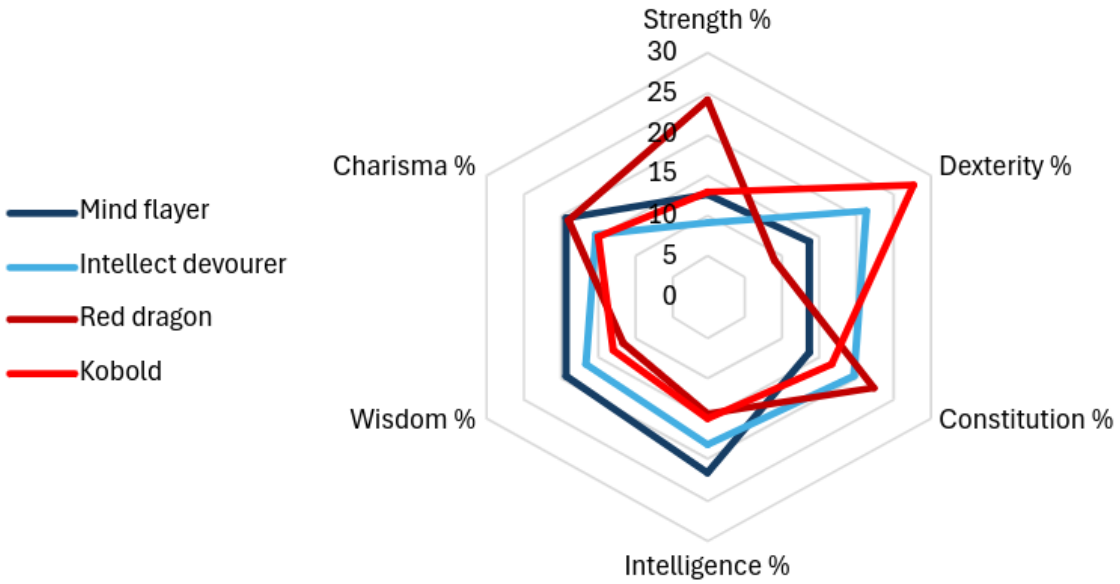


Figure 1: Comparison of creature statistics for two boss monsters (Mind Flayer and Red Dragon) and two minions (Intellect Devourer and Kobold)

other hand, a party going in a boat being attacked by a lion is an encounter that does not align with the lore. In D&D, one player enacts a special role named the *Dungeon Master* (DM) who takes the responsibility of being the lead story teller. Thus, the time-consuming task of selecting monsters for encounters according to the desired theme and difficulty is one of the responsibilities of the DM.

The *Challenge Rating* (CR) of a monster is an estimation of the threat posed by the particular monster and there are a few tools already available for calculating the difficulty of the encounter based on the challenge rating of the monsters to a particular party of players. These tools can help DM with the mathematical aspect of an encounter, but offers no help in the consistency of lore. In fact, there are no existing tools for the automatic generation of encounters with the consideration of the written lore. Currently, the DMs are expected to have near encyclopedic knowledge about the lore by themselves to make sure that aspect of the encounter is sound. Therefore, there is an existing necessity for a tool that can be used as a Dungeon master's assistant to select monsters according to the environment or to select monsters that are coherent with each other as described in the lore. Also, we analyzed the necessity by conducting a survey on reddit whether a tool with this functionality is preferable to DMs. Figure 2 shows the results of the survey. However, we had to end the poll early due to the AI related rules on the particular subreddit[6]. Therefore the poll at the time of closure had a total of 40 responses. According to the best of our knowledge, this is the first work on the specific problem. This work mainly focuses on generating coherent encounters consistent with the lore for 5th edition of D&D (5e). The language models that are currently used for the general domain cannot be used for a fantasy domain such as D&D given that a significant portion of jargon does not make sense in the general domain. Even when they do, the semantics of the words may be quite different (Peiris and de Silva, 2022). As a solution for this, we propose converting the problem of the abundance of the D&D lore into the solution itself by instruction tuning LLMs using automatically generated questions and answers from the lore documents.

For this study, we selected `Mistral7BInstruct v0.2` which is made by instruct tuning `Mistral 7B` (Jiang et al., 2023) as the base model and instruct tuned it with a set of domain specific datasets that we generated to obtain a set of models. After that, prompting experiments were done on all these models to select the best model and the best prompt. Then, we selected the final model with the best prompting technique to list the encounter and tested the best model extensively with 241 prompts. The LLM outputs from the best model for these 241 prompts were judged by 2 humans and 3 LLMs. All the links to our instruction-tuned models and

[6] https://www.reddit.com/mod/DnD/rules/

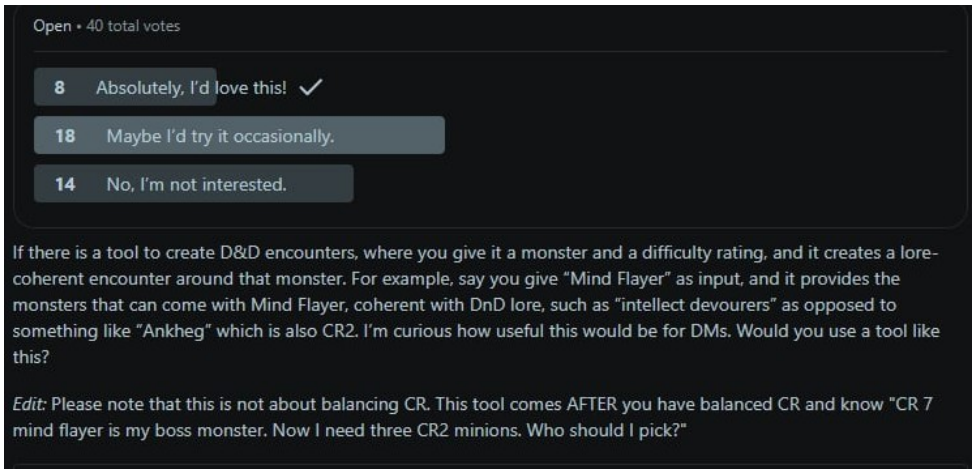


Figure 2: Reddit Survey taken for identifying whether the DMs would prefer a tool to generate encounters

the data used for instruction-tuning are given in Section 3.

2 Related Works

2.1 Existing Data Sets and Research on D&D

There is a considerable amount of official lore about D&D spread among copyrighted and non-copyrighted publications. The largest free and publicly accessible repository of relevant lore can be found on the Forgotten Realms Wiki[7] which is a Wikipedia-style collection of the lore written in the in-world perspective. For example, the article on *Mordenkainen* in ForgottenRealms wikia[8] starts as "Mordenkainen was a prolific archmage..." as opposed to the article[9] on Wikipedia for the same character which starts with "Mordenkainen is a fictional wizard...". Using the data from the wikia, Peiris and de Silva (2022) created the `FRW-dataset`[10] collection that includes 11 datasets at various types and levels on preprecessing applied to the data. This collection includes `FRW-alpaca.jsonl` which is a dataset of 41106 instructions and outputs in the alpaca format (Taori et al., 2023), which can be directly used for instruction tuning. This dataset contains instructions asking for descriptions as well as specific questions on events, places, monsters, and other concepts in D&D.

Monster statistics, which is not completely available on the *ForgottenRealms* wikia, can be accessed from `5etools`[11]. A full data dump of around 2079 monsters with their characteristics tabulated as a 32-column table can be downloaded as a csv file from `5etools`. This data can be used for generating instructions related to the mechanical aspect of the game.

Finally, the official source, `DnDbeyond`[12] contains a large amount of data. However, they do not provide direct access to any of the data (including the non-copyrighted data) in any form other than viewing on a web browser. Therefore, this source can only be used as a reference for our human experts to learn about the game and not as a data source.

There have been a number of studies using D&D data, especially in dialogue and discourse analysis (Rameshkumar and Bailey, 2020; Callison-Burch et al., 2022; Louis and Sutton, 2018). Further works have also explored the possibility of AI playing the game as DMs or players (Ellis and Hendler, 2017; Martin et al., 2018). There are some studies (Weerasundara and de Silva, 2023; Sivaganeshan and De Silva, 2023) that focus on the extended Named Entity Recognition (NER) task of identifying Dungeons and Dragons entities from text. Image (Weerasundara and de Silva, 2024) and adventure (Peiris and de Silva, 2022) generation are also aspects that have been explored for D&D. But, up to now, there are no existing works regarding the automated generation of lore consistent encounters for Dungeons and Dragons. The only available

[7] https://forgottenrealms.fandom.com

[8] https://forgottenrealms.fandom.com/wiki/Mordenkainen

[9] https://en.wikipedia.org/wiki/Mordenkainen

[10] https://huggingface.co/datasets/Akila/ForgottenRealmsWikiDataset

[11] https://5e.tools/

[12] https://www.dndbeyond.com/

tools [13] [14] [15] for encounter building only considers the mathematical aspect of the encounters.

2.2 Low Rank Adaptation

Low Rank Adaptation (LoRA) (Hu et al., 2021) is useful for fine-tuning large models to downstream tasks without updating all the parameters of the existing model. Also, this method creates a separate `adapter` for the base model for the particular application. This is useful for reducing the cost of fine-tuning and also different adapters can be made for different tasks for a single base model. Further, an extended method named `QLoRa` (Dettmers et al., 2024) can be used to reduce the memory requirements of fine-tuning by quantizing the pretrained model to 4 bits and adding a small set of LoRA weights that are tuned by backpropagating gradients through the quantized weights.

2.3 Instruction tuning LLMs for domain specific downstream tasks

Instruction tuning is a computationally effective process for adapting an LLM to a specific domain without extensive retraining or architectural changes (Zhang et al., 2023). In this technique, the LLMs are further trained using `(INSTRUCTION, OUTPUT)` pairs such that `INSTRUCTION` denotes human instruction to the model and `OUTPUT` denotes the expected output.

LoRA-based methods can be used to further improve the computational efficiency of instruction tuning. `Alpaca` (Taori et al., 2023) dataset format is a standard structure for instruction tuning which represents the instructions as a `JSON` array with objects containing `(INSTRUCTION,INPUT,OUTPUT)`. There are several tools and frameworks for this purpose where `Axolotl`[16] is one of easy-to-use tool.

2.4 Prompt Engineering

The *prompt* is the input provided to the LLM to obtain the output. Empirically, it is shown that better prompts lead to better outputs across different tasks (Wei et al., 2022; Liu et al., 2023). Due to the growth and widespread use of LLMs, prompting has become an emerging field. Text-based prompting can be generally divided into categories of: In-Context Learning, thought generation, decomposition, ensembling, and self-criticism (Schulhoff et al., 2024).

2.5 LLM-as-a-Judge

LLMs are a compelling alternative to traditional expert driven evaluations due to their ability to process diverse data types and provide scalable, flexible and consistent assessments (Gu et al., 2024). There are multiple works where LLMs replace human judges or used together with human judges for rapid, scalable evaluation (Ashktorab et al., 2024; Bavaresco et al., 2025; Tseng et al., 2024).

3 Methodology

This section provides an overview of the collection and preparation of instruction datasets, instruction tuning `MistralInstructv0.2`[17] with the instruction dataset, prompt engineering and evaluation of fine-tuned LLM outputs.

3.1 Building the Instruction Datasets

A proper instruction dataset is essential for adapting an LLM to the user objective and the D&D domain. We collected and processed data from numerous publicly available sources. First, 41106 instructions were obtained from `FRW-J-Alpaca.jsonl` discussed in section 2.1. Let us call this instruction data set `FRW-I`.

5et-I Instruction Dataset: As further discussed in section 2.1, in order to obtain information that is not included in Forgotten Realms Wiki (and thus not in `FRW-I`), we use the data export from `5etools`. This includes the data on feature columns such as environment, size, alignment, type, speed, strength, and also contains descriptions of several features such as traits, actions, bonus actions and lair actions. Given that these are also traits that are intrinsic to the given monsters in the D&D domain, these features may impact the decision of whether a set of monsters can come together in an encounter.

Linguistic diversity of instructions is seen to help models generalize better (Zhang et al., 2024). So, we formed questions in 3-5 linguistically different formats for each column. Next, when building an instruction entry, for each monster we picked a random format out of the different formats of questions we created. Thus, an instruction in the alpaca

[13] https://www.dndbeyond.com/encounter-builder
[14] https://www.aidedd.org/dnd-encounter/
[15] https://www.kassoon.com/dnd/5e/generate-encounter/
[16] https://github.com/axolotl-ai-cloud/axolotl
[17] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

```
1. Describe the regional effects of {}?
2. Explain the effects that {} can have in its region?
3. Describe the effects that {} can have in its surroundings.
4. Explain the impact that {} can have in its surroundings.
5. Describe the impact that {} can have in its surroundings.
```

Prompt Template 1: Example Question Formats for Regional Effects of a Monster

```
{
   "instruction": "What are the monsters that can be together
        with an Ancient Red Dragon in an encounter?",
   "input": "",
   "output": "Kobolds, Dragon Cultists, Red Dragon Wyrmlings"
}
```

Prompt Template 2: Example of an instruction built from an encounter

format was then generated with the question as the instruction, with empty input and the column value as the output. In a similar way instructions were created from questions such that the environment is in the instruction and all relevant monsters in the output. In total, this process yielded an instruction tuning dataset with a total of 27,959 instructions using the data on 21 columns. Let us call this instruction data set 5et-I.

5et-I-All Instruction Dataset: For comparison in fine-tuning, instead of selecting a random question format when building an instruction entry, we used all the question formats created for a column to each column and a monster pair. From this, 3-5 instructions were formed for a column, monster pair. This process yielded a total of 110,089 instructions. Let us call this instruction data set `5et-I-All`.

Enc-I Instruction Dataset: Associations between monsters is a main factor that impacts in deciding whether a set of monsters can come together in an encounter. Thus, mining association rules from human expert-created encounters and making use of them in building an instruction set could be considered as useful.

For this, a total of 3786 human expert-created encounters were extracted from publicly available data from numerous online sources. From this, a total of 2587 monster itemsets that can come together in an encounter were obtained.

Then, frequent itemset mining was done using `Apriori` algorithm on the above 2587 itemsets to extract association rules from the encounter itemsets. Starting from higher support and confidence, association rules were obtained, manually examined and then support and confidence were reduced to get more association rules. Finally, 1096 association rules were obtained when the support is 0.0003 and confidence is 0.2. In this process, support was even tested with the very low value of 0.0003. The reason for this is because the itemset list is not from a balanced list of encounters, given that we sourced it from only the publicly available encounters. Thus, it is not meaningful to reject encounters looking for a higher support threshold. On the other hand, considering the domain, threshold of 0.2 was found to be acceptable. With this process, a list of 1096 association rules was finalized. Building the instructions from this was reasonably similar to the process used earlier to build `5et-I` using different question templates and selecting a random one. Prompt Template 2 shows an example of an instruction built from an association rule. Let us call this instruction data set `Enc-I`.

Enc-I-All Instruction Dataset: A similar process to creating `Enc-I` was followed with the only difference of using all the question formats created instead of using a random format. From this process, an instruction dataset of 3288 instructions was created. Let us call this data set `Enc-I-All`.

Aggregated Instruction Datasets: The instruction datasets `FRW-I`, `5et-I`, and `Enc-I` were merged into a single file named `FRW-dnd-encounter dataset`[18] containing 70161 instructions in alpaca format and the instruction order was randomized to make the instruction dataset more suitable for training. Similarly, instruction datasets `FRW-I`, `5et-I-All` and `Enc-I-All` were merged into a single file named `FRW-dnd-encounter-all dataset`[19] with 154,483 instructions and the instruction order was randomized. Further, based on the results after trying different prompts, it was found that role prompting worked better in comparison to the other prompt formats. Based on that, another dataset was created from `FRW-dnd-encounter dataset`. Every instruction in the dataset was prefaced with

[18] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools.jsonl

[19] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_all.jsonl

```
{
  "instruction": "You are a D&D expert. Provide an answer to
       the following question.",
  "input": "To which types of damage is an Aboleth Spawn
       resistant?",
  "output": "psychic damage"
}
```

Prompt Template 3: Example of an instruction in the FRW-dnd-encounter-role dataset

"*You are a D&D expert. Provide an answer to the following question*". The questions which were previously added as "*instruction*" were now changed to "*input*". By doing this, a new dataset was created named `FRW-dnd-encounter-role dataset`[20]. Prompt Template 3 shows an example of an instruction in this dataset. Similarly, the `FRW-dnd-encounter-role-all dataset`[21] is created by modifying `FRW-dnd-encounter-all dataset`.

3.2 Training the model

`Mistral7BInstructv0.2`[22] was taken as the base model to be fine-tuned with the D&D instruction dataset. This is due to the fact that the Mistral-Instruct model (Jiang et al., 2023) with 7 billion parameters, has been shown to be efficient and having comparable performance to the state of the art 13B parameter chat models (Jiang et al., 2023). Also, it has been shown to work well with `LORA` (Fujiwara et al., 2024). `Quantized LORA (QLORA)` is utilized to reduce the computing resources used for training. The training was done using `Axolotl` framework in an H100 2XM GPU cloud virtual machine from runpod.io[23] with 80 GB VRAM. 8 models were obtained by finetuning the base model, `Mistral7BInstructv0.2` with the generated instruction datasets. `Model1v0.1`[24] and `Model1v0.2`[25] were obtained by finetuning with instruction dataset `FRW-dnd-encounter` for 1 and 2 epochs respectively. Similarly, another 6 models were obtained by finetuning with the generated datasets for 1 and 2 epochs. `Model2v0.1`[26] and `Model2v0.2`[27] were obtained by finetuning with `FRW-dnd-encounter-role`. `Model3v0.1`[28] and `Model3v0.2`[29] were obtained by finetuning with the `FRW-dnd-encounter-all dataset`.`Model4v0.1`[30] and `Model4v0.2`[31] were from finetuning with the `FRW-dnd-encounter-role-all dataset`. QLoRa training configurations for fine-tuning every model available in the link given with each of the above models.

3.3 Prompt Engineering

Prompt engineering is a crucial task for optimizing the performance of a large language model on customized tasks (Schulhoff et al., 2024). The prompt types that were tried were based on role prompting, zero-shot prompting, few-shot prompting, self criticism and chain of thought reasoning. Prompt Template 4 shows the 10 prompt variations that were tested for encounter generation. These prompt engineering methods were taken from Schulhoff et al. (2024). As the first step, all the 8 finetuned models were tested with the basic prompts (Prompt 1) and (Prompt 2) for 20 frequently used monsters and the best 2 models were selected. Then, for these 2 models, 10 prompt types were tried out and tested with the task of asking the LLM to create encounters for 20 frequently used monsters that were picked from diverse categories.

The prompt and the best model that yielded the best result was selected, and the best model was tested with the best prompt format for 241 frequently used monsters D&D. Then, the model outputs given for the 241 monsters were evaluated by 2 human annotators and 3 LLMs. `GPT-4.1`, `Gemini-2.5-flash` and `DeepSeek-V3-0324` were the LLMs used as annotators. Also, different prompt structures were tried on LLMs, asking the LLM to provide judgements whether the model outputs are correct, partially correct or wrong. The prompt format for which the LLM answers show

[20] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_role.jsonl

[21] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_role_all.jsonl

[22] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[23] https://www.runpod.io/

[24] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2v0.1

[25] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2v0.2

[26] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2rolev0.1

[27] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2rolev0.2

[28] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2allv0.1

[29] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2allv0.2

[30] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2roleallv0.1

[31] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2roleallv0.2

Prompt Formats

1. **Prompt 1 (Zero shot basic prompt) :** Create an encounter that consists of a {}?
2. **Prompt 2 (Prompt with the same format as training instruction) :** What are the minion monsters that can come with an {} in an encounter?
3. **Prompt 3 (Prompt with the same format as training instruction :** Can you tell me the monsters that can go together with a {} in an encounter?
4. **Prompt 4 (Prompt asking for explanation) :** Give me the answer with explanation. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with.
5. **Prompt 5 (Prompt asking to walk through thinking process of LLM) :** Walk me through this context in manageable parts step by step, summarising and analysing as we go. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {}, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
6. **Prompt 6 (Repeating prompt in answer to enable generation) :** Repeat the following prompt in your answer. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
7. **Prompt 7 (Role prompting) :** You are a D&D expert. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
8. **Prompt 8 (Role prompt with repeating the prompt in answer) :** You are a D&D expert. Repeat the prompt in your answer. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
9. **Prompt 9 (Role prompt with examples in prompt) :** You are a D&D expert planning to create coherent D&D encounters. A coherent D&D encounter with a Mind flayer as the boss monster has Intellect Devourers and Thralls as minions. A coherent D&D encounter with a Red Dragon as the boss monster has Kobolds and dragon cultists as minions. Give a similarly coherent D&D encounter with a {} as the boss monster while ignoring the CR difference, but adapting the encounter theme to match the creature type and the typical environment of a {} lair.
10. **Prompt 10 (Role prompt with specifing output) :** You are a D&D expert planning to create coherent D&D encounters. A coherent D&D encounter with a Mind flayer as the boss monster has Intellect Devourers and Thralls as minions. A coherent D&D encounter with a Red Dragon as the boss monster has Kobolds and dragon cultists as minions. Give a similarly coherent D&D encounter with a {} as the boss monster and 4 different minion monsters while ignoring the CR difference, but adapting the encounter theme to match the creature type and the typical environment of a {} lair.

Prompt Template 4: Prompt formats used for encounter generation

higher Spearman correlation with the human annotations is selected as the best prompt format and the answers of the LLMs for the particular prompt format were taken as the final judgement of the particular LLM.

Table 1: Statistics obtained by finetuning the different models and the accuracy in obtaining a coherent monster list for a given monster

Fine-tuned Model	No of Epochs	Finetuning Time (Minutes)	Accuracy (%)
Model1v0.1	1	38.0	20.0
Model1v0.2	2	87.0	27.5
Model2v0.1	1	41.0	22.5
Model2v0.2	2	91.0	37.5
Model3v0.1	1	57.0	22.5
Model3v0.2	2	118.0	30.0
Model4v0.1	1	64.0	27.5
Model4v0.2	2	132.0	40.0

4 Results

The results of the model training, prompt engineering, and the analysis done on the outputs from the fine-tuned LLM for the best prompt are provided in this section. Table 1 shows the time that was taken for fine-tuning each of the 8 models and the percentage of correct answers obtained for the 20 frequently used monsters with the basic prompt formats, prompt 1 and prompt 2. It was seen that the models fine-tuned for 1 epoch provided comparatively very inaccurate answers compared to the models trained for 2 epochs. In addition, from the accuracies in the table, it can be seen that adding a role in the instruction dataset has resulted in the increase of the percentage of correct answers. Also, it could be observed that using all the question formats instead of selecting a random one provided slightly better results with the basic prompts. `Model2v0.2` and `Model4v0.2` were taken as the best models and tested with 10 prompt types with 20 frequently used monsters for the next experiment.

Results were obtained for the 2 best models (`Model2v0.2` and `Model4v0.2`) with 10 different types of prompt formats for a set of 20 monsters. From the results obtained, it can be seen that the best result obtained for the `Model4v0.2` was 50.0% which is less compared to the first 2 best results obtained for the `Model2v0.2` which are 70.0% and 55% respectively. Considering the overall results, it can be seen that prompt 9 and prompt 10 with the `Model2v0.2` yielded the best results. And it is noted from Prompt Template 4, that prompt 9 and prompt 10 are based on providing two examples of coherent monster sets in addition to the input

Table 2: Inter-Annotatement agreements between pairs of judges as measured by Spearman correlation. We also show the Spearman Correlation scaled by the Human 1 to Human 2 value to show the relative success of the LLMs.

	Human 1		Human 2		GPT-4.1		Gemini-2.5 flash	
	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled
Human 2	0.49	1.00						
GPT-4.1	0.39	0.80	0.44	0.90				
Gemini-2.5-flash	0.45	0.92	0.53	1.08	0.59	1.20		
DeepSeek-V3-0324	0.48	0.98	0.43	0.88	0.53	1.08	0.53	1.08

Table 3: Percentage of correct, partial and wrong answers for the set of 241 different monsters for the best prompting method according to different judges

Result	Human 1	Human 2	GPT-4.1	Gemini-2.5 flash	DeepSeek-V3 0324	Overall
Coherent monsters in output	37.3	48.1	45.2	37.8	28.6	39.4
Partially coherent monsters in output	21.2	21.2	11.6	31.5	47.7	26.6
Not coherent	41.5	30.7	43.2	30.7	23.7	33.9

query monster and also uses a *role* ("*You are a D&D expert planning to create coherent D&D encounters.*") in the prompt. Conversely, it can be seen when considering the other prompting methods, some commonly used techniques such as asking the LLM to walk through the steps (Prompt 5), asking for explanation (Prompt 4), asking to include the prompt in the answer (Prompt 6) did not work well for this task. From the above results, `Model2v0.2` was considered as the best model and Prompt 9 was taken as the best prompt for the next experiment. These are the final best model and the best prompt used for extensive experimentation. The full results of this experiment is given in Appendix A.

The best prompt was applied to the set of 241 different monsters in D&D and the outputs were obtained. Table 2 shows the Spearman correlations between the different judges for the best judgment prompt. Similar to Palpanadan et al. (2022) and Van Aswegen and Engelbrecht (2009), we use the ranges defined by Guilford (1950) to determine the strength of the correlation. The inter-annotator agreement between the human judges, as measured by Spearman correlation, is observed to be 0.49, which can be considered as a moderate correlation. Therefore, any AI-Human correlation that approaches this value may be taken as reasonable. In order to highlight this relative measure, we have added *scaled* columns to Table 2 where each of the results are scaled as a ratio over the Human-Human correlation value. Considering the best judgment prompt, average AI-AI agreement as measured by Spearman correlation is observed to be 0.54, and the average Human-AI agreement as measured by Spearman correlation is observed to be 0.47, which are also observed to be moderate correlations. When scaled by the Human-Human value, the average Human-AI agreement can be taken as 0.96. The AI-AI agreement, in fact, exceeds 100% when scaled by the Human-Human value.

Further, we conducted a judgment analysis where the results of the best prompt was judged by humans and other LLMs, following the *LLM as Judge* experiment regimen proposed by works such as Gunathilaka and de Silva (2025). Table 3 shows the results obtained by the judgment results for the finetuned LLM by 5 different judges and the overall results. We show the full results of this experiment in Appendix B.

When tested with the best prompt for a set of 241 monsters, some interesting observations were had. For some of the prompts, answers obtained were not wrong but general. Which means they did not specify any monster, but did provide the general name that represented a category of monsters. Also, in some other prompts pertaining to the monsters that are usually found alone, the LLM correctly provided the answer that the monster hunts alone. This by itself is proof that the LLM has correctly learnt the lore. An analysis on whether the output of the LLM is consistent across CR levels is shown in Fig 3. It shows the comparison of cumulative counts of success, partial success, and failure as a percentage across different challenge ratings. The analysis shows that for lower challenge ratings (<1), the cumulative success percentage shows a minor dip. But after CR 1, the trends stabilize. *This*

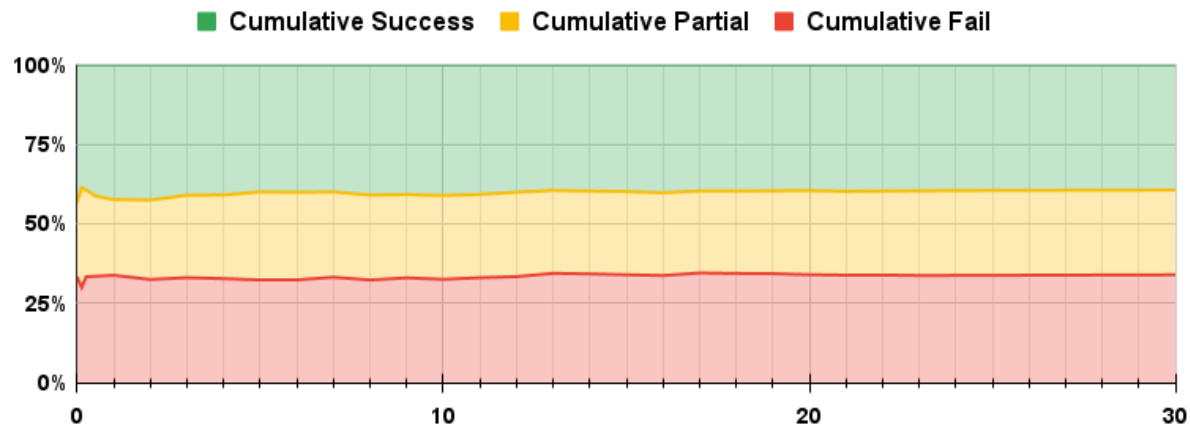


Figure 3: Comparison of cumulative counts of correct, partially correct and failure outputs across Challenge ratings of monsters

may be due to the *boss* and minion format in the prompt. Naturally, when given a monster as a *boss*, the LLM is trying to match them up with potential *minions* who by definition should be of lower power level (CR). But, these monsters being the lowest rung of monsters, the LLM was having a hard time proposing even lower-level monsters to be minions of them.

5 Conclusion and Future Directions

An LLM for generating encounters for a Dungeons and Dragons (D&D) was fine-tuned by instruction tuning an already instruction-tuned `Mistral7B` based LLM using `QLoRa`. The results show that 66.0% of encounters were cohesive or partially cohesive. We can see that further instruction tuning of an already instruction-tuned model is much effective at adapting LLMs to a different domain. The results of the subsequent prompt engineering show that role prompting with examples, which provide the monster combination for an encounter as boss and minions, was the most effective prompt for this particular application. It was also observed that in some instances, in place of the concrete monster name, the LLM provided the category of monsters the result. This may be an artefact of the presence of category level relations in the lore text (eg: "*Mind Flayers* may be seen with other creatures from the *Far Realm*").

Some of the results may also have been affected by the facts that: 1) Not all monsters can be candidates for the boss and minion format, 2) Some monsters are defined as solo creatures in the lore. In cases where these conditions were in play, the LLM would have provided an explanation of the impossibility of creating an encounter. Our current rigid evaluation criteria would have taken such instances as failures on the part of the LLM. However, we consider correcting this to be out of scope for this work and point out that this error results in an under-counting and not an over-counting. Thus, our reported accuracies are a strict lower limit to the actual possible human perceived accuracy. A human DM will find some of the results that we have currently rejected as wrong, to be reasonably acceptable.

As a further future direction, it is planned to augment our system to provide multiple potential encounters in a singular prompt and then, provide a ranking of the encounters based on the coherence to the lore. It is expected that providing the DMs with such a choice may lead to better usability of the system.

References

Justice Ramin Arman, Dan Dillon, and F. Wesley Schneider. 2023. *Planescape: Adventures in the Multiverse*. Wizards of the Coast.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *arXiv preprint arXiv:2410.00873*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael

Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393. Association for Computational Linguistics.

A Chowdhery and 1 others. 2022. Scaling language modeling with pathways.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2024. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*.

Jeremy Crawford, Christopher Perkins, and James Wyatt. 2014a. *Dungeon Master's Guide*. Wizards of the Coast LLC.

Jeremy Crawford, James Wyatt, and Keith Baker. 2019. *Eberron: Rising from the Last War*. Wizards of the Coast.

Jeremy Crawford, James Wyatt, Robert J Schwalb, and Bruce R Cordell. 2014b. *Player's Handbook*. Wizards of the Coast LLC.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 36.

Simon Ellis and James Hendler. 2017. Computers play chess, computers play go… humans play dungeons & dragons. *IEEE Intelligent Systems*, 32(4):31–34.

Haruka Fujiwara, Renta Kimura, and Tokuniki Nakano. 2024. Modify Mistral Large Performance with Low-Rank Adaptation (LoRA) on the BIG-Bench Dataset. *ResearchSquare*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Joy Paul Guilford. 1950. *Fundamental statistics in psychology and education*. McGraw-Hill.

Sadeep Gunathilaka and Nisansa de Silva. 2025. Automatic Analysis of App Reviews Using LLMs. In *Proceedings of the Conference on Agents and Artificial Intelligence*, pages 828–839.

Gary Gygax and Dave Arneson. 1974. *Dungeons & dragons*, volume 19. Tactical Studies Rules Lake Geneva, WI.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Adam Lee, James Introcaso, Ari Levitch, Mike Mearls, Lysa Penrose, Christopher Perkins, Ben Petrisor, Matthew Sernett, Kate Welch, Richard Whitters, and Shawn Wood. 2019. *Baldur's Gate: Descent into Avernus*. Wizards of the Coast.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713. Association for Computational Linguistics.

Lara J Martin, Srijan Sood, and Mark O Riedl. 2018. Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games. In *INTWICED*.

Sarala Thulasi Palpanadan, Toong Hai Sam, Khairunesa Isa, Nurliyana Md Rosni, Asokan Vasudevan, Kai Wah Cheng, and Xue Ruiteng. 2022. Relationship between knowledge level and online consumer purchasing attitude during covid-19 endemic phase. *resmilitaris*, 12(5):352–363.

Akila Peiris and Nisansa de Silva. 2022. Synthesis and evaluation of a domain-specific large data set for dungeons & dragons. *arXiv preprint arXiv:2212.09080*.

Akila Peiris and Nisansa de Silva. 2023. SHADE: Semantic Hypernym Annotator for Domain-Specific Entities-Dungeons and Dragons Domain Use Case. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6. IEEE.

Christopher Perkins, Adam Lee, Richard Whitters, and Jeremy Crawford. 2015. *Curse of Strahd*. Wizards of the Coast.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134. Association for Computational Linguistics.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, and 1 others. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Aravinth Sivaganeshan and Nisansa De Silva. 2023. Fine tuning named entity extraction models for the fantasy domain. In *2023 Moratuwa Engineering Research Conference (MERCon)*, pages 346–351. IEEE.

Kurt Squire. 2007. *Open-ended video games: A model for developing learning for the interactive age*. MacArthur Foundation Digital Media and Learning Initiative.

Eddo Stern. 2002. A touch of medieval: Narrative, magic and computer technology in massively multiplayer computer role-playing games. In *CGDC Conf.*

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are expert-level language models expert-level annotators? *arXiv preprint arXiv:2410.03254*.

Anja S Van Aswegen and Amos S Engelbrecht. 2009. The relationship between transformational leadership, integrity and an ethical climate in organizations. *SA Journal of Human Resource Management*, 7(1):1–9.

Gayashan Weerasundara and Nisansa de Silva. 2023. Comparative analysis of named entity recognition in the dungeons and dragons domain. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1225–1233. INCOMA Ltd., Shoumen, Bulgaria.

Gayashan Weerasundara and Nisansa de Silva. 2024. A Multi-Stage Approach to Image Consistency in Zero-Shot Character Art Generation for the D&D Domain. In *ICAART (3)*, pages 235–242.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Dylan Zhang, Justin Wang, and Francois Charton. 2024. Instruction diversity drives generalization to unseen tasks. *arXiv preprint arXiv:2402.10891*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

A Detailed Results of the prompting experiments

The detailed results of the prompting experiments that were used to compare the best 2 models from the initial experiments and also to select the best prompt format with the best model for the final experiment are given in Table 4. The selection of 20 monsters for this experiment was done manually to cover monsters belonging to different environments, types and alignments to ensure that the testing is unbiased. It can be observed that the combination of `Model2v0.2` and `Prompt 09` yields the best result while specifically for `Model4v0.2`, `Prompt 02` suits best. Overall, most prompts seem to struggle with monsters such as, *Storm Giant*, *Iron Golem*, and *Green Hag* while monsters such as *Lich*, *Sahuagin Baron*, and *Bandit Captain* seem to be easy for most prompts to handle.

B LLM-as-a-Judge Experiments

`GPT-4.1`, `Gemini-2.5-flash` and `DeepSeek-V3-0324` were used with the different judgement prompts to obtain judgements from the relevant LLMs. Prompt Template 5 shows the different prompts that were tried. The judgement prompts were created to impose different types of conditions to the judge the outputs of the fine-tuned LLM. For example, `Judgement Prompt 1` imposes several conditions that a human judges might look for in the fine-tuned LLM's output and `Judgement Prompt 6` simply asks the Judge LLM to rate the fine-tuned LLM's output without imposing any condition.

Results were obtained for the above experiment from each judge (2 human, 3 LLM). Based on this, agreement percentages and spearman correlation were calculated between judgements of each pair of judges to analyze the relationship between the

Table 4: Percentage of correct answers on the initial reference set of 20 monsters for different prompting methods along with basic monster statistics. The Alignment is given as AB where A={L: Lawful, N: Neutral, C: Chaotic} and B={G: Good, N: Neutral, E: Evil}

Prompt	Monster Stats										Prompt Results for `Model2v0.2`										Prompt Results for `Model4v0.2`									
	CR	Type	Strength	Dexterity	Constitution	Intelligence	Wisdom	Charisma	Alignment	Habitat	Prompt 01	Prompt 02	Prompt 03	Prompt 04	Prompt 05	Prompt 06	Prompt 07	Prompt 08	Prompt 09	Prompt 10	Prompt 01	Prompt 02	Prompt 03	Prompt 04	Prompt 05	Prompt 06	Prompt 07	Prompt 08	Prompt 09	Prompt 10
Skeleton	1/4	Undead	10	14	15	6	8	5	LE	Urban	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗
Bandit Captain	2	Humanoid	15	16	14	14	11	14	Any	Artic Coastal Desert Forest Hill Urban	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intellect Devourer	2	Aberration	6	14	13	12	11	20	LE	Underdark	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Hobgoblin Captain	3	Fey	15	14	14	12	10	13	LE	Any	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓
Green Hag	3	Fey	18	12	16	13	14	14	NE	Forest Hill Swamp	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Owlbear	3	Monstrosity	20	12	17	3	12	7	U	Forest	✗	✗	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Water Elemental	5	Elemental	18	14	18	5	10	8	N	Coastal Swamp Underwater	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Sahuagin Baron	5	Fiend	19	15	16	14	13	17	LE	Coastal Underwater	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
Hill giant	5	Giant	21	8	19	5	9	6	CE	Hill	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	✗	✗
Treant	9	Plant	23	8	21	12	16	12	CG	Forest	✗	✗	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗
Aboleth	10	Aberration	21	9	15	18	15	18	LE	Underdark Underwater	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Elder Oblex	10	Ooze	15	16	21	22	17	18	LE	Swamp Underdark Urban	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
Djinni	11	Elemental	21	15	22	15	16	20	CG	Coastal	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗
Beholder	13	Aberration	16	14	18	17	15	17	LE	Underdark	✗	✓	✗	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Storm Giant	13	Giant	29	14	20	16	20	18	CG	Coastal Underwater	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Iron Golem	16	Construct	24	9	20	3	11	1	U	Any	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Adult blue dragon	16	Dragon	25	10	23	16	15	19	LE	Coastal Desert	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Pit Fiend	20	Fiend	26	14	24	22	18	24	LE	Any	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
Lich	21	Undead	11	16	16	21	14	16	NE	Any	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓
Solar	21	Celestial	26	22	26	25	25	30	LG	Any	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗
Percentage											20.0	35.0	40.0	35.0	25.0	30.0	40.0	40.0	**70.0**	55.0	20.0	**60.0**	50.0	20.0	25.0	40.0	40.0	40.0	35.0	30.0

Table 5: Percentage of agreement between the judgements of different pairs of judges for different Judgement Prompts

Agreement (%)	Judgment Prompts					
	J-Prompt 1	J-Prompt 2	J-Prompt 3	J-Prompt 4	J-Prompt 5	J-Prompt 6
H1 vs H2	58.92	58.92	58.92	58.92	58.92	58.92
H1 vs GPT	53.94	52.70	50.62	56.43	53.11	55.19
H1 vs Gemini	48.96	51.45	46.47	55.60	51.87	57.68
H1 vs DeepSeek	44.81	49.38	46.06	48.13	46.06	38.59
H2 vs GPT	51.45	49.79	51.45	56.02	57.26	53.11
H2 vs Gemini	55.60	46.47	49.38	54.77	49.38	56.02
H2 vs DeepSeek	49.79	50.62	46.89	47.72	45.64	36.51
GPT vs Gemini	58.92	63.90	58.51	63.07	52.70	61.41
GPT vs DeepSeek	43.15	46.89	39.42	52.70	48.55	49.79
Gemini vs DeepSeek	44.40	48.13	41.49	54.36	54.36	48.13

Table 6: Spearman correlation between the judgements of different pairs of judges for different Judgement Prompts

Spearman Correlation	Judgment Prompts					
	J-Prompt 1	J-Prompt 2	J-Prompt 3	J-Prompt 4	J-Prompt 5	J-Prompt 6
H1 vs H2	0.49	0.49	0.49	0.49	0.49	0.49
H1 vs GPT	0.36	0.37	0.34	0.40	0.37	0.47
H1 vs Gemini	0.36	0.40	0.37	0.45	0.43	0.51
H1 vs DeepSeek	0.29	0.40	0.34	0.48	0.41	0.29
H2 vs GPT	0.33	0.29	0.31	0.44	0.45	0.46
H2 vs Gemini	0.42	0.36	0.39	0.53	0.40	0.49
H2 vs DeepSeek	0.29	0.33	0.27	0.43	0.38	0.18
GPT vs Gemini	0.56	0.57	0.48	0.59	0.51	0.65
GPT vs DeepSeek	0.14	0.24	0.15	0.53	0.42	0.32
Gemini vs DeepSeek	0.25	0.36	0.24	0.53	0.38	0.30

Prompt Formats

1. **Judgement Prompt 1 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct.
2. **Judgement Prompt 2 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct. Be a tough grader.
3. **Judgement Prompt 3 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct. Consider that the ratio of correct:partially correct:wrong answers is 90:51:100.
4. **Judgement Prompt 4 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. Answer should be only correct, wrong or partially correct.
5. **Judgement Prompt 5 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output. Answer should be only correct, wrong or partially correct.
6. **Judgement Prompt 6 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong. Answer should be only correct, wrong or partially correct.

Prompt Template 5: Prompt Formats given to LLMs for obtaining LLM judgements

judgment of different judges on the fine-tuned LLM outputs. Table 5 shows the agreement percentages between each pairs of judges. Table 6 shows the spearman correlations between the different pairs of judges.

Verb Phrase Idioms in Valency Alternation: A Selection-based Approach

Tomokazu Takehisa
Saitama Prefectural University
820 Sannomiya, Koshigaya
Saitama, Japan
takehisa-tomokazu@spu.ac.jp

Abstract

This paper discusses Japanese verb phrase idioms based on verbs alternating in valency and makes the following claims. First, the ditransitive/transitive alternation involves augmentation of an internal argument, unlike the transitive/intransitive alternation, which involves augmentation of an external argument. Second, given the structure of transitive/intransitive verbs, the transitive version of an intransitive idiom should in principle be available in terms of structure, yet some intransitive idioms lack their transitive counterparts. In defense of the selection-based approach to idioms, a possible solution is suggested where the absence of the transitive version is related to the interpretive properties of the base intransitive idiom, which might have structural reflexes. The paper also discusses the status and distribution of *ni*-phrases in transitive idioms.

1 Introduction

Idioms have been intensively studied since the early days of generative grammar, as they pose important challenges to the form-meaning correspondence in general. Phrasal idioms can sometimes behave as an atomic unit like a word, while acting as a complex phrase to varying degrees. In the case of verb phrase idioms, some idioms act as a complex predicate, and, if a nominal element is an essential part of an idiom, that element, being part of a verbal predicate, behaves as such and cannot be referential.

Another topic that has been intensively studied is verbs that alternate in valency. Since the early to mid-1980s, research on lexical semantics and its relation to syntax has centered around valency alternations such as the causative alternation, the double object alternation, the resultative construction, and the like, and has contributed to the current conceptions of argument structure in the principles-and-parameters framework.

When we consider verbs alternating in valency and verb phrase idioms together, there are three patterns as to the relation between alternating verbs and idioms, depending on the minimum size of a particular idiom: (i) idioms with both the valency-reduced and the valency-augmented alternants, (ii) idioms only with the valency-reduced alternant, (iii) idioms only with the valency-augmented alternant.

In this paper, I will discuss the alternation patterns of verb phrase idioms based on two types of alternating verbs in Japanese and argue that these two types involve different alternation processes. Specifically, I will argue that ditransitive/transitive verbs involve augmentation of an indirect internal argument, while transitive/intransitive verbs involve augmentation of an external argument. I will also discuss transitive idioms that are predicted to exist but not detected under the present approach and suggest possible solutions to their absence. Moreover, as an implication of the present analysis, I will discuss the status and distribution of *ni*-phrases in transitive idioms.

Before proceeding, I would like to briefly review the basic assumptions in this paper. One is a realizational theory of the morphology of Japanese transitive/intransitive verbs. Japanese has a series of idiosyncratic transitive-intransitive morpheme pairs, which are allomorphs of the same syntactic heads, and their distribution is determined by the particular lexical roots they are associated with (e.g., Jacobsen, 1990).

Under the theory of Distributed Morphology (Halle and Marantz, 1993; Embick, 2015), their distribution is handled by Vocabulary Insertion (VI) rules (Nishiyama, 1997; Miyagawa, 1998). In this paper, the (templatic) VI rules in (1)a and (1)b are assumed for the transitive and intransitive morphemes, respectively. They are assumed to be applied in the structure in (1)c, which is derived post-syntactically via head movement or a formally

equivalent operation.[1]

(1) a. v $\leftrightarrow s/X$__VOICE, where $X \in \{\sqrt{}\text{A}, \sqrt{}\text{B}, \textit{etc.}\}$
b. v $\leftrightarrow r/X$__, where $X \in \{\sqrt{}\text{A}, \sqrt{}\text{B}, \textit{etc.}\}$
c. After syntax:

$$\begin{array}{c} \tau\psi \\ \tau\psi \text{ VOICE} \\ \sqrt{} \quad \text{v} \end{array}$$

More concrete examples of the VI rules would be as in (2), for the transitive-intransitive pair *tuk-e(-ru)* 'attach (tr.)' and *tuk-∅(-u)* 'attach (in.)'.

(2) a. v $\leftrightarrow$ e/X__VOICE, where $X \in \{\sqrt{}\text{tuk}, \textit{etc.}\}$
b. v $\leftrightarrow \varnothing/X$__, where $X \in \{\sqrt{}\text{tuk}, \textit{etc.}\}$

Along with the realizational view of the verbal morphology, a relational theory of lexical causation is adopted, which assumes that the causative semantics of a lexical causative verb emerges from the combination of VOICE and its complement vP.[2] Specifically, the verbalizing functional head is interpreted as causative when combined with , as given in the semantic interpretation rule in (3), a simplified and slightly modified version of the one in Wood and Marantz (2017).

(3) $[\![\text{v}]\!] \leftrightarrow \lambda\text{P}\lambda\text{e}\exists\text{e}'[\text{P}(\text{e}') \,\&\, \text{CAUSE}(\text{e}, \text{e}')]$/__VOICE

Note that, although I adopt the causative semantics of lexical causation as relational, the syntactic causative predicate in Japanese, *-(s)ase*, is a single syntactic head specially dedicated to causation. Thus, these two kinds of causation in Japanese are similar yet ultimately different.

Yet another assumption adopted in this paper concerns phrasal idioms. Phrasal idioms are linguistic expressions with meanings that are unpredictable from those of their component parts (Nunberg *et al.*, 1994). It is standardly assumed (e.g., O'Grady, 1998; Bruening, 2010; Everaert, 2010) that an idiomatic interpretation is possible when a set of particular lexical items are related in a particular configuration, which can be understood as a chain of selectional relations between the lexical items involved. For example, an idiom like *shoot* the *breeze* is assumed to involve the following set of selectional relations, as given in (4)c.[3][4]

(4) *shoot the breeze*
a. [$_{\text{vP}}$ v $\sqrt{}$shoot[$_{\text{nP}}$[$_{\text{Det}}$ the][$_{\text{n}}$ n $\sqrt{}$breeze]]]
b. 'have a casual conversation; talk nonsense'
c. Selectional relations
i. v $\rightarrow \sqrt{}$shoot
ii. v $\rightarrow$ n
iii. n $\rightarrow \sqrt{}$breeze
iv. n $\rightarrow$ Det

A structure where all the required relations hold for an idiom can receive the relevant idiomatic interpretation,[5] so disrupting any one of them leads to its loss. Thus, if the object nP in (4)a is replaced with the PP *in the breeze*, the idiomatic interpretation (4)b becomes unavailable because the required selectional relation in (4)c-ii is replaced by two other relations irrelevant to the idiomatic interpretation: v $\rightarrow$ P and P $\rightarrow$ n.

Furthermore, there are cases where the nature of the event described matters as to the possibility of valency alternation. As Levin and Rappaport

[1] I assume that VOICE comes in two varieties: the active one, which introduces an external argument (Kratzer, 1996), and the passive one, which introduces an implicit agent (i.e., $[\![\text{VOICE}_{\text{PASS}}]\!] = \lambda e\, \exists x\, [\text{Agent}(e, x)]$) (e.g., Landau, 2010). I identify the active VOICE head as the basic argument introducer i*, proposed by Wood and Marantz (2017), which is category-neutral and inherits the first category it merges with. For the purposes of this paper, it can be taken to be a verbal functional head, i.e., as a flavor of v in the sense of Folli and Harley (2007). However, I diverge from the original proposal in that the head is not endowed with the function of closing off its projection, which is notated by "*." Thus, I omit the notation in what follows and use i instead of i*.

[2] See Hale and Keyser (1993) for the view of CAUSE as a syntactic configurational relation and Pylkkänen (2008) for the view of CAUSE as a syntactic head denoting a relation.

[3] For example, the relevant principle and constraint formulated in Bruening (2010: 532) are as follows:

(i) The Principle of Idiomatic Interpretation
X and Y may be interpreted idiomatically only if X selects Y.

(ii) Constraint on Idiomatic Interpretation
If X selects a lexical category Y, and X and Y are interpreted idiomatically, all of the selected arguments of Y must be interpreted as part of the idiom that includes X and Y. (Lexical categories are V, N, A, Adv.)

I assume that lexical categories in (ii) can be formed by categorizing heads such as v, n, and a.

[4] The precise treatment of weak definite articles in idioms is outside the scope of this paper, as Japanese does not have articles. While I simply assume that they are not obstacles to idiom formation, I follow Bruening (2010) in that determiners in English are modifiers to nPs (NPs in his paper), as given in the text. Note that I use the labels DP and nP(NP) interchangeably throughout this paper. See Gehrke and McNally (2019) for an approach to idioms that aims to integrate two different domains of semantics that are computed separately: reference and descriptive content.

[5] I assume that the Encyclopedia, a post-syntactic component in Distributed Morphology responsible for extralinguistic knowledge, is also responsible for the assignment of idiomatic interpretation.

Hovav (1995) argue, when the event described by the verb requires the intervention of an agent, the valency reduction of a transitive/intransitive verb is impossible, as shown in (5) below. In contrast, the causing argument of an externally caused event can be left unexpressed if the nature of the causing event is completely unspecified.

(5) a. He broke his promise/the world record.
b. .*His promise/The world record broke.
(Levin and Rapapport Hovav, 1995: 105)

In general, if a verb phrase describes an event whose nature implies the involvement of an agent, the presence of a VOICE head is effectively required. The same consideration applies to events described by idioms: If an event described by an idiom implies the involvement of an agent, it is not necessary to specify a VOICE head as its essential part, though it is possible to do so and there are indeed such cases.[6]

Lastly, the form-meaning correspondence in idioms can be one-to-many, aside from the non-idiomatic, literal interpretation: There can be more than one idiomatic interpretation per structure. In this paper, examples of idioms are presented with at least one idiomatic interpretation to show that they are idioms, but it should be kept in mind that there may be other idiomatic interpretations that are ignored in the discussion.[7]

With these backgrounds, we will turn to idioms based on verbs alternating in valency and the alternation patterns they display.

2 Idioms and Valency Alternation

2.1 Ditransitive/transitive verbs

There is a class of verbs in Japanese which alternate between ditransitive and transitive verbs. These verbs take nominative and accusative arguments in the dyadic use, and they can take an additional dative argument when they are turned into triadic by adding an argument-introducing morpheme *-se*, which I identify as APPL. [8]

We can find idioms based on verbs of this class: Some idioms are possible with both the transitive and the ditransitive alternants, as in (6), while other idioms are found only with one of the alternants, transitive (as in (7)), or ditransitive (as in (8)).[9][10]

(6) a. X-ga Y-o on-ni ki-∅-ru
X-NOM Y-ACC grace-DAT $\sqrt{}$put.on-V-NPST
'X is grateful/obliged for Y'
b. X-ni Y-o on-ni ki-∅-se-ru
X-DAT Y-ACC grace-DAT
ki-∅-se-ru
$\sqrt{}$put.on-V-APPL-NPST
'make X feel obliged for Y'

(7) a. X-ga baka-o mi-∅-ru
X-NOM nonsense-ACC $\sqrt{}$*textsee*-V-NPST
'X ends up losing out'
b. *X-ni baka-o mi-∅-se-ru
X-DAT nonsense-ACC $\sqrt{}$see-V-APPL-NPST
[No idiomatic interpretation] (see above)
c. X-ni baka-o mi-∅-sase-ru
X-DAT nonsense-ACC $\sqrt{}$see-V-CAUS-NPST
'make X end up losing out'

(8) a. *X-ga hiyamizu-o abi-∅-ru
X-NOM cold water-ACC $\sqrt{}$pour.on-V-NPST
[No idiomatic interpretation] (see below)
b. X-ni hiyamizu-o abi-∅-se-ru
X-DAT cold water-ACC
abi-∅-se-ru
$\sqrt{}$pour.on-V-APPL-NPST
'throw a wet blanket on X'
Lit.: 'pour cold water over X'

Since embedding an idiom should be fine as long as its prerequisite relations are kept intact, it is expected that (7)b should be possible, which is not the case. Given that embedding per se is not banned, as shown in (7)c, the unacceptability of (7)b suggests that something other than simple embedding is involved in the ditransitive/transitive alternation.

[6] See, for example, Schildmier Stone (2016) for English idioms that resist passivization (e.g., *kick the bucket*).

[7] When more than one idiomatic interpretation per structure is possible, the ambiguity is resolved contextually.

[8] It is more or less standard in the literature to treat this morpheme as a lexical causative (e.g., Matsuoka, 2003). The treatment in the text is justified by the success of the analysis presented therein.

[9] For space reasons, the nominative subject is omitted in most examples with the valency-augmented alternant.

[10] Abbreviations used in this paper are as follows: ACC(usative), APPL(locative), CAUS(ative), DAT(ive), DIM(unitive), DV = dummy verb, GEN(itive), HON(orific), IN = intransitive, INST(rumental), LOC(ative), N = nominalizer, NEG(ative), NOM(inative), NPST = nonpast, PASS(ive), PST = past, TOP(ic), TR = transitive, V = verbalizer, $\sqrt{}$noun = nominal root, $\sqrt{}$verb = verbal root.

To account for the difference in question, I argue that the transitive and the ditransitive alternants have the following structures, respectively, where *i* is assumed to introduce a thematically underspecified argument (Takehisa, 2018; cf. Wood and Marantz, 2017):[11]

(9) a. Transitive
$[_{vP2}\ DP_{NOM}\ [[_{vP1}\ DP_{ACC}\ [_{v}\ \sqrt{}verb\ v\ ]]\ i\ ]]$
b. Ditransitive
$[_{vP2}\ DP_{NOM}\ [[_{APPLP}\ DP_{DAT}\ [\ vP1\ APPL\ ]]\ i\ ]]$

As depicted in (9)b, it is assumed that APPL, a head introducing an internal argument, is responsible for deriving the ditransitive alternant, which is crucially different from previous analyses (e.g., Inoue, 1976), where *-se* is assumed to be a causative and introduce an external argument.

Under this applicativization view of ditransitives, an idiom displays the ditransitive/transitive alternation only when vP1 in (9) contains all the essential parts of that idiom; if an idiom requires material outside vP1 as essential, it does not alternate in ditransitivity and only the transitive or ditransitive version is available as an idiom.

Thus, while the examples in (6) involve an idiom whose essential parts are all contained in vP1, the idioms in (7) and (8) require material outside vP1. More specifically, the idiom involved in (7) requires at least the head of vP2 (i.e., *i*) as well as the material within its immediate complement vP1 (as in (9)a), which makes the ditransitive version in (7)b impossible as an idiom. Likewise, the idiom in (8)b requires at least the APPL head as well as the material within its immediate complement vP1 (as in (9)b), which makes the transitive version in (8)a impossible. Hence, the selection-based approach, combined with the applicativization view, can account for the alternation patterns of idioms based on ditransitive/transitive verbs.

Given the correspondence between the transitive nominative subject and the ditransitive dative object of the alternating verb, as in (10) below, it might appear natural to assume that the alternation involves causativization, as widely assumed in the literature. However, these two arguments are crucially different in that the transitive subject is variable in interpretation and can be either agentive or non-agentive, as shown in (11) and (12) below, respectively, each of which is a continuation of (10)a: (11) involves the process of *soo suru* ('do so') replacement, which is only possible with volitional agents, ensuring that the subject in (10)a can only be agentive; on the other hand, the subject in (10)a is forced to be non-agentive by (12), where the agenthood of the subject is negated in the presence of an empathic reflexive.

(10) a. Taro-ga mizu-o abi-∅-ta
T.-NOM water-ACC $\sqrt{}$pour.on-V-PST
i. 'Taro poured water over himself.'
ii. 'Taro got water poured over himself.'
b. Ziro-ga Taro-ni mizu-o
Z.-NOM T.-DAT water-ACC
abi-∅-se-ta
$\sqrt{}$pour.on-v-APPL-PST
'Ziro poured water over Taro.'

(11) Ziro-mo {soo si-ta / abi-∅-ta}
Z.-also so do-PST/ $\sqrt{}$pour.on-V-PST
'Ziro did so, too./ Ziro poured it, too.'

(12) kedo zibun-de-wa abi-∅-nak-ar-ta (> -at-ta)
but self-INST-TOP $\sqrt{}$pour.on-V-NEG-DV-PST
'but he didn't pour it himself.'

In contrast, the ditransitive dative object, as in (10)b, is not variable in thematic interpretation and can only be non-agentive.

The difference in this respect is reflected in the argument introducers involved: *i* introduces a thematically underspecified argument, but APPL introduces a thematically specified, non-agentive argument. Note that *i* also introduces the ditransitive subject, but it can only be agentive. This is because the presence of the non-agentive dative object prevents the subject from being non-agentive, as required by thematic uniqueness.[12]

Given the above discussion, it is predicted that the ditransitive version is unavailable when the selection of *v* by *i* is among an idiom's required relations. The most transparent case is idioms with

[11] Building on the work by Wood and Marantz (2017), Takehisa (2018) assumes that i introduces a thematically underspecified argument: an argument which can be interpreted either as agentive or non-agentive. To borrow Newman's (2024) notation and extend it into covering a non-agentive role, the denotation of i can be represented as follows: $[\![i]\!] \leftrightarrow \lambda e \lambda x[\text{-ER}(e, x) \vee \text{-EE}(e, x)]$, where the theta role labels "-ER" and "-EE" correspond to an external theta role and an internal theta role, respectively (cf. Dowty, 1990). The argument introduced by i ends up with one of the two interpretations at the interface, as a result of processes resolving thematic underspecification. See also footnote one.

[12] See Landman (2000), a.o., for thematic uniqueness.

transitive subjects as agents, as in (13)a, but those with non-agentive subjects, as in (7)a above, also count as evidence, since the subject interpretation is not necessarily variable in the case of idioms: It may be specified either as agentive or non-agentive as part of an idiomatic interpretation.

(13) a. X-ga neko-o kabur-∅-u
X.-NOM cat-ACC √put.on-V-NPST
'X pretends to be quiet; X puts it on'
b. *X-ni neko-o kabur-Ø-se-ru (> kabu-se-)
X-DAT cat-ACC √put.on-V-APPL-NPST
[No idiomatic interpretation] (see above)

Conversely, idioms whose subject is variable in thematic interpretation are based solely on material within vP1 in (9), and thus can alternate in ditransitivity, as shown in (14). The interpretational variability of the subject in (14)a is illustrated by (15) and (16), each of which is a continuation of (14)a and serves to disambiguate the subject.

(14) a. Taro-ga doro-o kabur-∅-ta (> kabut-ta)
T.-NOM mud-ACC √put.on-V-PST
i. 'Taro played a thankless role.'
ii. 'Taro was blamed (for someone).'
Lit.: 'Taro got covered with mud.'
b. Ziro-ga Taro-ni doro-o kabur-∅-se-ta (> kabu-se-ta)
Z.-NOM T.-DAT mud-ACC √put.on-V-APPL-PST
i. 'Ziro made Taro play a thankless role.'
ii. 'Ziro blamed Taro for him.'

(15) Ziro-mo {soo si-ta / kabur-∅-ta (> kabut-)}
Z.-also so do-PST/ √put.on-V-PST
'Ziro did so, too./ Ziro played it, too.'
Lit.: 'Ziro did so, too./ Ziro got covered, too.'

(16) kedo zibun-de-wa kabur-∅-anak-ar-ta (> -at-)
but self-INST-TOP √put.on-V-NEG-DV-PST
'but he didn't blame himself.'
Lit.: 'but he didn't get himself covered.'

Therefore, we conclude from these alternation patterns that the ditransitive alternant does not contain its transitive counterpart.

2.2 Transitive/intransitive verbs

Given the selection-based approach to idioms and that the transitive/intransitive alternation involves complementation of an intransitive vP by *i*, it is predicted that, if an intransitive idiom is possible, then its transitive counterpart is always possible as long as the idiom's required relations are kept intact. This prediction can be immediately countered by examples where idioms based on the intransitive alternant of an alternating verb have no transitive counterparts. In defense of the selection-based approach, I turn to such examples and present possible solutions under the present approach.

First, consider (17), where a different kind of idiom is formed by transitivization, in contrast to those in (18), (22), and (25) below, where the transitive and the intransitive idioms are related.

(17) a. (X-ni) atama-ga sag-ar-u
X-LOC head-NOM √lower-IN-NPST
'appreciate X; take one's hats off to X'
Lit.: 'One's head lowers to X'
b. (X-ni) atama-o sag-e-ru
X-LOC head-ACC √lower-TR-NPST
'apologize to X; beg X (for something)'
Lit.: 'bow to X'

The fact that (17)a cannot be retained in (17)b can be straightforwardly accounted for under the selection-based approach if we add the following: When a chain of selectional relations between particular lexical items is interpreted idiomatically, maximize the number of the relations involved. Thus, in (17)b, the selection of *v* by *i* must be counted as one of the idiom's essential relations.

Next, there are cases where transitivization destroys intransitive idioms. First, observe that transitive/intransitive verbs can form idioms that alternate in transitivity, or ones with only one of the alternants, transitive or intransitive. Consider (18), (19), and (20) below.

(18) a. (X-no)ryuu.in-ga sag-ar-u
X-GEN gastric juices-NOM √lower-IN-NPST
'{become/ X becomes} relieved/satisfied'
b. (X-ga) ryuu.in-o sag-e-ru
X-NOM gastric juices-ACC √lower-TR-NPST
'{have/ X has} one's grudge satisfied'

(19) a. ude-ga tat-∅-u
arm-NOM $\sqrt{}$stand-IN-NPST
'be skilled/competent'

b. *ude-o tat-e-ru
arm-ACC $\sqrt{}$stand-TR-NPST
[No idiomatic interpretation] (see above)

(20) a. *(o-)tya-ga nigo-r-u
HON-tea-NOM $\sqrt{}$muddy-IN-NPST
[No idiomatic interpretation] (see below)
Lit.: 'the tea gets cloudy'

b. (o-)tya-o nigo-s-u
HON-tea-ACC $\sqrt{}$muddy-TR-NPST
'give an evasive answer'
Lit.: 'make the tea cloudy'

The verbs involved in these examples are the ones whose transitive alternant takes two DPs as arguments, and idioms based on these verbs involve at least a verb and an internal argument as their essential parts. In addition, there are other types of idioms to consider. Specifically, there are transitive/intransitive verbs that take a locative PP as an argument. Idioms based on verbs of this class also involve a verb and an internal argument as their essential parts, but the internal argument may be DP or PP,[13] as depicted in the following, with an (intransitive) idiom underlined:

(21) a. DP as part of an idiom
$[\ldots[_{vP}\ \mathrm{PP}\ \underline{[\ \mathrm{DP}\ [_{v}\ \sqrt{}\ \ v\ ]]]}\ \ldots]$

b. PP as part of an idiom
$[\ldots[_{vP}\ \mathrm{DP}\ \underline{[\ \mathrm{PP}\ [_{v}\ \sqrt{}\ \ v\ ]]]}\ \ldots]$

With this distinction in mind, consider the following examples: The idioms in (22)-(24) have an internal DP argument as their essential part, while those in (25)-(26) have a locative PP as their essential part.

(22) a. X-ni hakusya-ga kak-ar-u
X-LOC spur-NOM $\sqrt{}$hook-IN-PST
'X gains impetus'

b. X-ni hakusya-o kak-e-ru
X-LOC spur-ACC $\sqrt{}$hook-TR-NPST
'give impetus to X'

(23) a. X-ni (Y-no) me-ga todok-∅-u
X-LOC Y-GEN eye-NOM $\sqrt{}$reach-IN-NPST
'X is within (Y's) sight; One/Y can see X'

b. *X-ni (Y-no) me-o todok-e-ru
X-LOC Y-GEN eye-ACC $\sqrt{}$reach-TR-NPST
[No idiomatic interpretation] (see above)

(24) a. *X-ni in.nen-ga tuk-∅-u
X-LOC fate-NOM $\sqrt{}$attach-IN-PST
[No idiomatic interpretation] (see below)

b. X-ni in.nen-o tuk-e-ru
X-LOC fate-ACC $\sqrt{}$attach-TR-NPST
'make a false charge against X'

(25) a. X-ga (Y-no) ki-ni kak-ar-u
X-NOM Y-GEN mind-LOC $\sqrt{}$hook-IN-NPST
'{be/ Y is} concerned about X'

b. (Y-ga) X-o ki-ni kak-e-ru
Y-NOM X-ACC mind-LOC $\sqrt{}$hook-TR-NPST
'{be/ Y is} concerned about X; care about X/ Y cares about X'

(26) a. X-ga (Y-no) hana-ni tuk-∅-u
X-NOM Y-GEN nose-LOC $\sqrt{}$attach-IN-NPST
'X gets up {one's/ Y's} nose; {be/ Y is} fed up with X'
Lit.: 'X sticks to {one's/ Y's} nose'

b. *(Y-ga) X-o hana-ni tuk-e-ru
Y-NOM X-ACC nose-LOC $\sqrt{}$attach-TR-NPST
[No idiomatic interpretation] (see above)

(27) a. *X-ga ko-mimi-ni hasam-ar-u
X-NOM DIM-ear-LOC $\sqrt{}$put.in-IN-NPST
[No idiomatic interpretation] (see below)

b. X-o ko-mimi-ni hasam-∅-u
X-ACC DIM-ear-LOC $\sqrt{}$put.in-TR-NPST
'overhear X; happen to hear X'

The idioms in (18), (22), and (25) alternate in transitivity, behaving as predicted by the present approach. Those in (20), (24), and (27) do not have the intransitive version to begin with and are thus irrelevant to the prediction in question. However, the fact that the transitive versions cannot be formed out of the intransitive idioms, as in (19), (23), and (26), may present a serious challenge to the present approach and needs to be taken care of.

I do not have a fully worked-out analysis of idioms based on transitive/intransitive verbs that only have the intransitive version at present. Yet what is crucially relevant in blocking the transitive idioms in question is that the base intransitive idioms have stative, non-episodic interpretations.[14] Specifically, (19)a is an individual-level predicate

[13] When a PP is an essential part of an idiom, it is lower than an accusative argument. This is what Kishimoto (2008) calls (internal) APPLP, distinct from the one higher than vP. For Kishimoto, it is APPLP rather than PP because *-ni* cannot be replaced by a postposition like *-e*. However, since *-ni* may be fixed as such due to being part of an idiom, I remain agnostic about the nature of *-ni* that comprises an idiom, though I agree that this use of *-ni* is special.

[14] For the following solution to go through, the aspectual/aktionsart properties of idioms in general and alternating idioms in particular need further investigation, which I leave for future research. I am grateful to an anonymous reviewer for bringing up this point.

denoting the property of being skilled in something;[15] (23)a describes a state in which something is within sight or ascribes an ability to see something to someone; and (26)a ascribes a disposition to someone/something. Moreover, the idioms in (28) below, which involve result state interpretations, could be added to the above data set.

(28) a. X-ga doo-ni ir-$\varnothing - u$
X-NOM temple-LOC $\sqrt{}$go.in-IN-NPST
'X reaches a masterly level'

b. too-ga tat-$\varnothing - u$
flower stalk-NOM $\sqrt{}$stand-IN-NPST
'have passed one's prime'
Lit.: 'a flower stalk lengthens/comes out'

These observations suggest that some element contributing to their stative meaning (e.g., a generic or aspectual operator, or other TAM-related elements) must be structurally related to the intransitive vP for the relevant idioms to be formed. If this is the case, it would immediately explain why their transitive counterparts are not available as idioms: Embedding an idiom in these cases is not possible without disrupting a required relation for an intransitive idiom (e.g, $H_{T/A/M} \rightarrow$ v, where $H_{T/A/M}$ is a TAM-related head).

However, the solution I am suggesting needs more elaboration and refinement, and more careful inquiry is no doubt needed about the data set under discussion before making any substantial claims. I leave the matter for future research.

3 Discussion: the status of *-ni*

Discussing two-place unaccusative verbs entering into the transitivity alternation (ones with DP and PP as internal arguments), Takano (2011) argues that a *ni*-phrase in unaccusatives is a PP headed by the postposition *-ni*, while that in (non-idiomatic) ditransitives (some of which are transitives in our analysis) may be marked by the dative case marker.

Given the discussion of the present study, it trivially follows that, if a *ni*-phrase is an integral part of an alternating intransitive idiom, which is of the form [PP V], it is a PP in the transitive version as well. The idioms in (25) are such examples.

Thus, the *ni*-phrase may be a dative DP only in the case of transitive idioms that require a nominal element as essential (i.e., are of the form [DP V]). This is also congenial to the assumption that, in ditransitive contexts, a *ni*-marked animate(-like) argument higher than an accusative argument may be a dative argument (Miyagawa and Tsujioka, 2004).[16] Examples of this class of idioms are given below, with (24)b repeated as (29)a.

(29) a. X-ni in.nen-o tuk-e-ru (= (24)b)
X-LOC fate-ACC $\sqrt{}$attach-TR-NPST
'make a false charge against X'

b. X-ni indoo-o wata-s-u
X-LOC last words-ACC $\sqrt{}$pass-TR-NPST
'give X one's last word (on something)'
Lit.: 'address the last words to X
(X: a newly deceased person)'

c. X-ni yak-i-o ir-e-ru
X-LOC $\sqrt{}$burn-N-ACC $\sqrt{}$go.in-TR-NPST
'teach X a lesson; discipline X; torture X'
Lit.: 'temper X (X: swords, *etc.*)'

d. X-ni kugi-o sas-$\varnothing$-u
X-LOC nail-ACC $\sqrt{\text{sting}}$-TR-NPST
'give X a warning'
Lit.: 'drive a nail into X'

For these idioms to be interpreted as such, the following selectional relations should be minimally satisfied, where $\sqrt{\text{noun}}$ and $\sqrt{\text{verb}}$ represent the relevant nominal and verbal roots, respectively:

(30) Selectional relations for the idioms in (29)

a. v $\rightarrow \sqrt{}$verb
b. v $\rightarrow$ n
c. n $\rightarrow \sqrt{}$noun

Note that, though they are all obligatorily transitive, the selection of v by *i* need not be specified (though it may be), since it is effectively required by the fact that the events described are externally caused through the intervention of an agent (cf. (17)b).

These assumptions would indeed give some room for the possibility of APPL selecting v in the idioms in (29) (with a necessary revision of or addition to the VI rule in (1)a above in order to accommodate cases where APPL selects v) because no required selectional relations in (30) would be disrupted. (Recall the discussion in section 2.)

However, it is notoriously hard to pin down the dative *-ni* as opposed to the postpositional *-ni* in the

[15] Levin and Rappaport Hovav (1995: 96) note that, unlike stage-level properties, individual-level properties, which describe permanent properties, typically cannot be externally caused.

[16] Lexical meanings of verbs also matter in the selection between the dative case marker or the postposition. In general, change-of-possession verbs take dative arguments, while change-of-location verbs take locative PPs. See Kishimoto (2001, 2008, 2010) and Kimura and Morita (2021) for proposals regarding the classification of Japanese ditransitive verbs.

cases under consideration. This is because these idioms allow no more than one *ni*-phrase, unlike "two-goal" constructions, discussed by Miyagawa and Tsujioka (2004), where the high goal and the low goal tend to be animate (i.e., recipient/goal) and inanimate (i.e., goal location), respectively.

However, though an animacy restriction holds for arguments of APPL, P can also take animate arguments, affected or not. That is, P can take as its complement whatever APPL can, making it even harder to pin down which use of *-ni* is involved in the context where only one *ni*-phrase is possible.

Thus, to argue for the possible presence of APPL in the idioms in (29), one should demonstrate that a *ni*-phrase cannot be PP, using a diagnostic like numeral quantifier float, which is possible with DP, but not with PP (Miyagawa, 1989). However, though it needs to be more thoroughly investigated and is left for future research, such examples are hard to construct, as things stand.

Given these considerations, I conclude that, although APPL showing up in place of P with idioms as in (29) is indeed a possibility, it has yet to be seen whether this possibility is instantiated.

Proponents for the dative analysis of *ni*-phrase might object to this conclusion by citing the example pair as in (31) below as evidence that the *ni*-phrase in the active is dative, provided that the *ni*-phrase in (31)a cannot be the passive subject, as in (31)b, if it is a PP.

(31) a. Yakuza-ga tensyu-ni
yakuza-NOM shop owner-LOC
in.nen-o tuk-e-ta
fate-ACC $\sqrt{}$attach-TR-PST
'A yakuza made a false charge against the shop owner.'

b. Tensyu-ga yakuza-kara
shop owner-NOM yakuza-from
in.nen-o tuk-e-rare-ta
fate-ACC $\sqrt{}$attach-TR-PASS-PST
'The shop owner had a false charge made against by a yakuza.'

An alternative analysis without APPL is possible, however. Specifically, it is possible to analyze the *ni*-phrase in the active as a PP argument of vP1, as depicted in (32)a, and the nominative subject in the passive as an argument introduced by *i*, which extends vP1, as in (32)b. This analysis, without APPL, can account for the idiomatic interpretation being kept intact in the passive and the fact that the goal argument can become a passive subject, as (31)b shows. Moreover, as depicted in (32)b, the passive nominative subject introduced by *i* is interpreted as non-agentive (goal or affected goal, in this case) in the presence of the passive VOICE head, as required by thematic uniqueness.[17] Note that, as shown in (32)c, no stacking of more than one argument introducer of the same kind is allowed, which is also due to thematic uniqueness, so no "extra" dative *ni*-phrase is allowed, and thus the *ni*-phrase in the active is unambiguously PP. Therefore, we can explain the dative-like behavior of the PP argument without invoking APPL.

(32) a. Active
$[\text{vP3 DP}_{\text{NOM}}\ [\text{vP1 PP}\ \underline{[\text{vP1 DP}_{\text{ACC}}[\text{v}\ \sqrt{}\ \text{v}\ ]]]}\ \underset{\text{-ER/-EE}}{i}\ ]$

b. Passive
$[\text{PASSP (PP}_{\text{-ER}})\ [\text{vP2 DP}_{\text{NOM}}\ \underline{[\text{vP1}\ ]}\ \underset{\text{-ER/-EE}}{i}\ ]\ \underset{\text{-ER}}{\text{VOICE}_{\text{PASS}}}]$

c. Active (thematic uniqueness violation)
$*[\text{vP3 DP}_{\text{NOM}}\ [\text{vP2 DP}_{\text{DAT}}\ \underline{[\text{vP1}\ ]}\ \underset{\text{-ER/-EE}}{i}\ ]\ \underset{\text{-ER/-EE}}{i}\ ]$

4 Summary

In light of the selection-based approach to idioms, we have argued for an applicativization analysis of the ditransitive/transitive alternation in Japanese and the relevance of stative interpretations to some non-alternating intransitive idioms, which may be amenable to a structural analysis under the present approach. We have also presented an analysis of the status and distribution of *ni*-phrases in transitive idioms, where *i*, instead of APPL, is utilized.

Acknowledgments

I am grateful to the anonymous reviewers for their feedback. The usual disclaimers apply.

References

Bruening, Benjamin. 2010. Ditransitive asymmetries and a theory of idiom formation. Linguistic Inquiry, 41(4):519–562.

Dowty, David. 1991. Thematic proto-roles and argument selection. Language 67(3):547–619.

Embick, David. 2015. The Morpheme: A Theoretical Introduction. De Gruyter Mouton, Boston and Berlin.

Everaert, Martin. 2010. The lexical encoding of idioms. In Malka Rapapport Hovav, Edit

[17] See footnotes one and eleven above. Interpretations that do not survive through inference driven by the constraint on thematic uniqueness are represented by a ~~double strikethrough~~.

Doron, and Ivy Sichel, eds., Lexical Semantics, Syntax, and Event Structure, pp.76–98. Oxford University Press, Oxford.

Folli, Raffaela and Heidi Harley. 2007. Causation, obligation, and argument structure: on the nature of little v. Linguistic Inquiry, 38(2):197–238.

Gehrke, Berit and Louise McNally. 2019. Idioms and the syntax/semantics interface of descriptive content vs. reference. Linguistics, 57(4):769–814.

Hale, Kenneth and Samuel J. Keyser. 1993. On argument structure and the lexical expression of syntactic relations. In Ken Hale and Samuel J. Keyser, eds., The View from Building 20, pp.53–109. MIT Press, Cambridge, MA.

Halle, Morris and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Ken Hale and Samuel J. Keyser, eds., The View from Building 20, pp.111–176. MIT Press, Cambridge, MA.

Inoue, Kazuko. 1976. Henkei-Bunpoo-to Nihongo [Transformational Grammar and Japanese], volume 2. Taishuukan, Tokyo.

Jacobsen, Wesley M. 1992. The Transitive Structure of Events in Japanese. Kurosio Publishers, Tokyo.

Kimura, Hiroko and Chigusa Morita. 2021. Lexical meanings of ditransitive verbs in Japanese. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp.99–107, Shanghai, China. Association for Computational Linguistics.

Kishimoto, Hideki. 2001. The role of lexical meanings in argument encoding: double object verbs in Japanese. Gengo Kenkyu, 120:35–65.

Kishimoto, Hideki. 2008. Ditransitive idioms and argument structure. Journal of East Asian Linguistics, 17(2):141–179.

Kishimoto, Hideki. 2010. The semantic basis of dative case marking in Japanese. Kobe Papers in Linguistics, 7:19–39.

Kratzer, Angelika. 1996. Severing the external argument from its verb. In Johan Rooryck and Lorie Zaring, eds., Phrase Structure and the Lexicon, pp.109–137. Kluwer, Dordrecht.

Landau, Idan. 2010. Saturated adjectives, reified properties. In Malka Rapapport Hovav, Edit Doron, and Ivy Sichel, eds., Lexical Semantics, Syntax, and Event Structure, pp.204–225. Oxford University Press, Oxford.

Landman, Fred. 2000. Events and Plurality: The Jerusalem Lectures. Kluwer, Dordrecht.

Levin, Beth and Malka Rappaport Hovav. 1995. Unaccusativity: At the Syntax-Lexical Semantics Interface. MIT Press, Cambridge, MA.

Matsuoka, Mikinari. 2003. Two types of ditransitive constructions in Japanese. Journal of East Asian Linguistics, 12(2), 171–203.

Miyagawa, Shigeru. 1989. Structure and Case Marking in Japanese. Academic Press, San Diego, CA.

Miyagawa, Shigeru. 1998. (S)ase as an elsewhere causative and the syntactic nature of words. Journal of Japanese Linguistics, 16(1):67–110.

Miyagawa, Shigeru and Takae Tsujioka. 2004. Argument structure and ditransitive verbs in Japanese. Journal of East Asian Linguistics, 13(1):1–38.

Newman, Elise. 2024. When Arguments Merge. MIT Press, Cambridge, MA.

Nishiyama, Kunio. 1998. The morphosyntax and morphophonology of Japanese predicates. Ph.D. Thesis. Cornell University.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. Language, 70(3):491–538.

O'Grady, William. 1998. The syntax of idioms. Natural Language and Linguistic Theory, 16(2):279–312.

Pylkkänen, Liina. 2008. Introducing Arguments. MIT Press, Cambridge, MA.

Schildmier Stone, Megan. 2016. The difference between bucket-kicking and kicking the bucket: understanding idiom flexibility. Ph.D. Thesis. University of Arizona.

Takano, Yuji. 2011. Double complement unaccusatives in Japanese: puzzles and implications. Journal of East Asian Linguistics, 20(3):229–254.

Takehisa, Tomokazu. 2018. On thematically underspecified arguments. McGill Working Papers in Linguistics, 25(1):404–413.

Wood, Jim and Alec Marantz. 2017. The interpretation of external arguments. In Roberta D'Alessandro, Irene Franco, and Ángel J. Gallego, eds., The Verbal Domain, pp.255–278, Oxford University Press, Oxford.

Improving the Efficiency of Long Document Classification using Sentence Ranking Approach

Prathamesh Kokate[1,3], Mitali Sarnaik[1,3], Manavi Khopade[1,3], and Raviraj Joshi[2,3]
[1]Pune Institute of Computer Technology, Pune
[2]Indian Institute of Technology Madras, Chennai
[3]L3Cube Labs, Pune

Abstract

Long document classification poses challenges due to the computational limitations of transformer-based models, particularly BERT, which are constrained by fixed input lengths and quadratic attention complexity. Moreover, using the full document for classification is often redundant, as only a subset of sentences typically carries the necessary information. To address this, we propose a TF-IDF-based sentence ranking method that improves efficiency by selecting the most informative content. Our approach explores fixed-count and percentage-based sentence selection, along with an enhanced scoring strategy combining normalized TF-IDF scores and sentence length. Evaluated on the MahaNews Long Document Classification (LDC) dataset of long Marathi news articles, the method consistently outperforms baselines such as first, last, and random sentence selection. With MahaBERT-v2, we achieve near-identical classification accuracy with just a 0.33 percent drop compared to the full-context baseline, while reducing input size by over 50 percent and inference latency by 43 percent. This demonstrates that significant context reduction is possible without sacrificing performance, making the method practical for real-world long document classification tasks.

1 Introduction

Long document classification is vital in NLP applications such as research, legal, news, and reviews. Transformer models like BERT achieve strong results but are limited by input size and high attention costs (Park et al., 2022; Zaheer et al., 2020), often requiring truncation or complex hierarchical processing (Wagh et al., 2021; Devlin et al., 2018). We propose a data-driven approach (Minaee et al., 2021) that ranks and selects key sentences using TF-IDF (Qaiser and Ali, 2018), treating each sentence as a document and summing term scores (Das and Chakraborty, 2020; Kim and Gil, 2019; Liu et al., 2018a; Das et al., 2023). High-ranking sentences capture domain-specific context while minimizing input length, enabling efficient classification with reduced computational overhead (Figure 1).
We explore multiple strategies for selecting sentences, including the following:

1. Fixed-length selection involves choosing a predefined number of top-ranked sentences, with evaluations conducted for 1, 2, 3, 4, and 5 sentences.

2. Percentage-based selection refers to the selection of a specific percentage of top-ranked sentences, varying from 10% to 100% in increments of 10%.

3. Weighted ranking combines normalized TF-IDF scores with sentence length to balance importance and informativeness, exploring different weighting factors to identify the optimal configuration.

To evaluate the effectiveness of these strategies, we conduct extensive experiments on the MahaNews[1] dataset (Mittal et al., 2023; Aishwarya et al., 2023), a corpus of long Marathi news articles categorized by topic. Using MahaBERT[2](marathi-bert-v2) (Joshi, 2022), we train and test models on reduced-context versions of the dataset and compare the classification performance across different selection methods. Our results demonstrate that TF-IDF-based ranking significantly outperforms simpler selection strategies, such as choosing the first, last, or randomly sampled sentences. Additionally, integrating length-aware weighting further enhances accuracy, while context reduction leads

[1]https://github.com/l3cube-pune/indic-nlp/tree/refs/heads/main/L3Cube-IndicNews/Marathi/LDC
[2]https://huggingface.co/l3cube-pune/marathi-bert-v2

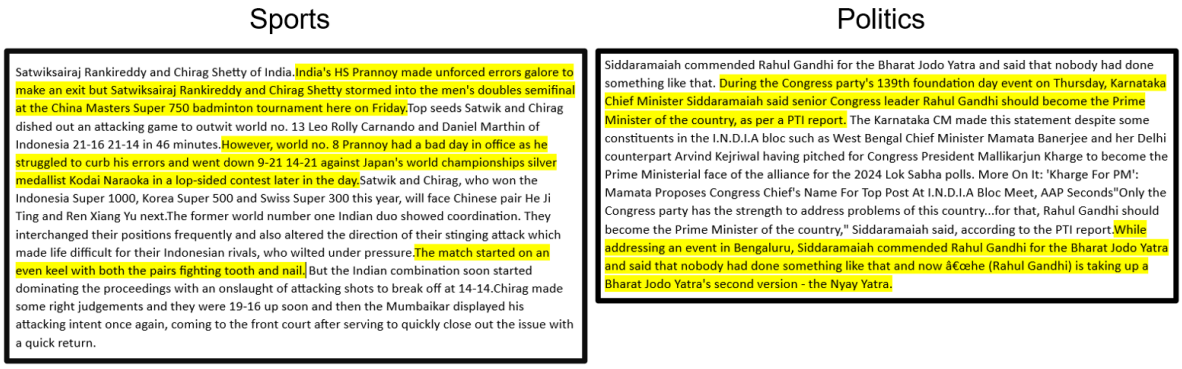


Figure 1: Illustration of key idea — selective sentence processing for efficient document classification. The figure presents two example paragraphs, representing only a portion of the long documents: one related to sports and the other to politics. In each case, the most semantically relevant and contextually informative sentences are highlighted. These highlighted sentences contain domain-specific cues (e.g., sports activities or political entities) that enable accurate classification without processing the full document. This demonstrates that selective sentence extraction can preserve classification performance while reducing computational overhead.

to a substantial decrease in inference time without compromising performance.

1.1 Key Contributions

- We propose a novel TF-IDF-based sentence ranking and context reduction strategy to improve the efficiency of BERT models for long document classification without altering the model architecture, significantly reducing processing time for large text inputs.
- We evaluate multiple sentence selection techniques such as fixed-length, percentage-based, and weighted ranking, analyzing their trade-offs in balancing efficiency and classification accuracy.
- Experiments on the MahaNews dataset show that ranked selection consistently outperforms naive approaches while maintaining accuracy and significantly reducing inference time. Specifically, the performance of selection strategies follows the order: ranked > first > random > last. Notably, selecting sentences from the beginning of the document serves as a strong baseline.
- Our findings reveal an optimal balance between input length, accuracy, and computational efficiency, demonstrating that selecting a subset of ranked sentences can achieve near-full-document classification performance.

By systematically analyzing context reduction techniques, our work provides a practical and efficient alternative to architectural modifications for long document classification in transformer-based models.

2 Related Work

Long documents contain extensive information, making direct processing with traditional classification models computationally expensive and time-consuming. To improve efficiency, existing methods generally fall into two categories: data-based approaches and model-based approaches.

2.1 Model Based Approaches

Handling long document classification efficiently requires balancing model complexity with computational feasibility. Model-based techniques addressing this challenge include sparse attention mechanisms, quantization, recurrent architectures, and normalization strategies. Sparse attention mechanisms enable transformer models to process significantly longer inputs while retaining the advantages of full-attention models (Pham and The, 2024). By incorporating global tokens for capturing overall context, local tokens for nearby interactions, and random tokens to enhance global coverage, these mechanisms effectively reduce memory and computation costs from quadratic to linear (Martins et al., 2020), making them particularly useful for handling extensive input sequences.
Beyond attention mechanisms, reducing computational demand can be achieved through quantization, which lowers the precision of model weights to save memory. For example, Q8 BERT employs

8-bit weights instead of the standard 32-bit, using techniques such as quantization-aware training (Zafrir, 2019). This approach significantly reduces model size while maintaining accuracy, making it suitable for deployment in resource-constrained environments. Recurrent architectures like Long Short-Term Memory (LSTM) networks have also been explored for capturing long-term dependencies (Teragawa et al., 2021; Putri and Setiawan, 2023). While LSTMs excel at preserving sequential information, their sequential nature limits parallelization, giving transformer-based models an advantage in scalability.

To further improve the stability and efficiency of transformer models, pre-layer normalization is applied. This technique normalizes activations before the attention mechanism, mitigating gradient instability and accelerating convergence (Beltagy et al., 2020). By improving training dynamics, pre-layer normalization enhances the robustness of deep transformer architectures, making them more suitable for long document classification. Combining sparse attention for efficiency, quantization for reduced computational demand, recurrent mechanisms for sequence retention, and pre-layer normalization for stability enables modern NLP models to effectively process long documents while optimizing performance and resource utilization (Al-Qurishi, 2022).

2.2 Data Based Approaches

Unlike model-based approaches that improve architectures and algorithms, data-centric methods optimize the training and testing data pipeline to boost performance without altering the model. For example, Discriminative Active Learning (DAL) reduces labeling effort by selecting informative instances near the decision boundary, ensuring labeled and unlabeled data distributions align in the learned space (Bamman and Smith, 2013). Another strategy tackles transformer input limits by splitting long documents, processing chunks individually, and aggregating results via hierarchical models (Yang et al., 2016; Khandve et al., 2022) or hierarchical attention (Yang et al., 2016). We adopt a data-centric approach for its easy integration, domain-agnostic nature, robustness against model biases, and scalability (Song, 2024; Moro, 2023). By minimizing contextual information during training and inference (He, 2019; Liu et al., 2018b; Tay et al., 2021) through selective input curation, our method adapts across domains and models without architectural changes (Li et al., 2018; Prabhu et al., 2021; Sun et al., 2020).

3 Methodology

The datasets utilized in our experiments are sourced from L3Cube's IndicNews corpus a multilingual text classification dataset curated for Indian regional languages. The MahaNews corresponds to the Marathi subset of the IndicNews dataset. The corpus covers news headlines and articles in 11 prominent Indic languages, with each language dataset encompassing 10 or more news categories. We have made use of the LDC dataset which consists of full articles with their categories (Mittal et al., 2023; Aishwarya et al., 2023).

Our methodology focuses on optimizing input size while preserving classification performance using the Marathi LDC dataset, which consists of full-length articles in Marathi (Jain et al., 2020). We begin by tokenizing each article into individual sentences, followed by computing the TF-IDF score for each sentence. The sentences are then ranked based on their scores, and the context is reduced by selecting top-ranked sentences. To achieve this, we explore various sentence selection strategies. Instead of using the entire article, the selected sentences are fed into the MahaBERT model for classification.

3.1 Training and Testing

We aim to improve classification efficiency by reducing input text during training and inference while maintaining performance comparable to full-document processing. On the full LDC dataset, the MahaBERT model, fine-tuned on L3Cube-MahaCorpus and other public Marathi datasets achieved 94.706% accuracy. Our objective is to approach this accuracy using reduced context inputs. The Marathi LDC dataset contains 20,425 training samples, 2,550 testing samples, and 2,548 validation samples used to enhance model accuracy.

Sentence Selection Techniques

We evaluated several sentence selection strategies, ranging from simple selection methods to a novel TF-IDF-based method. These approaches aim to retain the most informative parts of each document which are used to train the model and subsequently test it.

- **First Few Sentences Selection:** In this

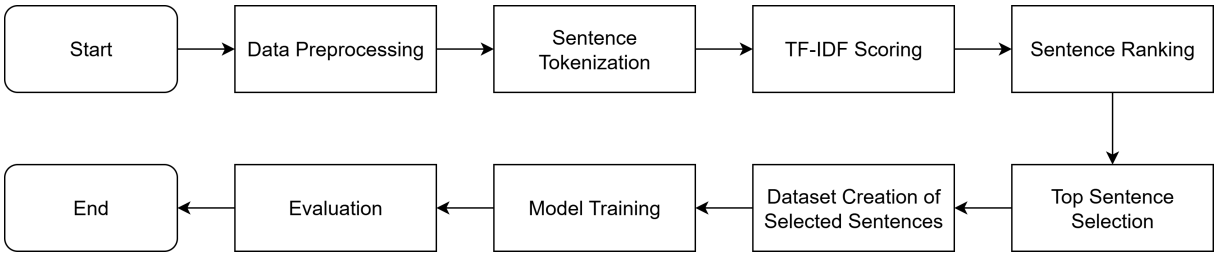


Figure 2: Workflow of ranked approach for sentence selection — The diagram illustrates a ranked sentence selection workflow starting from raw text input. Data is preprocessed and split into sentences, which are scored using TF-IDF. Top-ranked sentences are selected to create a training dataset. This dataset is used for model training and evaluation before concluding the process.

method, only a specified percentage of the initial sentences from each article is selected. This leverages the observation that the opening sentences often contain summaries or key contextual information critical for classification.

- **Last Few Sentences Selection:** Conversely, this method selects only the last portion of sentences from each article. The rationale is that concluding sentences often include detailed analysis or summaries, which may also be useful for accurate classification.

- **Random Sentences Selection:** Here, sentences are randomly selected from across the article. While this approach is computationally efficient and allows for diverse content selection, it is unreliable, as critical information may be excluded, leading to inconsistent classification performance across samples.

While these methods are straightforward and easy to implement, they can fail to consistently capture the document's most relevant content, as important sentences may appear in various parts of the text.

3.2 TF-IDF-Based Ranking and Selection

Random sentence selection, though efficient, ignores sentence importance based on factors like distinctiveness and semantic relevance, leading to inconsistent accuracy in long document classification. We address this with a novel TF-IDF based sentence selection method that ranks sentences by informative value, reducing training and inference time while maintaining high accuracy.

General Flow of the Method

The sentences of each article are preserved in their original order and tokenized using a language-specific tokenizer, such as the Indic NLP tokenizer for the IndicNews dataset.

For TF-IDF score calculation, each sentence is treated as an individual document in the context of determining Term Frequency (TF) and Inverse Document Frequency (IDF). This approach identifies terms that occur frequently within a sentence but are rare across others in the same article, thereby quantifying the importance of each term.

The TF-IDF scores computed for each sentence result in an ordered array where the most informative sentences appear at the top.

Score Computation

The score of a sentence S_i can be calculated as the sum of the TF-IDF scores of all terms t_j within the sentence.

Formally, the score $\text{Score}(S_i)$ is defined as:

$$\text{Score}(S_i) = \sum_{t_j \in S_i} \text{TF-IDF}(t_i)$$

Where,

$$\text{TF-IDF}(t_i) = \text{TF}(t_i) \cdot \text{IDF}(t_i)$$

Definitions

1. **Term Frequency (TF):**

$$\text{TF}(t_j) = \frac{\text{Frequency of } t_j \text{ in } S_i}{\text{Total number of terms in } S_i}$$

2. **Inverse Document Frequency (IDF):**

$$\text{IDF}(t_j) = \log\left(\frac{N}{1 + \text{Sentence frequency of } t_j}\right)$$

where N is the total number of sentences in the article, and the document frequency is the number

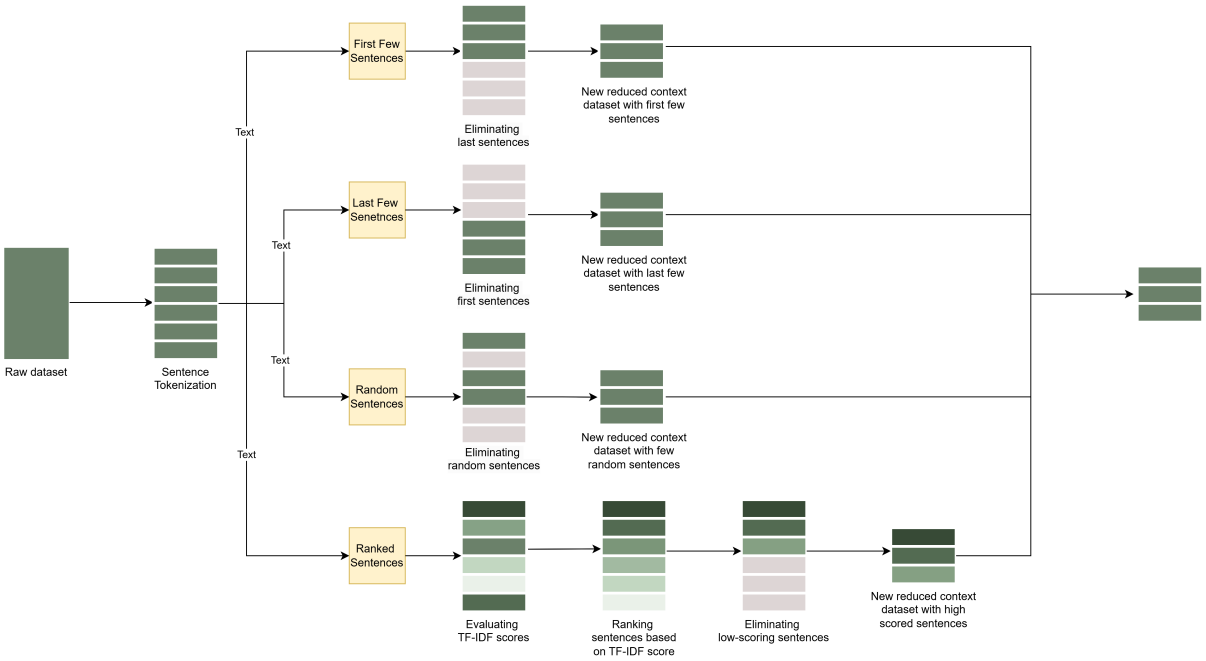


Figure 3: Sentence selection approaches — The image illustrates various sentence selection approaches used for context reduction. Methods include selecting the first few sentences, last few sentences, random sentences, and ranked sentences. In the ranked approach, sentences are scored using TF-IDF and selected based on informative context.

of sentences containing t_j.
This formula ensures that the importance of each sentence is derived from the significance of its terms within the context of the article.
The above approaches, to select sentences, for model are as depicted in Figure 3.

Optimal Number Sentence Selection
Selecting the optimal number of sentences involves balancing efficiency and classification accuracy. Several approaches are considered for determining the most informative subset of sentences:

- **Top-ranked sentence selection:** The highest-ranked sentence, based on its TF-IDF score, is used to evaluate the effectiveness of minimal context in classification.
- **Incremental context expansion:** The top two, three, four, and five ranked sentences are examined to assess the impact of increasing contextual information on classification accuracy and to identify the point of diminishing returns.
- **Percentage-based selection:** Top-ranked sentences are progressively selected in increments of 10%, aiming to find an optimal balance between efficiency and performance. This method is particularly effective for documents.

These approaches help refine sentence selection strategies to enhance both computational efficiency and model performance while minimizing unnecessary information.

3.3 Length Normalization

Normalization scales features to a common range, ensuring fair contribution and preventing dominance due to scale differences. In our case, it adjusts sentence TF-IDF scores to avoid bias toward longer sentences, which would otherwise rank higher simply due to more terms rather than higher informative content.
To ensure fair sentence ranking, different approaches balance sentence length and TF-IDF scores:

- **Length Normalization:** Divides the total TF-IDF score by sentence length to prevent longer sentences from being unfairly ranked higher.
- **Weighted Balancing:** Uses a dynamic weighted formula to balance TF-IDF score and sentence length.

Length normalization (dividing total TF-IDF by token count) ranks sentences by average term im-

portance, allowing fair comparison across different lengths. However, after analyzing the selected sentences it was observed that normalization introduced an inverse bias toward shorter sentences. To address this, we introduce a weighted balancing approach that incorporates an additional factor for more balanced and meaningful scoring.

Balancing Length Factor

To achieve a fair ranking, we needed a mechanism that dynamically adjusts the influence of TF-IDF scores and sentence length. This approach creates a flexible ranking mechanism, where the relative importance of each factor can be controlled to ensure an optimal trade-off between uniqueness and context.

To balance this bias and achieve a trade-off between the two extremes, the following formula was introduced:

$$\text{Score} = (\lambda_1 \cdot \text{Normalized\_TF\_IDF}) + (\lambda_2 \cdot \text{length})$$

$$\text{Where, } \lambda_1 = 1 - \lambda_2 \quad \text{and} \quad 0 \leq \lambda_1, \lambda_2 \leq 1$$

These are weights to control the relative importance of **Normalized_TF_IDF** and **length** in the final ranking.

- $\lambda_1 > \lambda_2$: Focus on sentences with unique terms (higher TF-IDF score).
- $\lambda_2 > \lambda_1$: Prioritize sentences with more context (lengthier ones).

This formula effectively balances the biases introduced by normalization and sentence length by distributing the total weight between the two factors. Since λ_1 = 1 - λ_2 , increasing the weight on one factor automatically reduces the influence of the other, ensuring a controlled trade-off. If λ_1 is higher, the ranking favors sentences with higher TF-IDF scores, emphasizing term uniqueness. Conversely, if λ_2 is higher, longer sentences with more contextual richness are prioritized. This dynamic weighting mechanism allows for fine-tuning based on the specific needs of the classification task, preventing extreme biases toward either short or long sentences.

4 Results and Discussion

4.1 Number of Sentences Approach

The results in Table 1 and Figure 4 show that accuracy improves as more sentences are selected, with the order: ranked > first > random > last. In the ranked method, accuracy peaks at three sentences. Selecting a fixed number of top-ranked sentences keeps inference time nearly constant and achieves **94.19%** accuracy with just 5 sentences only **0.544%** below the full-context baseline of 94.706%. This shows that substantial context reduction is possible with minimal accuracy loss. To further refine this, we introduce a normalization strategy that combines normalized TF-IDF scores with sentence length to capture both relevance and informational content.

Sentence(s)	First	Last	Random	Ranked
1	**90.70%**	81.64%	75.76%	90.35%
2	93.01%	87.60%	90.31%	**93.17%**
3	93.17%	89.76%	91.49%	**93.64%**
4	92.82%	91.09%	91.72%	**94.00%**
5	93.56%	91.64%	92.70%	**94.19%**

Table 1: Sentence-wise Accuracy Results — The table shows accuracy across different sentence selection strategies (first, last, random, ranked) for 1 to 5 selected sentences. Results indicate that the ranked approach performs best, followed by first, random, and last, highlighting the importance of the selection method and sentence count on model performance.

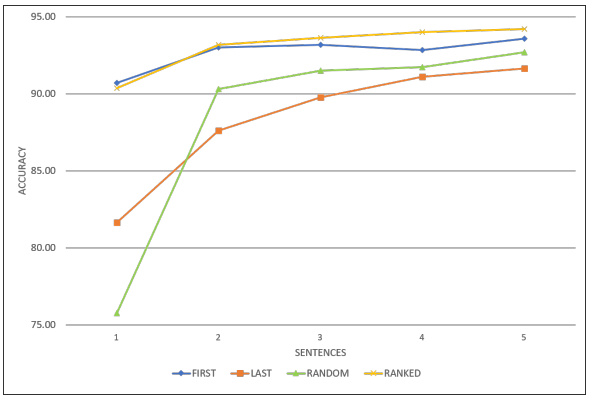


Figure 4: Sentence-wise accuracy graph — The graph visualizes sentence-wise accuracy for different selection methods: first, last, random, and ranked. It plots accuracy against the number of selected sentences (1 to 5). The graph shows that ranked > first > random > last.

Normalization Results

Table 2 shows that combining normalized TF-IDF scores with sentence length yields peak accuracy at λ_2 = 0.7, achieving **94.07%** with 4 sentences. Longer sentences prove more informative in minimal contexts, while λ_2 values of 0.2 and 0.7 best balance relevance and length. Accuracy stabilizes as sentence count increases.

Sentence(s)	0.2 (λ_2)	0.5 (λ_2)	0.7 (λ_2)	1.0 (λ_2)
1	89.11%	88.94%	89.10%	**89.26%**
2	**92.82%**	91.86%	92.73%	92.23%
3	93.79%	93.81%	93.78%	**93.82%**
4	93.95%	93.36%	**94.07%**	93.59%
5	93.47%	93.56%	**93.67%**	93.32%

Table 2: Normalized sentence-wise accuracy results — The table presents normalized sentence-wise accuracy results for the ranked sentence selection method. It shows results across different values of λ_2 ranging from 0.2 to 1.0. $\lambda_2 = 0.7$ provides optimal performance for different sentence counts.

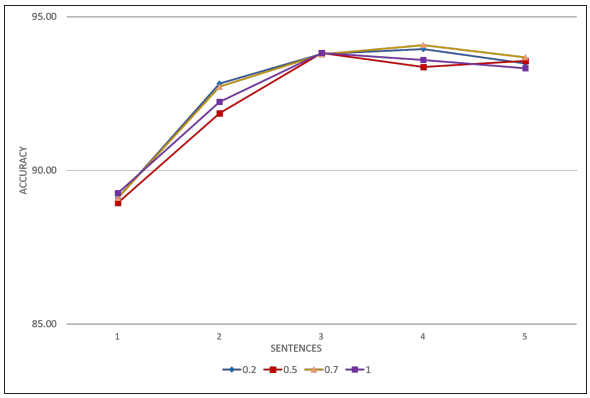


Figure 5: Normalized sentence-wise accuracy graph — The graph shows normalized sentence-wise accuracy for the ranked selection method. The x-axis represents the number of selected sentences, while the y-axis shows the corresponding accuracy. Each line corresponds to a different λ_2.

Sentence(s)	Ranked	Ranked Normalized
1	90.35%	89.11%
2	93.17%	92.82%
3	93.64%	93.82%
4	94.00%	94.07%
5	94.19%	93.67%

Table 3: Comparison of Ranked and Ranked-Normalized Results — The table compares the accuracy of sentence selection using ranked and ranked-normalized methods. It shows that normalization has little impact when the number of selected sentences is low.

Table 3 compares basic ranked selection with normalized ranking, showing minimal improvement when few sentences are selected.

4.2 Data Percentage Approach

Table 4 and Figure 6 illustrate the accuracy achieved by selecting first, last, random, and ranked percentages of sentences from documents. Using the full-length documents for and testing yields an

Percentage	First	Last	Random	Ranked	Ranked Normalized
10%	90.74%	71.76%	86.00%	**91.80%**	91.14%
20%	93.25%	87.88%	90.31%	93.41%	**93.64%**
30%	93.19%	90.31%	92.22%	93.29%	**93.80%**
40%	93.58%	91.96%	92.82%	93.98%	**94.39%**
50%	94.19%	92.98%	93.33%	**94.51%**	94.04%
60%	**94.31%**	92.86%	93.05%	**94.31%**	93.60%
70%	94.62%	94.19%	94.31%	93.92%	**94.47%**
80%	**94.43%**	93.45%	94.23%	94.15%	94.15%
90%	94.50%	93.56%	94.03%	94.11%	**94.90%**
100%	94.35%	94.35%	94.11%	94.63%	**94.78%**

Table 4: Percentage-wise accuracy results — The table presents percentage-wise accuracy results for sentence selection approaches at coverage levels from 10% to 100%. It compares First, Last, Random, Ranked, and Ranked Normalized methods. At 100% coverage, minor accuracy variations (<1%) arise from sentence reordering in Random and Ranked methods, whereas First and Last preserve the original order, yielding identical results.

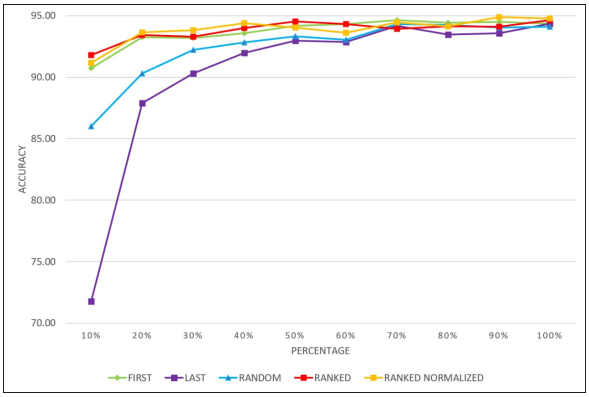


Figure 6: Percentage-wise accuracy graph — The graph visualizes percentage-wise accuracy for different sentence selection methods as sentence coverage increases from 10% to 100%. The x-axis represents the percentage of selected sentences, while the y-axis shows accuracy. Each line corresponds to a method: First, Last, Random, Ranked, and Ranked Normalized. The graph highlights how accuracy improves with more context and which methods are most effective.

accuracy of 94.706%, which serves as the baseline for comparison. At 100% coverage, accuracies are nearly identical across methods, with variations of less than 1%. These slight differences arise from sentence reordering in Random and Ranked selections, which alters the semantic flow and impacts model interpretation, whereas First and Last preserve the original order, yielding identical results. Importantly, by reducing the context to just 40 to

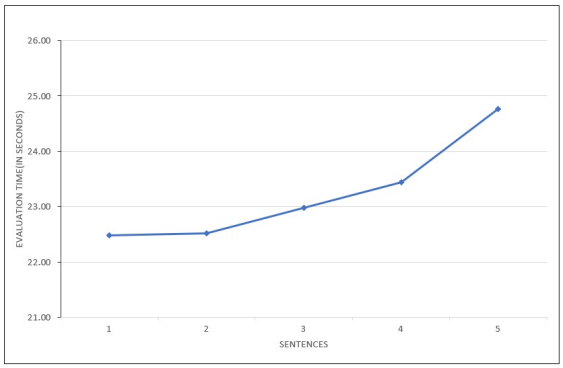


Figure 7: Evaluation time graph for sentence-wise selection — The graph represents the relationship between evaluation time (in seconds) and the number of sentences considered during sentence-wise selection. The x-axis denotes the number of sentences, while the y-axis shows the corresponding evaluation time required. The graph typically highlights a trend where evaluation time increases as the number of sentences grows

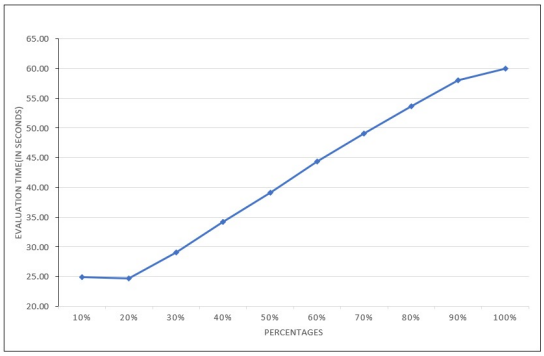


Figure 8: Evaluation time graph for percentage-wise selection — The graph illustrates the relationship between evaluation time (in seconds) and the percentage of sentences selected during percentage-wise selection. The graph demonstrates a trend where evaluation time changes based on the percentage selected.

50 percent of the original document, we are still able to achieve an impressive accuracy of **94.39%**, remarkably close to the base accuracy of 94.706%, demonstrating that its performance is competitive with approaches that use the full document context. With reduced context sizes, the Ranked Selection method consistently outperforms other techniques, such as First, Last, and Random selection. As the context size increases, the performance of all methods converges, yielding similar results. This convergence indicates that the ranked selection method is particularly effective in enhancing accuracy when operating with smaller context windows. In this setting, normalization shows a positive impact, enhancing performance in most cases.

4.3 Inference Time

Context reduction aims to minimize inference time while preserving accuracy. Using TF-IDF, we dynamically adjust sequence length to match document content, avoiding inefficiencies from fixed limits like BERT's 512 tokens. Figures 7 and 8 show testing time variations across different context lengths.

Figure 7 shows a positive correlation between the number of sentences and evaluation time. While the increase is modest from 1 to 3 sentences, it becomes more pronounced from sentence 4 onward, indicating that evaluation time grows increasingly with higher sentence counts.

Figure 8 shows evaluation time stays stable from 10–20%, then rises sharply from 30%, peaking at 100%. At 40% context, our method reaches 94.39% accuracy, only 0.33% below the full-context baseline, while cutting inference latency by 43%. This highlights an efficient speed–accuracy trade-off for scalable, real-world use.

5 Conclusion

We propose an efficient approach to long document classification using sentence selection techniques that reduce input size while maintaining accuracy comparable to full-context models. Strategies include first/last sentence selection, random sampling, and TF-IDF-based ranking.

On the Marathi LDC dataset from L3Cube's IndicNews collection, our method significantly reduces computational costs without sacrificing performance, with TF-IDF ranking proving especially effective. We examined the trade-off between input size and accuracy, finding that selecting a proportion of high-ranking sentences yields better efficiency–performance balance than a fixed number, and that normalization can further improve results. Overall, our approach is scalable, resource-efficient, and adaptable, with potential for domain-specific selection or hybrid models to refine input representation in future work.

6 Limitations

While our MahaBERT-based model captures deep semantics from selected sentences, the TF-IDF based selection relies solely on term frequency,

ignoring contextual relationships.
Future work could incorporate semantic embeddings to improve relevance and reduce redundancy. Additionally, discourse parsers or coherence models may help maintain logical flow, preserve essential context, and enhance interpretability.

Acknowledgement

This work was undertaken with the mentorship of L3Cube, Pune. We sincerely appreciate the invaluable guidance and consistent encouragement provided by our mentor during this endeavor.

References

Mirashi Aishwarya, Sonavane Srushti, Lingayat Purva, Padhiyar Tejas, and Joshi Raviraj. 2023. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 442–449.

Muhammad Al-Qurishi. 2022. Recent advances in long documents classification using deep-learning. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*.

David Bamman and Noah Smith. 2013. New alignment methods for discriminative book summarization.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Bijoyan Das and Sarit Chakraborty. 2020. An improved text sentiment classification model using tf-idf and next word negation.

Mamata Das, K. Selvakumar, and P.J.A. Alphonse. 2023. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jun He. 2019. Long document classification from local word glimpses via recurrent attention learning. *IEEE Access*.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, and 1 others. 2020. Indic-transformers: An analysis of transformer language models for indian languages.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Snehal Ishwar Khandve, Vedangi Kishor Wagh, Apurva Dinesh Wani, Isha Mandar Joshi, and Raviraj Bhuminand Joshi. 2022. Hierarchical neural network approaches for long document classification. In *Proceedings of the 2022 14th International Conference on Machine Learning and Computing*, pages 115–119.

Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*.

Chao Li, Yanfen Cheng, and Hongxia Wang. 2018. A novel document classification algorithm based on statistical features and attention mechanism.

C. z. Liu, Y. x. Sheng, Z. q. Wei, and Y. Q. Yang. 2018a. Research of text classification based on improved tfi-df algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*.

L. Liu, K. Liu, Z. Cong, J. Zhao, Y. Ji, and J. He. 2018b. Long length document classification by local convolutional feature aggregation. *Algorithms*.

André F. T. Martins, António Farinhas, Marcos Treviso, and 1 others. 2020. Sparse and continuous attention mechanisms.

S. Minaee, N. Kalchbrenner, E. Cambria, and 1 others. 2021. Deep learning based text classification: A comprehensive review.

Saloni Mittal, Vidula Magdum, Sharayu Hiwarkhedkar, Omkar Dhekane, and Raviraj Joshi. 2023. L3cube-mahanews: News-based short text and long document classification datasets in marathi. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 52–63. Springer.

G. Moro. 2023. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Linh Manh Pham and Hoang Cao The. 2024. Lnlf-bert: Transformer for long document classification with multiple attention levels. *IEEE Access*.

Sumanth Prabhu, Moosa Mohamed, and Hemant Misra. 2021. Multi-class text classification using bert-based active learning.

Bella Adriani Putri and Erwin Budi Setiawan. 2023. Topic classification using the long short-term memory (lstm) method with fasttext feature expansion on twitter.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: Use of tf-idf to examine the relevance of words to documents. *IJCA*.

B. Song. 2024. State space models based efficient long documents classification. *Journal of Intelligent Learning Systems and Applications*.

Chi Sun, Xipeng Qiu, Yige Xu, and 1 others. 2020. How to fine-tune bert for text classification?

Y. Tay, M. Dehghani, S. Abnar, and 1 others. 2021. Long range arena: A benchmark for efficient transformers.

Shoryu Teragawa, Lei Wang, and Ruixin Ma. 2021. A deep neural network approach using convolutional network and long short term memory for text sentiment classification. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*.

Vedangi Wagh, Snehal Khandve, Isha Joshi, Apurva Wani, Geetanjali Kale, and Raviraj Joshi. 2021. Comparative study of long document classification. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, pages 732–737. IEEE.

Zichao Yang, Diyi Yang, Chris Dyer, and 1 others. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

O. Zafrir. 2019. Q8bert: Quantized 8bit bert.

M. Zaheer, J. Ainslie, G. Guruganesh, and 1 others. 2020. Big bird: Transformers for longer sequences. In *Proceedings of NeurIPS*, pages 702–709.

Diverse Target Representations for Language Models with Word Difference Representations

DongNyeong Heo
Handong Global University
Pohang, South Korea
sjglsks@gmail.com

Daniela Noemi Rim
Handong Global University
Pohang, South Korea
danielarim@handong.ac.kr

Heeyoul Choi
Handong Global University
Pohang, South Korea
hchoi@handong.edu

Abstract

Language modeling (LM) serves as the foundational framework underpinning remarkable successes of recent text generation tasks. In this paper, we study the potential benefits of incorporating diverse and contextualized target representations during LM training, in contrast to conventional approaches that rely on fixed target representations for words. We hypothesize that the use of diverse target representations can enhance the generalizability of LM training. To examine this hypothesis, we introduce the word difference representation (WDR) transformation function, which provides diverse target representations for words in LM. In addition, we propose a simple N-gram prediction framework and an ensemble method to facilitate the WDR approach, and manifest the potential of the WDR approach. Through extensive experiments across various model architectures—including large language models and diffusion models—and multiple benchmark datasets, we empirically validate our hypothesis and demonstrate the practical benefits of our proposed methodologies in terms of improved text generation performance.

1 Introduction

With the remarkable advancements in deep learning techniques (Radford et al., 2019; Brown et al., 2020), language modeling (LM) has become a central component in text generation tasks, such as neural machine translation (NMT) and question answering chatbots. In general, an LM model processes given context words through a neural network to output a representation. The inner product between this representation and the weight matrix of the final logit layer produces scores for each word (Bengio et al., 2000). During training, in order to achieve a high score for the target word after the inner product, the neural network is optimized to output a representation similar to the target word's assigned representation in the vector space of the logit layer's weights.

The assigned representation of each word solely represents the target of the word. In other words, the LM task is a sequential prediction of the unique target representation of each word within a sentence. In contrast, tasks such as image generation that has almost infinite variety of target image representations, the variety of the target text representation is limited to the vocabulary size regardless of the context. Based on this limitation, we pose the following question: *"Would it be beneficial to the training process if we* ***provide different target representations*** *for the same word depending on the context?"*. Regarding the practical benefit of this question, we expect that incorporating diverse target representations would lead to greater diversity in gradient computations during backpropagation. Based on prior analysis (Yin et al., 2018), the increased gradient diversity can enhance the generalizability of the model.

In this paper, we explore the question above by proposing a contextual transformation function - word difference representation (WDR). WDR contextualizes subsequent words (N-gram) through an arithmetic subtraction operation, and we incorporate WDRs as additional, diverse target representations alongside the original ones. To verify the effectiveness of the WDR approach in contrast to baseline approaches, we first develop a simple N-gram prediction framework with minimal modifications to the conventional LM model, and then we apply the WDR approach to the simple N-gram framework. Furthermore, we propose an ensemble method that leverages N-gram predictions to enhance next-word prediction, which is the primary objective of the conventional LM model.

We experimented with our proposed methodologies across multiple baseline model architectures, such as conventional Transformer-based models (Vaswani et al., 2017), large language models (GPT-

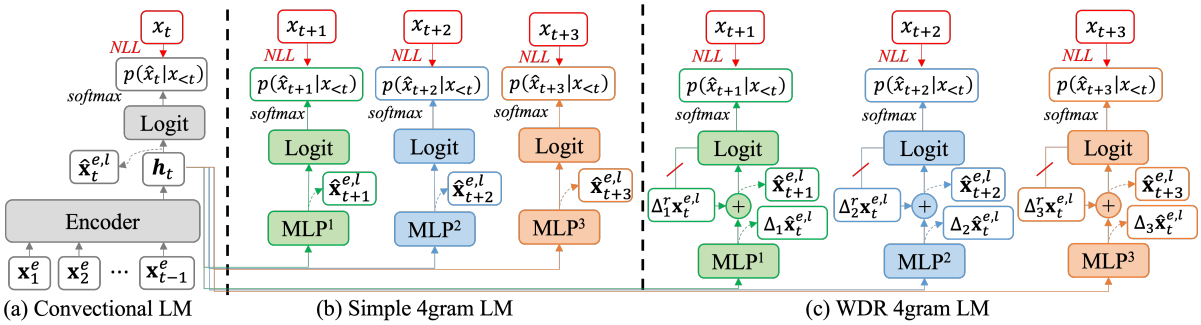


Figure 1: Model illustrations of (a) conventional LM, (b) simple N-gram LM, and (c) WDR N-gram LM when $N = 4$. Note that all of the drawn logit layers above the MLP layers share the parameters. Red diagonal lines in (c) on lines from logit layer to $\Delta_i^r \mathbf{x}_t^{e,l}$ indicate gradient detaching operation.

NEO series) (Black et al., 2021), and the text diffusion model (Ye et al., 2023; Gao et al., 2022), using multiple benchmark datasets for LM and various conditional text generation tasks. In the experiment results, we empirically validate that applying WDR enhances gradient diversity and boosts performance while causing only a slightly increased parameter count and computational cost. These findings provide evidence that incorporating diverse target representations is beneficial for general text generation tasks. Additionally, our results indicate that our ensemble method is also advantageous compared to conventional LM models.

2 Background

2.1 Language Modeling

As background knowledge, in this section we describe the conventional training framework of neural network based LMs.

A sentence consists of words $X = \{x_1, x_2, \cdots, x_T\}$, where $x_t \in \mathcal{V}$, and T and $\mathcal{V}$ indicate the length of the sentence and the vocabulary set, respectively. Conventional LMs compute the likelihood of a word conditioned on its preceding words in the sentence, $p(x_t|x_{<t})$.

First, words are mapped to embedding vectors (Mikolov et al., 2013), and the encoded hidden state at time-step t is formulated as follows:

$$\mathbf{h}_t = Enc_\theta(\{\mathbf{x}_1^e, \mathbf{x}_2^e, \cdots, \mathbf{x}_{t-1}^e\}) \in \mathbb{R}^d, \quad (1)$$

where $\mathbf{x}_t^e \in \mathbb{R}^d$ means the embedded vector of x_t, Enc_θ is an encoder model with its parameter set θ, and d is the dimension of the encoded hidden state and the embedding vector spaces. Recently, most LMs use Transformer (Vaswani et al., 2017) as their encoder architecture. After encoding, the hidden state is linearly transformed to a logit value of each word in $\mathcal{V}$. Finally, the likelihood of the predicted word is calculated as follows:

$$p(\hat{x}_t|x_{<t};\theta) = softmax(\hat{\mathbf{x}}_t^l),$$
$$\hat{\mathbf{x}}_t^l = \mathbf{W}^l \mathbf{h}_t = \mathbf{W}^l \hat{\mathbf{x}}_t^{e,l}, \quad (2)$$

where $\mathbf{W}^l \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the weight matrix of the logit layer.

Throughout this paper, we consider the logit layer's weight matrix as a word embedding set, $\mathbf{W}^l = [\mathbf{x}_1^{e,l}, \mathbf{x}_2^{e,l}, \cdots, \mathbf{x}_{|\mathcal{V}|}^{e,l}]^\top$, where each d-dimensional embedding vector is aligned with the target word. The superscript notation (e, l) means that it is an embedding vector at the logit layer. From this point of view, the hidden state, $\mathbf{h}_t$, could be understood as the predicted word embedding, $\hat{\mathbf{x}}_t^{e,l}$. The inner product of $\mathbf{W}^l$ and $\hat{\mathbf{x}}_t^{e,l}$ outputs the predicted score of each word based on the similarity between the embedding and predicted vectors.

Then, the model learns to minimize the negative log-likelihood (NLL) loss as follows:

$$\mathcal{L}(X, \theta) = -\sum_{t=1}^{T} \log p(\hat{x}_t = x_t|x_{<t};\theta). \quad (3)$$

This process is illustrated in Fig.1(a).

2.2 N-gram Prediction

The conventional LM training framework follows the '*next word prediction*' approach that predicts a word given the whole previous words. Despite the successes, this approach might lead models to overfit to local dependencies rather than capturing long-term dependencies between words. This tendency arises from the strong dependencies found in some phrases or word pairs, such as "*Barack Obama*" and "*Harry Potter*" (Qi et al., 2020).

A way of mitigating this problem involves predicting not solely the next word but also subsequent words in later time-steps such as N-gram prediction. Researchers (Sun et al., 2019; Joshi et al., 2020; Xiao et al., 2020; Qi et al., 2020) have adopted this N-gram prediction methodology for the masked language modeling during the pre-training phase of large language models (Devlin et al., 2018). Similar approaches have been applied to the NMT task (Shao et al., 2018; Ma et al., 2018; Shao et al., 2020). To utilize the N-gram prediction method, previous works significantly modified the model architecture, the loss function, or the vocabulary set. In this paper, as the development base of our main idea (diverse target representations with WDR), we introduce a 'simple N-gram prediction' method that requires the least modifications from the conventional LM so that is transparent to analyze the main idea's advantage.

3 Proposed Methods

In this section, we introduce all of our proposed methodologies: (1) a simple N-gram prediction framework that becomes the development's base, (2) the main WDR idea and its application to several LM architectures, and (3) ensemble method applicable to N-gram prediction.

3.1 Simple N-gram Prediction

The core idea of our proposed simple N-gram prediction method is adding a multi-layer perception (MLP) layer to predict a future word's embedding (logit layer's) given the same hidden state of the conventional LM. This process is formulated as follows:

$$\hat{\mathbf{x}}^{e,l}_{t+n} = MLP^n(\mathbf{h}_t). \quad (4)$$

For instance, assuming N is 4, three MLP layers, MLP^1, MLP^2 and MLP^3, are employed to predict $\hat{\mathbf{x}}^{e,l}_{t+1}$, $\hat{\mathbf{x}}^{e,l}_{t+2}$, and $\hat{\mathbf{x}}^{e,l}_{t+3}$ respectively, as shown in Fig.1(b). Note that the conventional LM model predicts $\hat{\mathbf{x}}^{e,l}_{t}$, so a total 4-gram words are predicted. Then, we compute the likelihoods of the future target words, $p(\hat{x}_{t+1}|x_{<t};\theta)$, $p(\hat{x}_{t+2}|x_{<t};\theta)$, and $p(\hat{x}_{t+3}|x_{<t};\theta)$, following each logit layer and the softmax function. Instead of using separate logit layers for each future word prediction, we share the parameters across all logit layers, including the conventional LM model's original logit layer. Therefore, this approach increases just a small amount of parameters for each additional MLP layer. Finally, the individual word prediction loss and the total loss are formulated as follows:

$$\mathcal{L}_n = -\sum_{t=1}^{T-n} \log p(\hat{x}_{t+n} = x_{t+n}|x_{<t};\theta), \quad (5)$$

$$\mathcal{L}^{tot}_N = \frac{1}{2}\mathcal{L}_0 + \frac{\alpha}{2N-2}\sum_{n=1}^{N-1}\mathcal{L}_n, \quad (6)$$

where $\alpha < 1$ is a hyperparameter to control the additional loss term's influence. Note that $\mathcal{L}_0$ is the same as the conventional LM's loss, Eq.(3).

The framework of the simple N-gram prediction method is quite similar to the recent speculative decoding approaches proposed by (Gloeckle et al., 2024; Cai et al., 2024) which also add additional heads on top of the conventional LM's encoder to predict future words with shared encoder and logit layer like our method. As independently developed, however, there are several differences. First, we employ an MLP layer with ReLU activation expecting that the limited capacity of the MLP layer appropriately regularizes the main encoder, Enc_θ, to find simultaneously informative hidden states for all N-gram predictions. Second, since the next word typically has stronger dependencies with the preceding words than the other future words do, we multiply the original loss with only $1/2$ while we multiply the future words' losses with $\frac{1}{2N-2}$ and the scalar hyperparameter α in Eq.(6).

3.2 Word Difference Representation (WDR)

In this section, we explain the concept of WDR and how to provide diverse target representations with WDR to simple N-gram prediction LMs and diffusion model-based LMs (Li et al., 2022; Gao et al., 2022; Ye et al., 2023).

3.2.1 Definition of n-level WDR

As we mentioned, WDR is a function that transforms subsequent N words' embedding vectors into a contextualized form. WDR's transformation is based on a form of word embedding compositions: the difference vector, $\mathbf{x}^e_{t+1} - \mathbf{x}^e_t$ as its name implies. Since (Mikolov et al., 2013) demonstrated that arithmetic compositions of learned word embedding can convey semantic meanings, many researchers have explored the word embedding compositionality (Xu et al., 2015; Hartung et al., 2017; Poliak et al., 2017; Scheepers et al., 2018; Li et al., 2018; Frandsen and Ge, 2019). Their studies employed composed word embeddings as model inputs instead of original word embeddings. In con-

trast, we use the composed word embeddings as the target representation.

Given an embedding vector sequence $\{\mathbf{x}_1^e, \mathbf{x}_2^e, \cdots, \mathbf{x}_T^e\}$, the 1-level WDR at the time-step t is defined as follows:

$$\Delta_1 \mathbf{x}_t^e = \begin{cases} \mathbf{x}_{t+1}^e - \mathbf{x}_t^e & \text{if } 1 \leq t < T, \\ \mathbf{x}_T^e & \text{if } t = T. \end{cases} \quad (7)$$

In an inductive manner, the n-level WDR at the time-step t when $n > 1$ is defined as follows:

$$\Delta_n \mathbf{x}_t^e = \begin{cases} \Delta_{n-1}\mathbf{x}_{t+1}^e - \Delta_{n-1}\mathbf{x}_t^e & \text{if } 1 \leq t < T, \\ \Delta_{n-1}\mathbf{x}_T^e = \mathbf{x}_T^e & \text{if } t = T. \end{cases} \quad (8)$$

Based on the definitions of Eqs. (7) and (8), the n-level WDR can be represented by a linear combination of original word embeddings. For example, the 2- and 3-level WDRs at time-step t can be represented as follows: $\Delta_2 \mathbf{x}_t^e = \mathbf{x}_{t+2}^e - 2\mathbf{x}_{t+1}^e + \mathbf{x}_t^e$ and $\Delta_3 \mathbf{x}_t^e = \mathbf{x}_{t+3}^e - 3\mathbf{x}_{t+2}^e + 3\mathbf{x}_{t+1}^e - \mathbf{x}_t^e$, respectively. In this manner, we can derive the formulation of n-level WDR as follows:

$$\Delta_n \mathbf{x}_t^e = \sum_{i=0}^{n} \binom{n}{i} (-1)^i \mathbf{x}_{t+(n-i)}^e, \quad (9)$$

where $\binom{n}{i} = \frac{n!}{(n-i)!i!}$ is the binomial coefficient. This equation holds for every positive integer of n and for every time-step t when $t \leq T - n$. See Appendix A.1 for a proof of this equation. From this equation, n-level WDR can be easily reconstructed to the original word embedding. For example, the 1-level WDR, $\mathbf{x}_{t+1}^e$ can be reconstructed by adding $\mathbf{x}_t^e$ to $\Delta_1 \mathbf{x}_t^e$. Likewise, $\mathbf{x}_{t+n}^e$ can be reconstructed by adding $-\sum_{i=1}^{n} \binom{n}{i} (-1)^i \mathbf{x}_{t+(n-i)}^e$ to $\Delta_n \mathbf{x}_t^e$ (note that the first term of the right-hand side of Eq.(9) is $\mathbf{x}_{t+n}^e$). For simplicity, we use a new notation for the conjugate term that reconstructs the original embedding by addition to the n-level WDR as follows:

$$\Delta_n^r \mathbf{x}_t^e = -\sum_{i=1}^{n} \binom{n}{i} (-1)^i \mathbf{x}_{t+(n-i)}^e, \quad (10)$$

which leads to $\Delta_n \mathbf{x}_t^e + \Delta_n^r \mathbf{x}_t^e = \mathbf{x}_{t+n}^e$.

The subtracting operations of generating n-level WDRs, Eq.(8), and their reconstruction process (note the reconstruction conjugate term, Eq.(10), is the subtraction of the original embedding and n-level WDR) are parallelizable, so that they do not impose computational overhead in long sequence.

3.2.2 WDR on Simple N-gram LM

The application of WDR to simple N-gram LM (we call it 'WDR N-gram LM') follows the idea of adding additional MLP layers and predicting future words. However, in WDR N-gram LM, the MLP^n layer outputs $\Delta_n \hat{\mathbf{x}}_t^{e,l}$ instead of $\hat{\mathbf{x}}_{t+n}^{e,l}$. Then we produce its corresponding reconstruction conjugate term, $\Delta_n^r \mathbf{x}_t^{e,l}$, based on the logit layer's embedding matrix and target words. Adding those two, $\Delta_n \hat{\mathbf{x}}_t^{e,l} + \Delta_n^r \mathbf{x}_t^{e,l}$, yields $\hat{\mathbf{x}}_{t+n}^{e,l}$ as in the simple N-gram LM. Then, we take the processes of the logit, likelihood, and loss computations as in the simple N-gram LM.

An essential design of this framework is detachment of the produced reconstruction conjugate term, $\Delta_n^r \mathbf{x}_t^{e,l}$, during the backpropagation process. The absence of this detachment might lead the model to adjust the logit layer's weight matrix in a distorted manner, because the input of the logit layer will be recursively produced from itself.

Unlike the simple N-gram LM, the individual NLL loss, Eq.(5), decreases when the model predicts more accurate n-level WDR, $\Delta_n \hat{\mathbf{x}}_t^{e,l}$, because it will output more accurate future word's embedding after adding the reconstruction conjugate term $\Delta_n \hat{\mathbf{x}}_t^{e,l} + \Delta_n^r \mathbf{x}_t^{e,l} = \hat{\mathbf{x}}_{t+n}^{e,l}$. Since the reconstruction conjugate term is detached, the model would learn to predict $\Delta_n \mathbf{x}_t^{e,l}$, which is true n-level WDR. In other words, WDR N-gram LM's training framework can train LM with diverse and contextualized target representations given the same target word. The entire process of WDR 4-gram LM example is illustrated in Fig.1(c).

3.2.3 WDR on Diffusion-based LM

Since their great success in image generative models (Ho et al., 2020; Ramesh et al., 2022), diffusion models have been applied to text generation tasks (Li et al., 2022). However, compared to image diffusion models, text diffusion models' performances are less impressive, and some prior studies have investigated the reasons behind the limited progress of text diffusion models, particularly focusing on the diffusion and denoising processes on the word embedding space (Ye et al., 2023). A common argument from their analyses starts from the discreteness of the embedding vectors which form a finite number of clusters within the high-dimensional embedding space. Since the small noises of early diffusion steps are not significant enough to move an embedding vector from one cluster to another, the denoising process that trains the model becomes

trivial. We conjecture that providing diverse target representations can explicitly mitigate this problem because diversifying word embedding vectors will reduce the discreteness.

Given this perspective, we apply our WDR idea to the diffusion-based LM, such as DINOISER (Ye et al., 2023) and Differmer (Gao et al., 2022). As in the application of WDR to simple N-gram LM, we add extra heads to predict WDRs beside the original head that predicts the original embedding given t-times diffused embeddings. Similarly, we compute the reconstruction conjugate term from the shared logit layer with detachment, and then we reconstruct the original embedding. Finally, the diffusion loss for each head's prediction is formulated as follows (the original diffusion loss is when $n = 0$):

$$\mathcal{L}_n^d = \sum_{t=n+1}^{T} \| \hat{\mathbf{x}}_{t,0}^{e,l} - \mathbf{x}_{t,0}^{e,l} \|_2^2, \tag{11}$$

$$\hat{\mathbf{x}}_{t,0}^{e,l} = \begin{cases} f(\mathbf{x}_{t,k}^{e,l}; \theta) & \text{if } n = 0, \\ MLP^n(f(\mathbf{x}_{t,k}^{e,l}; \theta)) + \Delta_n^r \mathbf{x}_{t,0}^{e,l} & \text{if } n > 0, \end{cases} \tag{12}$$

where $f(\mathbf{x}_{t,k}^{e,l}; \theta)$ is the diffusion model's output given k-times diffused embedding vector, $\mathbf{x}_{t,k}^{e,l}$, with parameter set, θ. Analogously to the total loss of simple N-gram LM, we average the original and additional diffusion losses. See the previous works (Ye et al., 2023; Gao et al., 2022) for more details on the diffusion and denoising processes as well as the final loss. We followed their methodologies except for the changes to apply the WDR.

3.3 Ensemble Method to Leverage N-gram Predictions for Next Word Prediction

In this section, we propose a new ensemble method to incorporate the N-gram predictions into the process of the next word prediction. The encoded hidden state $\mathbf{h}_t$ represents the computed hidden state given the inputs up to time-steps $(t-1)$. During testing, in addition to the predicted embedding $\hat{\mathbf{x}}_t^{e,l}$ from the conventional LM, the MLP^n layer of simple N-gram LM can estimate the target word for time t given $\mathbf{h}_{t-n}$. In total, we can get N predicted embeddings for the current time-step. We ensemble these predicted embeddings just before the logit layer using the following formulation:

$$\hat{\mathbf{x}}_{t,ens}^{e,l} = (1-\lambda)\hat{\mathbf{x}}_t^{e,l} + \frac{\lambda}{N-1} \sum_{n=1}^{N-1} MLP^i(\mathbf{h}_{t-n}), \tag{13}$$

where λ is a scalar value between 0 and 1, which controls the influences of future word predictions (but derived from past time-steps) on the current word prediction. Similar to the rationale behind the dominance of the original NLL loss in its total loss formulation, Eq.(6), we do not equally average the original predicted embedding with others. In the case of WDR N-gram or diffusion LM, we ensemble $MLP^i(\mathbf{h}_{t-n}) + \Delta_n^r \mathbf{x}_{t-n}^{e,l} = \hat{\mathbf{x}}_t^{e,l}$ in the summation part in Eq.(13).

After this ensemble computation, we feed it to the logit layer and compute the next word's likelihood. During testing, this ensemble likelihood result is used to compute perplexity (PPL) in LM tasks or serving as candidate scores for beam search in NMT tasks. It is natural to expect that our ensemble method is beneficial if the additional heads predict independent (and effective) information from the original heads.

Table 1: Word-level PPL results of the preliminary experiment with Transformer decoder-based LMs on the PTB dataset. We tried several λ values in Eq.(13).

Model	Test PPL			
	λ=0.0	0.2	0.4	0.6
TF	161.0	-	-	-
TF+Sim N=2	150.8	134.6	135.3	156.3
N=3	153.3	134.4	133.0	151.9
N=4	158.1	133.6	129.1	147.1
TF+WDR N=2	149.0	136.5	129.8	**128.1**
N=3	153.1	136.1	128.2	128.8
N=4	150.5	131.6	124.1	127.5

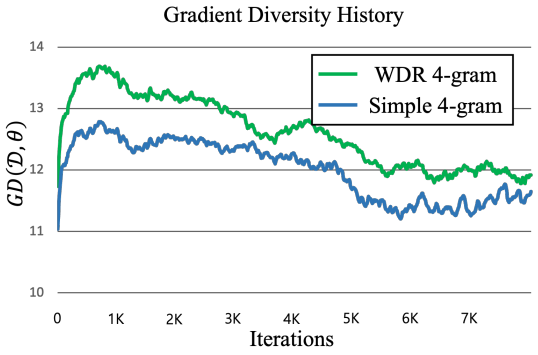


Figure 2: Gradient diversity comparison between simple 4-gram LM and WDR 4-gram LM.

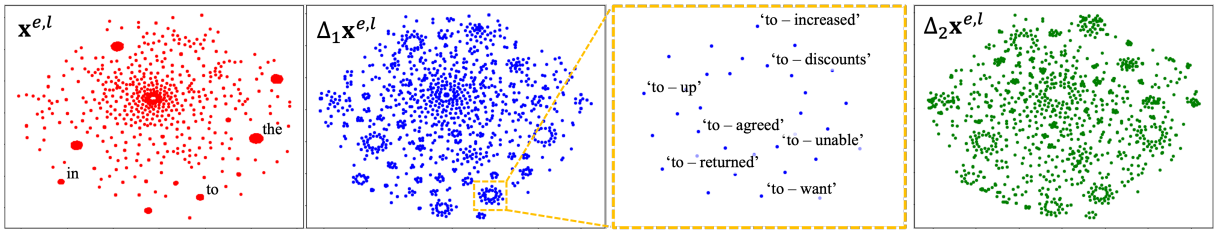


Figure 3: From the left-to-right, they are visualizations of the original embeddings (first), 1-level WDR and the plot zoomed in around the original word 'to' (second and third), and 2-level WDR (last), respectively. In the third plot, ('to-*word*') means the 1-level WDR vector, that is $\mathbf{x}^{e,l}_{to} - \mathbf{x}^{e,l}_{word}$ based on the word '*to*' fragment within the sentence.

4 Experiments and Results

We conducted preliminary analysis and main experiments. During the preliminary analysis, we trained the basic LM model on the Penn TreeBank (PTB) dataset (Marcus et al., 1993) to verify the expectations of our proposed methodologies. During the main experiments, we conducted LM and various conditional text generation tasks with multiple benchmark datasets and multiple baseline models, including an LLM, to demonstrate practical advantages of our proposed methodologies. Due to the page limit, we describe the details of the dataset, model architecture, and training scheme in Appendices A.2 and A.3.

4.1 Preliminary Analysis

As a preliminary analysis, we trained Transformer (TF) decoder-based LM models by applying simple N-gram and WDR N-gram ('TF+Sim' and 'TF+WDR') on the PTB dataset. During testing, we applied our ensemble method with varying the λ value. The total number of parameters of the TF baseline is 12M, and our proposed simple and WDR methods increase only 0.1M parameters for additional MLP layer. The details of the model architecture and training method for this experiment are described in Table 6 in Appendix A.3 as 'Small Dec. TF LM'. Also refer to Section A.2 for the details of PTB data preprocessing.

4.1.1 Perplexity of the Models

Table 1 presents the outcomes of the experiments. We trained the model of each configuration five times with different random seeds, and we report the average PPL scores. As demonstrated in previous N-gram prediction approaches (Sun et al., 2019; Joshi et al., 2020; Xiao et al., 2020; Qi et al., 2020; Gloeckle et al., 2024), both 'TF+Sim' and 'TF+WDR' outperform the conventional LM baseline. More interestingly, we found that 'TF+WDR' is generally better than 'TF+Sim' in various N settings. Our ensemble method consistently improves performance compared to the non-ensemble ones (where λ=0.0). More importantly, 'TF+WDR' models bring greater improvements with the ensemble method than 'TF+Sim'. For example, 'TF+WDR' models' improvements with the ensemble method is 24.07 on average, while 'TF+Sim' models' is 21.83. We interpret that WDR method trained the additional heads to predict more independent and effective information.

4.1.2 Comparison of Gradient Diversity

As mentioned in the introduction section, diverse target representations can bring a higher gradient diversity during training. To verify this expectation, during training, we measured '*gradient diversity (GD)*' (Yin et al., 2018) as follows:

$$\begin{aligned} GD(\mathcal{D},\theta) &= \frac{\sum_{i=1}^{|\mathcal{D}|} ||g_i||_2^2}{||\sum_{i=1}^{|\mathcal{D}|} g_i||_2^2}, \\ &= \frac{\sum_{i=1}^{|\mathcal{D}|} ||g_i||_2^2}{\sum_{i=1}^{|\mathcal{D}|} ||g_i||_2^2 + \sum_{i\neq j}\langle g_i, g_j\rangle}, \quad (14) \\ g_i &= \nabla_\theta \mathcal{L}_N^{tot}(X_i,\theta), \end{aligned}$$

where $\mathcal{D} = \{X_1, X_2, \cdots, X_{|\mathcal{D}|}\}$ is a mini-batch, $||\cdot||_2^2$ is the squared L^2 norm operation, $\langle\cdot,\cdot\rangle$ is the inner product operation, and ∇_θ is gradient operator with respect to θ. This metric is large when the inner product terms in the denominator are small, which means the gradients are different from each other.

Fig.2 demonstrates the GD history of 'TF+Sim N=4' and 'TF+WDR N=4' models during training. 'TF+WDR N=4' generally received higher GD than 'TF+Sim N=4'. As the stochastic property of stochastic gradient descent is known to enhance generalizability compared to full-batch gradient descent (Hardt et al., 2016; Yin et al., 2018), higher GD may offer similar advantages due to

higher stochasticity. Given this understanding, we believe that WDR-based training could be beneficial to improve generalizability.

4.1.3 Visualization of the Representations

To gain a more profound understanding of WDR's effect on target representations, we collected 1,270 actual target representations of the conventional LM model's training, which are the logit layer's embedding vectors corresponding to target words from the PTB testset. Also, we computed 1- and 2-level WDRs with the collected embeddings, and added them to the collection, resulting in 3,810 representations in total. Finally, we reduced the dimension of the total collection to 2-dimension with the t-SNE algorithm (Van der Maaten and Hinton, 2008).

Fig.3 shows the collected representations in a 2-dimensional space. The first plot illustrates the original embeddings, $\mathbf{x}^{e,l}$. Note that the representations of frequent words (e.g., 'in', 'to', and 'the') may be included more times than other words in the collection. We interpret that this is the reason why t-SNE places frequent words distant from other less frequent words to resemble the non-uniform distribution of the collection. On the other hand, the 1-level WDR representations, $\Delta_1\mathbf{x}^{e,l}$, look more diverse compared to the original embeddings as in the second plot. For example, by composing adjacent words such as 'want', 'unable', 'returned', into the frequent word 'to', it diversifies the embedding representations according to its previous word as in the third plot which is zoomed in. The 2-level WDR looks more diverse than the 1-level WDR as in the last plot. Based on this analysis, we believe that WDR N-gram LM actually receives diverse target representations.

4.2 Language Modeling Experiments

Our main LM experiments consist of training conventional LM models on multiple benchmark datasets and fine-turning the pre-trained LLM. We trained Tensorized Transformer (TT) (Ma et al., 2019) and Reformer (RF) (Kitaev et al., 2020) models on PTB, WikiText-2 (W2), Text8 (T8), and WikiText-103 (W103) datasets (Mikolov et al., 2014; Merity et al., 2016) for the conventional LM experiments. For the fine-tuning experiments, we fine-tuned GPTNEO 1.3B model (Black et al., 2021) on the W2 dataset. We note that our simple and WDR methods increased around 3.43% of total parameters out of the baselines' on average. For more details, refer to Appendices A.2 and A.3.

Table 2: Results of the conventional LM experiments.

Model	Test Word-level PPL			
	PTB	W2	T8	W103
TT (baseline)	55.0	56.1	121.4	20.1
TT+Sim	51.6	62.0	106.5	17.1
Ensemble	45.5	56.0	**89.5**	17.9
TT+WDR	47.5	57.7	91.7	**16.8**
Ensemble	**44.4**	**53.8**	90.2	16.9
RF (baseline)	28.0	31.6	64.3	50.3
RF+Sim	27.8	31.6	**62.1**	43.1
Ensemble	26.4	31.0	62.2	43.4
RF+WDR	26.0	31.5	62.2	**41.8**
Ensemble	**25.9**	**30.8**	**62.1**	41.9

Table 3: Results of LLM fine-tuning experiments.

Model	WikiText-2 PPL (w/ Ens.)
GPTNEO 1.3B	13.25
GPTNEO 1.3B + Sim	13.13 (12.93)
GPTNEO 1.3B + WDR	**13.00** (**12.91**)

Table 2 presents the entire results of the conventional LM experiments. The results show that, with the exception of TT-based models on W2, our proposed N-gram LM models consistently either match or surpass the baselines, even without the ensemble method. Remarkably, WDR N-gram LM models generally improve performance on top of the simple N-gram LM models. Upon applying our proposed ensemble method, they generally exhibit improvements over their non-ensemble counterparts, except the models trained on W103. Notably, the effect of the ensemble method is relatively significant in the smaller datasets (PTB and W2) rather than the larger datasets (T8 and W103).

Table 3 presents the results of our LLM fine-tuning experiments. Each configured model was trained three times with different random seeds. Notably, our WDR model outperformed the simple baseline. Specifically, the average PPL of WDR model, 13.00, surpasses the 95% confidence interval of the simple model's result, which is 13.02 (without ensemble). Furthermore, we observed that our ensemble method generally leads to improved performance. Based on these LM results, we argue that providing diverse target representations can offer advantages over the conventional reliance on fixed target representations alone.

Table 4: NMT results of conventional Transformer models on several benchmark datasets. The left and right numbers of '/' mean En-to-*(De or Tr)* and *(De or Tr)*-to-En translation results, respectively.

Model	BLEU Scores		
	IWSLT	WMT14	WMT18
TF	27.6/32.5	26.5/30.4	**11.9**/18.2
TF+Sim	28.0/33.0	26.2/30.9	11.6/18.2
Ensemble	**28.3**/33.4	26.3/31.0	11.6/18.3
TF+WDR	27.9/33.5	**26.7**/31.1	11.8/18.5
Ensemble	**28.3/34.0**	**26.7/31.2**	**11.9/18.8**

Table 5: Results of WDR applications on diffusion models: DINOISER and Differmer.

Baseline Arch.	Task	BLEU Scores	
		Baseline	+WDR
DINOISER	IWSLT14 En2De	25.76	26.26
	IWSLT14 De2En	31.26	31.83
Differmer	QQP	28.62	29.73
	WikiAuto	34.33	37.53

4.3 Conditional Text Generation Experiments

To further investigate the benefits of our proposed methodologies, we conducted multiple conditional text generation tasks, which can be regarded as conditional language modeling tasks. Our experiments are divided into two categories: (1) training conventional Transformer models on NMT datasets such as IWSLT14 En-De (Hwang and Jeong, 2023), WMT14 En-De (Vaswani et al., 2017), and WMT18 En-Tr (Bojar et al., 2018); and (2) training text diffusion models such as DINOISER and Differmer on IWSLT14 En-De, QQP (text paraphrasing), and WikiAuto (text simplification) (Gao et al., 2022). We note that both the simple and WDR methods increased the total number of parameters by approximately 2.45% over the baselines on average. For further details on models and datasets, see Appendices A.2 and A.3.

Table 4 presents the conventional Transformer-based NMT experiment results based on SacreBLEU (Post, 2018) evaluation metric. While simple N-gram method, 'TF+Sim', is sometimes worse than 'TF' baseline, WDR N-gram method, 'TF+WDR', always outperforms or is similar to the baseline. Notably, the integration of the ensemble method from either of 'TF+Sim' or 'TF+WDR' further increases performances. Specifically, we note that 'TF+WDR' with ensemble method improves performances by 0.7$\sim$1.5 BLEU scores compared to 'TF' baseline on both translation directions of 'IWSLT14 En-De', and German-to-English translations of 'WMT14 En-De' testsets.

The N-gram prediction approaches are more effective for De-En translation compared to En-De translation in 'IWSLT14 En-De' and 'WMT14 En-De' experiments. We believe that the difference in word variety between the two languages plays a key role. We analyzed the 'WMT14 En-De' training dataset (subword-level tokenized) and found that English has around 33.6K unique unigrams and 6.7M unique bigrams, while German has around 34.9K unique unigrams and 9.3M unique bigrams. This suggests that De-En translation might have simpler local dependencies to learn compared to En-De translation due to the lower number of unique bigrams. Considering simple local dependencies might lead to the over-fitting problem, we believe that this is a potential reason why N-gram prediction approaches, which can help mitigate over-fitting to local dependencies, are more effective for De-En translation.

Table 5 presents the experimental results of baseline diffusion models and our WDR models ('+WDR'), evaluated using BLEU scores. Note that we did not apply the simple N-gram method, since generic text diffusion models operate in a non-autoregressive decoding manner, predicting all words simultaneously. By applying our WDR method, we aim to leverage diverse target representations. Notably, our WDR models improved BLEU scores by 0.50, 0.57, 1.11, and 3.20 on the IWSLT14 En2De, IWSLT14 De2En, QQP, and WikiAuto datasets, respectively. Considering the results across all conditional text generation tasks, our WDR approach is also beneficial for a wide range of practical conditional LM tasks.

5 Limitations

During the hyperparameter search in the preliminary analysis (Sec. 4.1), we observed that when N is larger than 4, both simple and WDR N-gram LMs consistently performed worse. We hypothesize that predictions far into the future are too difficult to learn, and thus may have regularized the encoder in a disadvantageous way. Unfortunately, our WDR method could not overcome this limitation under the current experimental setting. In future work, we aim to develop a novel approach that leverages diverse target representations to address the issues associated with large N, or an alternative approach that does not rely on the N-gram

prediction framework.

6 Conclusion

In this work, we explored the potential of using contextualized target representations for training language models. We proposed WDR function, which transforms word fragments into contextualized forms and uses them as auxiliary target representations alongside the original targets. Building on our simple N-gram prediction framework, we applied WDR and validated its practical advantages across various models and datasets in both language modeling and conditional language modeling tasks. We found that applying WDR increases gradient diversity, which in turn improves generalization and overall performance.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443930, Development of Human-Level Robot Intelligence Capable of Real-Time Environmental Adaptation), and by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program(RS-2024-00431394) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abraham Frandsen and Rong Ge. 2019. Understanding composition of word embeddings via tensor decomposition. *arXiv preprint arXiv:1902.00613*.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. 2022. Empowering diffusion models on the embedding space for text generation. *arXiv preprint arXiv:2212.09412*.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*.

Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64.

DongNyeong Heo and Heeyoul Choi. 2023. Shared latent space by both languages in non-autoregressive neural machine translation. *arXiv preprint arXiv:2305.03511*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Soon-Jae Hwang and Chang-Sung Jeong. 2023. Integrating pre-trained language model into neural machine translation. *arXiv preprint arXiv:2310.19680*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In *Proceedings of the second workshop on subword/character level models*, pages 38–48.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018. Bag-of-words as target for neural machine translation. *arXiv preprint arXiv:1805.04871*.

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. 2019. A tensorized transformer for language modeling. *Advances in neural information processing systems*, 32.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Armand Joulin, Sumit Chopra, Michael Mathieu, and Marc'Aurelio Ranzato. 2014. Learning longer memory in recurrent neural networks. *arXiv preprint arXiv:1412.7753*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Adam Poliak, Pushpendre Rastogi, M Patrick Martin, and Benjamin Van Durme. 2017. Efficient, compositional, order-sensitive n-gram embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 503–508.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Thijs Scheepers, Evangelos Kanoulas, and Efstratios Gavves. 2018. Improving word embedding compositionality using lexicographic definitions. In *Proceedings of the 2018 World Wide Web Conference*, pages 1083–1093.

Chenze Shao, Yang Feng, and Xilin Chen. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. *arXiv preprint arXiv:1809.03132*.

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2020. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 198–205.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. *arXiv preprint arXiv:2010.12148*.

Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. 2015. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7:226–240.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. 2023. Dinoiser: Diffused conditional sequence learning by manipulating noises. *arXiv preprint arXiv:2302.10025*.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. 2018. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1998–2007. PMLR.

A Appendix

A.1 Proof of Eq.(9)

We provide a proof of Eq.(9) with the induction method. To avoid confusion, we temporarily change the notation of $\Delta_n \mathbf{x}_t^e$ in conjecture Eq.(9) to $\hat{\Delta}_n \mathbf{x}_t^e$ until it is proved. Based on the definitions of the 1 and n-level WDR, Eq.(7) and Eq.(8), we can verify the initial condition, that is $n = 1$, holds as follows:

$$\begin{aligned}\Delta_1 \mathbf{x}_t^e &= \mathbf{x}_{t+1}^e - \mathbf{x}_t^e \\ &= \binom{1}{0}(-1)^0 \mathbf{x}_{t+1}^e + \binom{1}{1}(-1)^1 \mathbf{x}_t^e \\ &= \sum_{i=0}^{1} \binom{1}{i}(-1)^i \mathbf{x}_{t+(1-i)}^e \\ &= \hat{\Delta}_1 \mathbf{x}_t^e.\end{aligned}$$

Therefore, the conjecture holds for the initial condition. Then, by following the induction method, we assume the conjecture at n-level is true, that is $\hat{\Delta}_n \mathbf{x}_t^e = \Delta_n \mathbf{x}_t^e$. Then, the $(n+1)$-level WDR from the definition Eq.(8) is derived to $\Delta_{n+1} \mathbf{x}_t^e = \Delta_n \mathbf{x}_{t+1}^e - \Delta_n \mathbf{x}_t^e = \hat{\Delta}_n \mathbf{x}_{t+1}^e - \hat{\Delta}_n \mathbf{x}_t^e$. Each term is derived as follows:

$$\begin{aligned}\hat{\Delta}_n \mathbf{x}_{t+1}^e &= \binom{n}{0}(-1)^0 \mathbf{x}_{t+n+1}^e + \binom{n}{1}(-1)^1 \mathbf{x}_{t+n}^e \\ &\quad \cdots \binom{n}{n-1}(-1)^{n-1} \mathbf{x}_{t+2}^e + \binom{n}{n}(-1)^n \mathbf{x}_{t+1}^e, \\ -\hat{\Delta}_n \mathbf{x}_t^e &= \binom{n}{0}(-1)^1 \mathbf{x}_{t+n}^e + \binom{n}{1}(-1)^2 \mathbf{x}_{t+n-1}^e \\ &\quad \cdots \binom{n}{n-1}(-1)^n \mathbf{x}_{t+1}^e + \binom{n}{n}(-1)^{n+1} \mathbf{x}_t^e, \\ \hat{\Delta}_n \mathbf{x}_{t+1}^e - \hat{\Delta}_n \mathbf{x}_t^e &= \binom{n}{0}(-1)^0 \mathbf{x}_{t+n+1}^e + \\ &\quad \left(\binom{n}{0} + \binom{n}{1}\right)(-1)^1 \mathbf{x}_{t+n}^e + \\ &\quad \cdots \left(\binom{n}{n-1} + \binom{n}{n}\right)(-1)^n \mathbf{x}_{t+1}^e + \\ &\quad \binom{n}{n}(-1)^{n+1} \mathbf{x}_t^e \\ &= \sum_{i=0}^{n+1} \binom{n+1}{i}(-1)^i \mathbf{x}_{t+(n+1-i)}^e \\ &= \hat{\Delta}_{n+1} \mathbf{x}_t^e.\end{aligned}$$

Note that the binomial coefficient, $\binom{n}{i}$, is the n-th row and i-th value of Pascal's triangle, and it satisfies $\binom{n}{i-1} + \binom{n}{i} = \binom{n+1}{i}$. Based on this outcome, the conjecture holds for $(n+1)$-level if the n-level is true. Therefore, the conjecture is proved.

A.2 Dataset Details

In this section, we describe the details of the datasets that we utilized in LM and conditional text generation experiments (Sections 4.2 and 4.3).

A.2.1 Language Modeling Dataset Description

For the training of conventional LM models, we utilized four datsets, such as PTB (-, 0.9M tokens, 10K vocabulary), WikiText-2 (W2, 2M tokens, 33K vocabulary), Text8 (T8, 15M tokens, 254K vocabulary), and WikiText-103 (W103, 103M tokens, 268K vocabulary) (Mikolov et al., 2014; Merity et al., 2016). We followed the open sources for data-related processes (e.g., download, tokenization, vocabulary, and train/valid/test sets splitting). Specifically, the W2 and T8 datasets were sourced from the GitHub repository[1], while the PTB and W103 datasets were sourced from the Tensorized Transformer (Ma et al., 2019)'s GitHub repository[2]. For the fine-tuning LLM task, we utilized W2 dataset with different data processes based on GPTNEO models' open source from Huggingface[3].

A.2.2 Conditional Text Generation Dataset Description

For the training of the conventional Transformer model for NMT, we utilized three datasets, such as 'IWSLT14 English-German'(En-De, 160K training pairs) (Hwang and Jeong, 2023), 'WMT14 English-German'(En-De, 3.9M training pairs) (Vaswani et al., 2017), and 'WMT18 English-Turkish' (En-Tr, 207K training pairs) (Bojar et al., 2018). We used the same preprocessing, tokenization, and subword byte-pair encoding methods with (Ott et al., 2019). We collected 10K, 10K, 32K most frequent subwords to organize vocabularies for datasets, respectively. For test sets, we used translations of TED and TEDx talks for IWSLT14 En-De. Also, we used Newstest18 and Newstest14 for WMT18 En-Tr and WMT14 En-De, respectively. For the IWSLT14 En-De dataset used for the training of DINOISER model, we downloaded the open preprocessed dataset of the Github[4] as the DINOISER baseline used. For the QQP (145K training pairs,

[1] https://github.com/chakki-works/chazutsu

[2] https://github.com/szhangtju/The-compression-of-Transformer

[3] https://huggingface.co/EleutherAI/gpt-neo-1.3B

[4] https://github.com/shawnkx/Fully-NAT#dataset

Table 6: Model and optimizer configurations of Transformer architectures used in the preliminary experiment of LM and NMT tasks. We used the same notation for model configurations as in (Vaswani et al., 2017), except for the number of layers (# of Layers) and multi-head attention's heads (# of Heads). 'ISRS' means the inverse square root learning rate scheduler (Ott et al., 2019) and '# of Tokens' indicates the total number of tokens in a mini-batch at each iteration.

Config.	Small Dec. TF LM	Small Enc-Dec TF NMT	Base Enc-Dec TF NMT
d_{model}	256	512	512
d_{ff}	2100	1024	2048
$d_k = d_v$	64	64	64
P_{drop}	0.3	0.3	0.1
ϵ_{ls}	0.1	0.1	0.1
# of Layers	6	6	6
# of Head	4	4	8
Optimizer	Adam	Adam	Adam
Learning Rate	0.00025	0.0005	0.001
Scheduler	None	ISRS	ISRS
# of Tokens	4K	4K	25K
Patience	50	50	50

text paraphrasing) and WikiAuto (678K training pairs, text simplification) datsaets that we used for the training Differmer models, we mainly followed the instructions of their official Github[5].

A.3 Model and Training Details

We explain the details of the model and training scheme that were used in our experiments (Sections 4.2 and 4.3) in this section.

A.3.1 Language Modeling Experiments

For the models and training of the conventional LM experiments, including 'TT', 'RF' baselines, and applications of our simple and WDR N-gram LM models, we followed the configurations reported in the previous works' papers (Ma et al., 2019; Kitaev et al., 2020) with several changes as described in Table 7. The total parameters of (TT, RF) baselines according to datasets are (6.7M, 15.3M) for PTB and W2, (82.4M, 236.6M) for T8 and W103, respectively. Both of our proposed simple and WDR methods increased the number of parameters by 0.1M for TT and 0.5M for RF per an additional MLP layer regardless of the type of dataset. The optimal hyperparameter settings of our proposed methods were found after the hyperparameter search. They are described in Table 8. Refer to the previous works for all of the details. The experiments of small datasets, PTB and W2, took around 3 hours on average based on a single GTX1080Ti GPU, while the experiments of large datasets, T8 and W103, took around 24 hours in average based on a single RTX3090 GPU. Unexpectedly, we found that the PPL of 'RF (baseline)' on W103 in Table 2 is unsatisfying compared to other results of PTB, W2, and T8 datasets. We trained 'RF' on W103 based on the same provided source code with the default configuration except for a few changes described in Table 7. Note that 'RF+Sim' and 'RF+WDR' models were trained under the same setting for fair comparisons.

For the fine-tuning of GPTNEO 1.3B pre-trained model, we downloaded the pre-trained models and their tokenizers with Huggingface API. We fully fine-tuned the pre-trained models on W2 dataset for 3 epochs with 8 batch size, AdamW optimizer (Loshchilov and Hutter, 2017), 1^{-5} initial learning rate, and linearly decreasing learning rate schedule (with 1K warmup). For our simple and WDR N-gram applications, we mainly followed the same training scheme of MEDUSA's fine-tuning (Cai et al., 2024). We shortly fine-tuned only for the additional heads before the main full fine-tuning because randomly initialized parameters of the additional heads on top of the pre-trained model can cause unstable training. Our simple and WDR applications increase the number of parameters by 8.39M per an additional MLP layer which are 0.65% out of the 1.3B total parameters. Similar to the conventional LM experiments, we conducted the hyperparameter search and we report the resulting optimal settings in Table 9. The fine-tuning experiments with baseline GPTNEO 1.3B pre-trained model took around 22.25 hours on average based on a single RTX3090 GPU. Our simple and WDR applications took 28.03 and 29.35 hours on average, respectively. We note that WDR method increased the total training times with 4.71% on top of the simple N-gram LM models.

A.3.2 Conditional Text Generation Experiments

We implemented the encoder-decoder Transformers (Vaswani et al., 2017) in different scales, small and base. We trained the small Transformer for the 'IWSLT14 En-De' and 'WMT18 En-Tr' datasets, and the base Transformer for the 'WMT14 En-De' dataset. Model and training configurations of these

[5]https://github.com/zhjgao/differmer

Table 7: Changed configurations from the original Tensorized Transformer and Reformer (Ma et al., 2019; Kitaev et al., 2020). We note that '# of Tokens' indicates the total number of tokens in a mini-batch at each iteration.

Dataset	Tensorized Transformer			Reformer	
	# of Tokens	# of Layers	Learning Rate	# of Tokens	Learning Rate
PTB	3,840	3	0.0025	16,384	0.0001
WikiText-2	3,840	3		8,192	
Text8	4,800	6		512	
WikiText-103	4,800	6		512	

Table 8: Configurations of our proposed N-gram approaches: N, α, and λ, used in the conventional LM and NMT experiments.

LM Task						NMT Task				
Model	Config.	Dataset				Model	Config.	Dataset		
		PTB	W2	T8	W103			IWSLT14	WMT14	WMT18
TT+Sim	$N/\alpha/\lambda$	2/1.0/0.2	4/1.0/0.2	3/1.0/0.2	2/1.0/0.1	TF+Sim	$N/\alpha/\lambda$	3/1.0/0.3	2/1.0/0.1	2/1.0/0.2
TT+WDR	$N/\alpha/\lambda$	2/1.0/0.4	4/1.0/0.3	3/1.0/0.1	2/1.0/0.1					
RF+Sim	$N/\alpha/\lambda$	4/1.0/0.2	2/1.0/0.2	3/1.0/0.1	4/1.0/0.1	TF+WDR	$N/\alpha/\lambda$	3/1.0/0.5	2/1.0/0.1	2/1.0/0.3
RF+WDR	$N/\alpha/\lambda$	4/1.0/0.1	2/1.0/0.3	3/1.0/0.1	4/1.0/0.1					

Table 9: Configurations of our proposed N-gram approaches: N, α, and λ, used in the LLM fine-tuning experiments.

Model	Config.	Value
GPTNEO 1.3B + Sim	$N/\alpha/\lambda$	2/0.10/0.08
GPTNEO 1.3B + WDR	$N/\alpha/\lambda$	4/0.10/0.04

Table 10: Configurations of our proposed N-gram approaches: N, α, and λ, used in the text diffusion model experiments.

Dataset	Config.	Value
IWSLT14 En2De	$N/\alpha/\lambda$	2/0.1/0.1
IWSLT14 De2En	$N/\alpha/\lambda$	3/0.1/0.1
QQP	$N/\alpha/\lambda$	3/0.5/0.2
WikiAuto	$N/\alpha/\lambda$	4/1.0/0.1

models are described in 'Small Enc-Dec TF NMT' and 'Base Enc-Dec TF NMT' columns of Table 6. The total number of parameters of small and base Transformer baselines are 32M and 77M, respectively. We applied our simple and WDR N-gram LM methods onto the decoder part. Each additional MLP layer in our simple and WDR methods required the number of parameters by around 0.5M. After hyperparameter search, we determined the optimal hyperparameters, N, α, and λ, and those are described in the 'NMT Task' column of Table 8. During training, we saved the best checkpoint based on the validation results. We early stopped the training whenever the model did not beat its previous best performance for the 'Patience' times on the validation (Heo and Choi, 2023). The experiments of small datasets, such as IWSLT14 En-De and WMT18 En-Tr, took 3 days in average based on 2 GTX1080Ti GPUs, while the experiments of large dataset, that is WMT14 En-De, took 3 days on average based on 4 RTX3090 GPUs. During testing, we applied beam search with 5 widths to output the final translation results.

For the experiments of text diffusion model, we heavily followed the model, diffusion process, and training configurations of the previous works (Ye et al., 2023; Gao et al., 2022). We refer to previous works for that information. Consequently, the total number of parameters of DINOISER(IWSLT14) / Difformer(QQP) / Difformer(WikiAuto) are 37M/109M/112M, respectively. As explained in Section 3.2.3, we added additional heads for WDR predictions on top of the denoising model's encoder (Transformer encoder). Notably, we did not apply the simple N-gram method, since diffusion-based text generative models usually follow non-autoregressive decoding so there is no need for future word prediction. Each added head increases by around 0.5M. We also conducted a hyperparameter search for WDR-related configurations, then we determined the optimal hyperparameters as demonstrated in Table 10. These experiments took 30 hours in average based on a single A4000 GPU.

How Is Context Important in Named Entity Recognition? A Comparison of Non-contextual and Contextual Word Embeddings

Dawid Smalcuga[1], Piotr Andruszkiewicz[1,2]

[1]Warsaw University of Technology, [2]IDEAS Research Institute
dawid.smalcuga.stud@pw.edu.pl, piotr.andruszkiewicz@pw.edu.pl

Abstract

The paper compares non-contextual and contextual word embeddings in Named Entity Recognition (NER) task. Word embeddings created by the GloVe, ELMo, BERT and RoBERTa models and the named entities predicted by the LUKE model have been tested. Models based on LSTM recursive neural network and convolutional neural network (CNN) have been created. They use vector representations and are being trained to recognize named entities in text. Using these models, the impact of word embeddings in the Named Entity Recognition task has been examined. The datasets used in the experiments are the Annotated Corpus for Named Entity Recognition and CoNLL-2003. We have investigated the importance of the size of context and which vector representation performs best in Named Entity Recognition. The experiments, we have conducted, prove that contextual word embeddings show their advantage if the context is longer than one sentence. Moreover, BERT and especially RoBERTa perform significantly worse than other models for entity types with small number of instances. Another finding is that cased BERT model achieves better results than its uncased counterpart.

1 Introduction

In this paper, Named Entity Recognition (NER), one of the tasks of Natural Language Processing (NLP), is examined. The NER tak is important in many applications of NLP, for instance, in analysis of user's utterances/commands passed to an intelligent agent. Such utterances are transformed with Automatic Speech Recognition (ASR) and then processed with NER to improve quality of tasks performed according to user's requests, e.g., Question Answering, making reservations, etc.

To investigate the impact of the size of the context on the NER task, we used two well-known benchmark datasets; namely, the Annotated Corpus for Named Entity Recognition (Walia) and CoNLL-2003 (Sang and Meulder, 2003). The main difference between these sets is the length of the context. The former dataset consists of single sentences, so that the context of each word is limited only to the sentence in which it occurs. The latter dataset consists of long documents where the meaning of each word depends on the broad context. The models that create the embedding vectors are: non-contextual GloVe (Pennington et al., 2014), contextual ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), which create word representations, and contextual LUKE (Yamada et al., 2020), which creates entity representations. These models differ in construction, in particular GloVe uses statistics. ELMo uses recursive neural networks, while the other three models, BERT, RoBERTa and LUKE, are based on transformers.

The aim of this work is to examine the impact of the choice of a model that creates representations of words or entities in the NER task regarding the available context. Thus, models creating non-contextual and contextual word embedding vectors are to be examined. We use the word embedding vectors created by GloVe, ELMo, BERT and RoBERTa (words) and LUKE (entities), and then train a model based on neural networks in the task of Name Entity Recognition. We focus on differences between non-contextual and contextual embeddings in regards to the available context in the data. Furthermore, we investigate the difference between cased and uncased models that produce word embedding vectors.

The source code and the data are available at the following website https://staff.elka.pw.edu.pl/~pandrusz/data/contexte/.

2 Related Work

Named Entity Recognition (NER) task has been recently studied in the context of Deep Neural Net-

Model	F1
LSTM-CRF (Lample et al., 2016)	91.0
ELMo (Peters et al., 2018)	92.2
BERT (Devlin et al., 2019)	92.8
Akbik, Blythe, Vollgraf (2018) (Akbik et al., 2018)	93.1
Baevski, Edunov, Liu, Zettlemoyer, Auli (2019) (Baevski et al., 2019)	93.5
RoBERTa (Liu et al., 2019)	92.4
LUKE (Yamada et al., 2020)	94.3

Table 1: Named Entity Recognition results from (Yamada et al., 2020).

works (Li et al., 2022). As different network architectures emerge, they are applied in NER task. Beneath the architectures, there are also word representations – embedding vectors – which are important in recognition of different entities types.

The first well-known representation using embedding vectors is Word2Vec (Mikolov et al., 2013). This solution creates similar vectors for semantically close words and uses two standard approaches: skip-gram and CBOW. After popularization of Word2Vec, more solutions appeared quickly, including very popular GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017).

In GloVe, learning is based on the probability of occurrence of a given word. To estimate this probability a table of mutual words occurrence is calculated. Then, we count how many times every word has appeared in the context of a given word. The context in the case is a "frame" with a width of 3 - the word preceding, a given word, and the following word.

The above models, produce the same vectors for a word no matter the context a word appears in. Thus, contextual vector embeddings were proposed. Contextual vectors are the extension of the described solutions. These models generate different vectors for the same word depending on the context of the word, i.e. words appearing next to the analyzed word.

ELMo (Peters et al., 2018) model uses recursive neural networks – LSTMs – to generate vectors. It process proceeding and following neighborhood by applying bidirectional LSTMs. A different type of contextual word embeddings models is based on transformers. The well known representatives of this type of solution are: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and LUKE (Yamada et al., 2020).

BERT drops recurrent mechanism and uses self attention. This model was later improved by, e.g., different pre-training procedure in RoBERTa and further improvements were incorporated in LUKE model.

LUKE treats words and entities in the text as independent tokens and generates their representations. It was tested on the CONLL-2003 dataset. This model achieved the highest F1 result at the time of its publication (please refer to Table 1). The model finds all possible ranges of entities in every sentence and then classifies them as one of the defined types of named entities or no entity. The representation of each entity is created by the alignment of the representation of the first and last word in the span and representation of the entity corresponding to the span. The maximum length of the entity span is 16 words, while the context is 512 tokens.

As there are many types of embeddings, a question arises – which type is the best in a given task. This aspect has been studied in the context of emotion detection (Polignano et al., 2019) and biomedical Natural Language Processing (Wang et al., 2018). We investigate the types of embedding vectors in Named Entity Recognition task for English.

3 Model for Word Embeddings Comparision

In this section, we present a description of the model we built for examining the impact of the method used for creating representations of words and entities in the task of NER. It uses the word embedding vectors created by the models described in previous sections, and then trains neural networks to recognize named entity. The number of entity types depends on the dataset we use and amounts to 8 for the Annotated Corpus for Named Entity Recognition called Kaggle (Walia) and 4 for CoNLL-2003 (Sang and Meulder, 2003).

Modern models, after appropriate training,

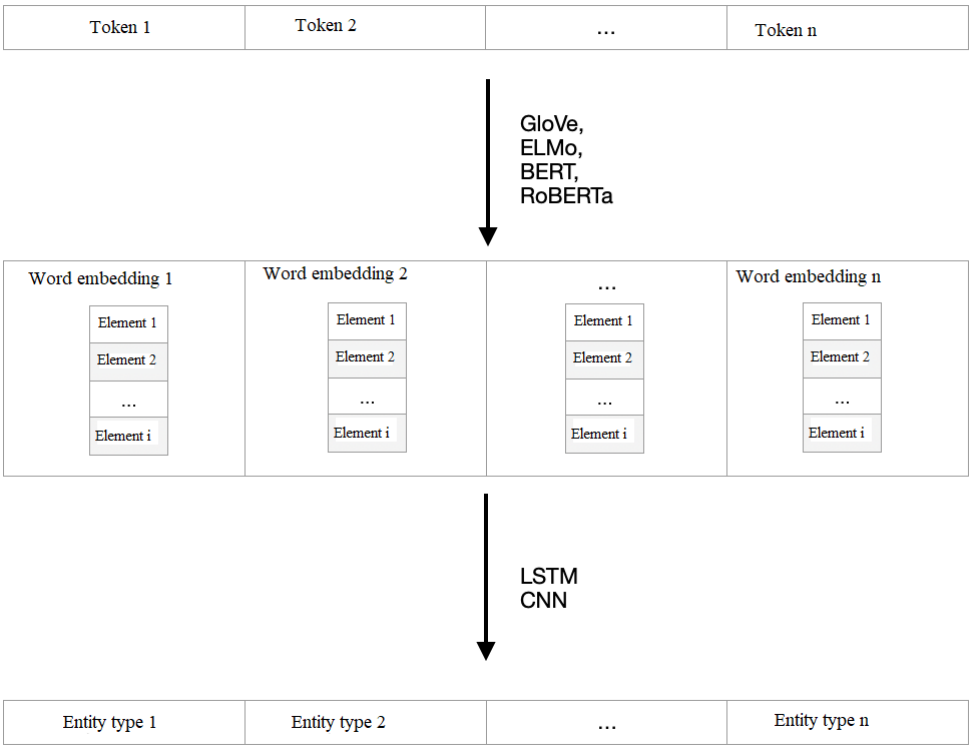


Figure 1: Model architecture using GloVe, ELMo, BERT or RoBERTa.

achieve very good results in the CoNLL-2003 ranking. For example, the LUKE achieves F1 score of 94.3, placing it at the top of the ranking. The idea in our research is to compare only the word embedding vectors, which are not adapted to NER task, not the whole models trained for NER task.

The architecture of the model we created is shown in Figure 1. LUKE, on the other hand, creates entity representations that are not available. For this reason, we use a version of the LUKE model that has been trained on the CoNLL-2003 dataset and its output, which is an entity type, instead of multidimensional vectors that are used to predict an entity type. The LUKE's output is processed in the same way as the model treats word embeddings – they are the input of the neural network.

3.1 Word Embedding Vectors

The first part of the model, we created, is a vector representation provided by one of the models:

GloVe trained on Wikipedia from 2014 and Gigaword 5 with the largest, 300-dimensional vector size.

ELMo and the following models generate contextual vectors. We used ELMo Medium, whose parameters are: total 28 million trainable parameters, 512-dimensional embeding vectors.

BERT is the most popular model in NLP. Thus, we decided to investigate the effect of two different versions of this model available in the Transformers library[1]: bert-base-cased and bert-base-uncased. The parameters of these models are as follows: 12 layers of transformers, 768 hidden units, a total of 110 million trainable parameters, 768-dimensional vectors. The *cased* model was trained on text consisting of lowercase and uppercase letters, while the text used to train the *uncased* model consisted only of lowercase letters. When not explicitly specified, BERT-cased has been used.

The RoBERT model is also from the Transformers library. The version we are using is roberta-base, which has the same number of parameters as bert-base.

The last model, LUKE, is also available in the Transformers library. It uses the *studio-ousia/luke-large-finetuned-conll-2003* version that was fine-tuned on the CoNLL-2003 dataset. It is not possible to use this model to classify the Kaggle dataset because the named entity types are already defined. There are no embedding vectors in the output, because the model is adapted to predict en-

[1]https://huggingface.co/docs/transformers/index

Word embeddings	Network type	Prec. macro	Recall macro	F1 macro	Prec. micro	Recall micro	F1 micro	Acc.
GloVe	LSTM	**68.32**	**60.79**	**62.83**	75.30	78.96	77.09	96.24
ELMo	LSTM	58.63	57.38	57.72	**78.26**	**80.67**	**79.45**	**96.60**
BERT	LSTM	56.49	52.15	52.88	75.42	77.15	76.27	96.00
RoBERTa	LSTM	47.00	41.99	42.95	63.73	65.93	64.81	94.25
GloVe	CNN	55.23	52.21	52.75	68.32	72.18	70.20	94.59
ELMo	CNN	56.84	53.03	53.90	73.90	76.63	75.24	95.74
BERT	CNN	46.70	46.02	45.48	71.03	74.11	72.54	95.28
RoBERTa	CNN	41.85	36.09	36.55	56.68	58.50	57.57	92.93

Table 2: Results for Kaggle set (the best results marked in bold).

Word embed.	Network type	art	eve	geo	gpe	nat	org	per	tim
GloVe	LSTM	**18.92**	38.71	82.55	**94.21**	**59.65**	56.86	72.86	78.85
ELMo	LSTM	2.44	25.32	**84.15**	92.54	37.21	**62.45**	**73.45**	**84.21**
BERT	LSTM	0.00	28.57	82.71	87.38	15.38	56.46	70.16	82.36
RoBERTa	LSTM	0.00	28.12	72.93	67.01	0.00	39.81	58.77	76.95
GloVe	CNN	0.00	**40.00**	78.90	93.20	34.38	44.69	60.37	70.50
ELMo	CNN	2.47	23.68	80.46	91.30	30.00	54.55	72.17	77.53
BERT	CNN	0.00	0.00	79.33	86.03	5.41	47.85	66.20	79.04
RoBERTa	CNN	0.00	11.54	66.12	65.21	0.00	30.57	48.29	70.66

Table 3: Results for entity types in Kaggle set.

tity types. The number at the model output, corresponding to the expected entity type, is treated as a 1-dimensional vector and is the input of the neural network. The idea behind this solution is to use all models in the same way, i.e., to train LSTM or CNN networks based on the created representations. Due to the lack of word embedding representation for the LUKE model, the result predicted by it was used.

3.2 Trainable Neural Network Part of Model

The word embedding vectors are prepared just before entering the neural network. They are then processed by recursive neural network LSTM or convolutional neural network (CNN). In the next step, the result goes through the dropout layer with a value of 0.3. Finally, the values are linearly transformed and the logarithm of the softmax function is returned.

Batch sizes were 4 for training and 2 for validation and test. The hidden layer size of LSTM and CNN was 100. The learning rate was 10^{-3}, while the number of iterations was 5. The Adam optimizer was used.

3.3 Differences Between the Proposed Solution and the Existing Models

It is worth emphasizing the differences between the proposed solution and the models trained to detect named entities, which are mentioned in Table 1. Figure 1 shows the architecture of the model, where we can see that the embedding vectors are used by the LSTM and CNN networks. The goal of such a solution is to use the same mechanism for all vector representations. It seems that such a procedure may significantly reduce the efficiency of Named Entity Recognition, because single layers of neural networks did not achieve good results in the NER task, which indicates the level of complexity of the models described in Table 1. Models presented in this table were finetuned on the CoNLL-2003 dataset. Nevertheless, the purpose of this work is to compare the effects of word embedding vectors only, not to achieve the best result.

4 Experiments

This section presents the results of experiments we conducted in our study. We present the results over two NER datasets: Annotated Corpus for Named Entity Recognition called Kaggle (Walia) and CoNLL-2003 (Sang and Meulder, 2003).

Word embeddings	Network type	Prec. macro	Recall macro	F1 macro	Prec. micro	Recall micro	F1 micro	Acc.
GloVe	LSTM	58.18	67.79	61.47	56.84	67.79	61.83	92.56
ELMo	LSTM	82.32	86.35	84.21	84.46	88.24	86.31	97.46
BERT	LSTM	80.00	79.63	79.61	81.15	83.25	82.19	96.16
RoBERTa	LSTM	67.72	70.78	69.12	70.71	74.58	72.59	94.51
LUKE	LSTM	**92.64**	**92.78**	**92.71**	**94.05**	**94.02**	**94.03**	**98.66**
GloVe	CNN	48.71	57.03	49.37	36.50	56.29	44.28	89.08
ELMo	CNN	76.08	80.56	78.05	75.54	82.08	78.68	96.51
BERT	CNN	71.80	72.92	71.90	70.08	75.43	72.66	94.46
RoBERTa	CNN	61.04	62.80	61.12	61.41	67.23	64.19	92.84
LUKE	CNN	85.98	88.71	87.26	88.83	92.10	90.44	98.33

Table 4: Results for CoNLL-2003 set.

4.1 Description of Tests

On the Kaggle dataset, the following combinations of the vector representation models: GloVe, ELMo, BERT, RoBERTa and the type of neural network: LSTM, CNN were tested. On the other hand, on the CoNLL-2003 dataset, in addition to the above-mentioned algorithms, LUKE was also tested. In each case, the precision, recall and F1 score were calculated, divided into micro and macro, and accuracy. In addition, the F1 results of each category were calculated in order to examine which labels the model copes with the worst.

We divided the dataset into three parts: 70% - training, 15% - validation, 15% - testing.

4.2 Evaluation method

The word embedding vectors created by GloVe, ELMo, BERT and RoBERTa represent words, not entities. For this reason, we used the IOB2 format.

Assuming that we count the detection of entire entities, we can evaluate the operation of the model, i.e., the entity must be accurately predicted to be successful. According to this scheme we calculate the precision and recall, and thus - the F1 measures. This is a method published with the CoNLL-2003 dataset. In all experiments in the paper, averages of 5 runs are presented. The best results are marked with bold.

4.3 Research Results on the Kaggle Dataset

The results of the experiments on the Kaggle dataset are presented in Tables 2 and 3.

The first observation during the analysis of the results for the Kaggle set is high accuracy of all cases. There are no big differences between its values in subsequent rows of the tables. Therefore, comparing models based on accuracy seems not to be a good idea. The F1 micro results oscillate between 64.81 and 79.45 for the LSTM neural network and 57.57 – 75.24 for the CNN, while the macro F1 results between 42.95 – 62.83 and 36.55 – 53.90. The difference between micro and macro is greater than 20%, which shows that there are large discrepancies in the number of elements of different entity types. Definitely better results are achieved by a model using a recursive type of neural network. This is not a surprise, because recursion is perfect for language learning, because a given word depends on those that occurred earlier.

In both cases – LSTM and CNN, the best micro results are achieved by the model using the embedding vectors created by ELMo. It is ranked first in the ranking of all the measures used, achieving F1 micro scores of 79.45 and 75.24, respectively.

However, in the case of macro average, the best F1 result is achieved by GloVe with LSTM equal to 62.83, and with CNN it is only 0.85 pp lower than the best result achieved by ELMo. It is worth noting the high value of GloVe macro precision with LSTM, which is equal to 68.32. The results of GloVe average macro show that the word embedding vectors have a greater impact on the ability to detect named entities of different types with more similar effectiveness, regardless of the number of elements in a given entity type. It may be surprising that the model using GloVe word embedding vectors seems to work so well. However, keeping in mind that the context of words in the Kaggle datasset is very limited, for this reason, context models do not have significant advantage over non-contextual models.

As for the BERT model, it ranks third for LSTM

Word emb.	Net.	loc	misc	org	per
GloVe	LSTM	67.62	52.48	55.07	70.72
ELMo	LSTM	89.05	70.05	81.89	95.86
BERT	LSTM	85.66	64.81	78.78	89.20
RoBERTa	LSTM	76.77	50.07	68.77	80.87
LUKE	LSTM	94.96	**84.60**	**93.50**	**97.77**
GloVe	CNN	74.96	62.18	32.49	27.85
ELMo	CNN	84.08	70.82	66.91	90.37
BERT	CNN	79.27	65.50	65.51	77.31
RoBERTa	CNN	71.29	45.08	58.03	70.08
LUKE	CNN	**95.12**	67.61	88.57	97.74

Table 5: Results for CoNLL-2003 set per entity type.

and second for CNN in micro results. It achieves the largest difference between the average micro and macro F1 results, which for the CNN case is as much as 27.06 pp. This shows how much it relies on a similar distribution of elements in entity types.

The worst results in all cases were achieved by the RoBERTa model. It is worth recalling that according to Table 1 the RoBERTa model, despite being an improved BERT model, performs worse in the task of NER compared to BERT. However, the difference is only 0.4 pp. Nevertheless, the results, we obtained, of the word embedding vectors created by RoBERTa are surprisingly low with a significant margin to BERT (around 9-13 pp.).

Table 3 shows the F1 results of each label for each case tested. For the entity type "art", the results of the most test cases are equal to 0. GloVe with LSTM achieved an F1 score significantly higher than other models, but still very low – 18.92%. The number of elements of the "art" category is only 59, which to some extent explains the difficulties in learning how to recognize this type of entity by the model. Two other categories with a small number of elements are "nat" and "eve", respectively 30 and 56. In these cases, the F1 scores of all models are definitely lower than for more numerous categories, leading to the same conclusion. In addition, it can be seen that the BERT and RoBERTa models achieve significantly lower results than other models in detecting "nat" entities.

In summary, the best embedding vector model for the Kaggle dataset is ELMo, because it achieves the best average micro scores. Nevertheless, the GloVe model is best at predicting different entity types, on average. The worst results were achieved by the RoBERTa model and, similarly to the BERT model, it performs significantly worse for entity types with small number of instances.

4.4 Research Results on the CoNLL-2003 Dataset

The first noticeable difference between the research results on the CoNLL-2003 dataset (Tables 4 and 5) and Kaggle is that the GloVe model achieves the lowest results in all categories. This is because the CoNLL-2003 dataset is divided into documents, so the context of each word is much broader than Kaggle's. Small context is the only advantage of non-contextual embedding vectors over contextual ones.

An additional model tested on the CoNLL-2003 dataset is LUKE. It is a model created for tasks dealing with entities, which achieved by far the best results in our study. The F1 result of the micro combination of LUKE with CNN is better than the second result for this type of network by as much as 11.76 pp. ELMo is in the second place and has a few percent advantage over the BERT model. The last but one place is occupied by the RoBERTa model, with much better results than the last GloVe. It is worth paying attention to the small discrepancy in the average micro and macro results. The difference between the extreme cases of F1 measure is 3.47 pp. for LSTM and 5.09 pp. for CNN. This is due to the similar number of elements in all categories, where the only category with a different number of elements is "MISC" (around twice smaller than other categories).

Table 5 presents F1 results for individual labels. The distribution for GloVe with LSTM is significantly different from that for GloVe with CNN, especially for "PER" type, where the difference is 42.87 pp. Other models achieve significantly better results for this type of entity, and the LUKE model

Word embeddings	Network type	Prec. macro	Recall macro	F1 macro	Prec. micro	Recall micro	F1 micro	Acc.
BERT-cased	LSTM	**56.49**	**52.15**	**52.88**	**75.42**	**77.15**	**76.27**	**96.00**
BERT-uncased	LSTM	50.89	47.64	48.38	73.02	73.56	73.29	95.07
BERT-cased	CNN	46.70	46.02	45.48	71.03	74.11	72.54	95.28
BERT-uncased	CNN	46.11	43.08	43.41	68.80	68.25	68.52	94.28

Table 6: Results of BERT-cased and BERT-uncased for Kaggle set.

Word embeddings	Network type	Prec. macro	Recall macro	F1 macro	Prec. micro	Recall micro	F1 micro	Acc.
BERT-cased	LSTM	**80.00**	**79.63**	**79.61**	**81.15**	**83.25**	**82.19**	**96.16**
BERT-uncased	LSTM	76.82	75.20	75.91	78.75	78.33	78.54	95.23
BERT-cased	CNN	71.80	72.92	71.90	70.08	75.43	72.66	94.46
BERT-uncased	CNN	70.26	71.51	71.12	70.86	71.51	71.18	94.01

Table 7: Results of BERT-cased and BERT-uncased for CoNLL-2003 set.

yields very high scores: 97.77 for LSTM and 97.74 for CNN. The worst scores can be observed for the "MISC" category. It is around twice as numerous as the others, and this is the reason for the worse results.

In conclusion, the LUKE model is by far the best choice of all the models tested. However, the system uses its ability to create entity representations. For word representations, the ELMo model achieves the best results, similarly to the Kaggle dataset. This model is followed by BERT, RoBERTa and GLoVe. The latter model yields significantly worse results as it is the only one model that creates non-contextual embeddings and the CoNLL-2003 dataset consists of long documents, thus long context is available.

Another finding is that RoBERTa underperforms compared to BERT. Despite being an optimized version of BERT, RoBERTa does not always outperform it in Named Entity Recognition (NER) tasks, please refer to Table 1. Several factors contribute to this. Firstly, there are differences in pretraining. RoBERTa removes the Next Sentence Prediction (NSP) task present in BERT. While studies have suggested that NSP is not critical for many NLP tasks, it might play a role in NER, where cross-sentence context can be important. Secondly, RoBERTa performs worse than BERT in NER tasks on both the CoNLL-2003 and Kaggle datasets. RoBERTa struggles particularly with less frequent entity types such as "nat" (nationalities) and "art" (artifacts). BERT, with NSP training, may be better at capturing global document-level

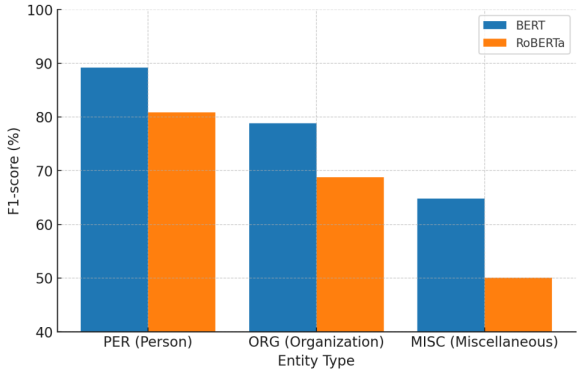


Figure 2: A bar chart comparing F1-scores for different entity categories with separate bars for BERT and RoBERTa on the CoNLL-2003 dataset.

context. Thirdly, RoBERTa employs an improved subword encoding mechanism – Byte-Pair Encoding at the byte level, which enhances generalization for rare words. However, in NER, this can lead to excessive fragmentation of named entities, making them harder to classify correctly.

An F1-score comparison, please refer to Figure 2, indicates that RoBERTa performs worse across all entity types on the CoNLL-2003 dataset.

RoBERTa's optimizations impact computational requirements in various ways. RoBERTa was trained on 160GB of raw text, whereas BERT used 13GB, requiring significantly more memory and compute power. Eliminating Next Sentence Prediction (NSP) should theoretically improve computational efficiency, but it does not necessarily enhance NER performance.

To conclude, while RoBERTa is designed as an

improved version of BERT, its modifications do not always lead to better performance in NER tasks. It is more computationally demanding yet does not always leverage its increased training data and longer context effectively for entity recognition.

4.5 Meaning of Capital Letters. Comparison of BERT-cased and BERT-uncased Results

At the first glance, the task of Named Entity Recognition seems to consist in detecting proper names. This is not true in all cases, because there are entities that have a defined type and are not proper names, for example, the date in the Kaggle dataset is of type "tim". Nevertheless, many named entities begin with capital letters, for example, the first name and surname of a person, or the name of an organization. Thus, we investigate whether the BERT-cased model, pre-trained on a dataset containing uppercase and lowercase letters, produces embedding vectors that outperform those produced by BERT-uncased, a model that has been pre-trained on an all-lowercase dataset.

The test results for BERT-cased and BERT-uncased are presented in Tables 6 and 7. The effects of these models on the Kaggle and CoNLL-2003 datasets were examined in the same way as the effects of the models in the previous sections.

The results of research on both datasets show that the BERT-cased model creates embedding vectors that work better in the task of detecting named entities. The only case where the BERT-uncased model performed better is CNN's BERT for the CoNLL-2003 dataset, where the micro precision is 0.78 pp. higher than that achieved by BERT-cased. For the Kaggle dataset, the largest difference in F1 micro scores is 4.02 pp. while the biggest difference in F1 macro results is 4.50 pp. The analogous values for the CoNLL-2003 dataset are 3.65 pp. and 2.7 pp., respectively. The differences in the results are significant, considering that the only difference is the size of the letters. The BERT-cased model produces embedding vectors that work better in the NER task.

5 Conclusions

In this paper, the influence of the choice of a model creating word embedding vectors on the results of the Named Entity Recognition, one of the widely used Natural Language Processing tasks, was investigated.

Five models creating word embeddings were used for the research: non-contextual GloVe, contextual ELMo, BERT, and RoBERTa, which create word representations, and contextual LUKE, which creates entity representations.

Based on the experiments, it can be concluded that in the case of Named Entity Recognition in single sentences (short context), the ELMo's embeddings perform the best. GloVe achieves slightly worse results. The context of the word is significantly less important when we are limited to just a single sentence, so using GloVe's word embedding vectors seems like a good idea. The BERT model also performs well, while the RoBERTa model ranks in the last position. However, in the NER task with the long context, the best performer is LUKE, which achieved much higher results than other models. Among the models creating representations of words, the best results were achieved by ELMo, and the worst by GloVe. This model creates non-contextual embedding vectors, which is why it does not perform well with long texts, i.e., long context. BERT again performs better than RoBERTa but worse than ELMo.

Another finding is that, BERT and especially RoBERTa perform significantly worse than other models for entity types with small number of instances for both short and long context. GloVe yields the best results in such a case for short context, especially with LSTM.

Finally, the effect of keeping/removing upper letters was investigated in the Named Entity Recognition task. Two types of BERT models were tested: BERT-cased, which was pre-trained on a dataset consisting of uppercase and lowercase letters, and BERT-uncased, which was pre-trained on a dataset consisting of only lowercase letters. As it turned out, such a small change in the dataset led to significantly better results of the BERT-cased model.

An interesting extension of this paper seems to be the study of the influence of the method of creating embedding vectors in other Natural Language Processing tasks, as well as the study of other models. Another possible continuation is to use the entity and words representations created by LUKE and examine their influence on the results of the experiments. For this purpose, it would be necessary to implement the LUKE model, because the embedding vectors created by LUKE are not available.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649. Association for Computational Linguistics.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5359–5368. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using bilstm, CNN and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 09-12, 2019*, pages 63–68. ACM.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

A. Walia. Annotated corpus for named entity recognition. *https://www.kaggle.com/datasets/abhinavwalia95/entity-annotated-corpus*.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul R. Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Informatics*, 87:12–20.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6442–6454. Association for Computational Linguistics.

IndicSQuAD : A Comprehensive Multilingual Question Answering Dataset for Indic Languages

Sharvi Endait[1,3], Ruturaj Ghatage[1,3], Aditya Kulkarni[1,3], Rajlaxmi Patil[1,3], Raviraj Joshi[2,3],
[1]Pune Institute of Computer Technology, [2]Indian Institute of Technology Madras, Chennai,
[3]L3Cube Labs, Pune,

Abstract

The rapid progress in question-answering (QA) systems has predominantly benefited high-resource languages, leaving Indic languages largely underrepresented despite their vast native speaker base. In this paper, we present IndicSQuAD, a comprehensive multi-lingual extractive QA dataset covering nine major Indic languages, systematically derived from the SQuAD dataset. Building on prior methods for constructing QA datasets in low-resource Indic languages, we adapt and extend translation techniques to ensure high linguistic fidelity and accurate answer-span alignment across diverse languages. IndicSQuAD comprises extensive training, validation, and test sets for each language, providing a robust foundation for model development. We evaluate baseline performances using language-specific monolingual BERT models and the multilingual MuRIL-BERT. The results indicate some challenges inherent in low-resource settings. Moreover, our experiments suggest potential directions for future work, including expanding to additional languages, developing domain-specific datasets, and incorporating multimodal data.

1 Introduction

Question Answering (QA) has been a cornerstone of natural language understanding (NLU), with datasets like SQuAD (Rajpurkar et al., 2016) driving advancements in machine learning models for extractive QA. While these datasets have enabled significant progress, most large-scale QA datasets are centered around English and a few high-resource languages, leaving many Indic languages underrepresented. Recent benchmarks highlight this disparity, with evaluations showing that multilingual Large Language Models (LLMs) perform substantially worse on low-resource languages compared to English. As a result, developing robust QA systems for Indic languages remains a challenge, despite the large native speaker population and the growing need for AI applications across diverse linguistic landscapes.

Indic languages, including Hindi, Bengali, Tamil, Telugu, Marathi, and others, are spoken by over a billion people. However, unlike English or Chinese, Indic languages lack extensive QA datasets, limiting the performance and adaptability of multilingual models in real-world applications. While initiatives such as TyDiQA (Clark et al., 2020) and XQuAD (Artetxe et al., 2019) have attempted to introduce non-English QA datasets, these remain either limited in size or do not comprehensively cover multiple Indic languages. The Indic-QA Benchmark (Singh et al., 2025) introduced in 2024 represents a significant advancement in this space, covering 11 major Indian languages from two language families and incorporating both extractive and abstractive question-answering tasks. This benchmark combines existing datasets with English QA datasets translated into Indian languages, demonstrating the growing recognition of the need for comprehensive multilingual QA resources. Nevertheless, the performance of multilingual models on this benchmark remains subpar, particularly for low-resource languages, underscoring the persistent challenges in this domain. A more targeted effort in this direction is the IndicQuest benchmark (Rohera et al., 2024), which is specifically designed to evaluate the factual accuracy of Indic LLMs.

The absence of large-scale, high-quality annotated QA datasets for Indic languages restricts their integration into information retrieval, education, healthcare services, governance applications, and AI-powered customer support systems. This limitation perpetuates digital inequality, where speakers of high-resource languages benefit more from technological advancements than those speaking low-resource languages.

To bridge this gap, we introduce IndicSQuAD, a comprehensive multi-lingual QA dataset covering

9 Indic languages, built by systematically translating and adapting the English SQuAD dataset while ensuring linguistic accuracy and answer-span alignment. Following earlier work on developing Marathi QA datasets, this study extends the methodology to nine additional Indic languages. Drawing from recent advances in alignment techniques and span retrieval methods, our approach addresses the challenges identified in previous translation efforts (e.g., morphological variations, syntactic differences, and maintaining contextual integrity). IndicSQuAD represents the largest multi-Indic QA resource to date, presenting the dataset along with additional baseline models, designed to facilitate research in low-resource language modeling and improve access to knowledge for Indic language speakers. The key contributions are as follows:

1. **Creation of IndicSQuAD dataset**[1] – A large-scale multi-lingual extractive QA dataset for 10 Indic languages, derived from SQuAD, ensuring high linguistic fidelity and comprehensive coverage across language families. The languages supported are Marathi, Hindi, Bengali, Telugu, Tamil, Gujarati, Punjabi, Kannada, Oriya, and Malayalam.

2. **Comprehensive Baseline Models and Benchmarking** – Strong baseline performances are established using fine-tuned monolingual BERT models for each Indic language, ensuring a tailored evaluation that captures language-specific nuances. Additionally, a comparative analysis with multilingual models like MuRIL Bert (Khanuja et al., 2021) is provided, assessing their effectiveness across diverse Indic languages. This evaluation framework addresses the unique challenges of each language, enabling meaningful comparisons across language families and resource availability.

2 Related Work

The development of question-answering systems for Indian languages has gained significant attention in recent years, though these languages remain resource-scarce compared to English. Several datasets have been created to address this gap using translation and native language approaches. This section provides a comprehensive overview of the existing work in this domain.

2.1 Translation Based Approaches

MahaSQuAD (Ruturaj et al., 2023) (Joshi, 2022b) represents the first comprehensive question-answering dataset specifically developed for Marathi. This work filled a critical gap in the language resources landscape. The paper details that MahaSQuAD consists of 118,516 training, 11,873 validation, and 11,803 test samples, accompanied by a gold test set of 500 manually verified examples. The work also presents a generic approach for translating SQuAD into any low-resource language, addressing the significant challenge of mapping answer translations to their spans in translated passages. In the current work, this approach is extended to nine more Indian languages.

(Kumar et al., 2022a) developed an extensive resource with 28,000 samples each for Hindi and Marathi by translating SQuAD 2.0, helping address data scarcity for these languages.

XQuAD (Artetxe et al., 2019) consists of 240 paragraphs and 1,190 question-answer pairs derived from SQuAD v1.1 and professionally translated into ten languages, including Hindi. This dataset has served as an important benchmark for cross-lingual question answering evaluation.

MLQA (Lewis et al., 2019) serves as another important benchmark with 4,918 context-question-answer triples available in Hindi. It enables the evaluation of cross-lingual generalization capabilities in multiple languages simultaneously.

2.2 Natively annotated datasets

The ChaII Dataset (Thirumala and Ferracane, 2022) features context-question-answer triples in Hindi and Tamil gathered directly without translation. Created by native speaker annotators, this dataset presents realistic information-seeking tasks focused on Wikipedia articles. The dataset includes 1,104 questions with the Hindi portion translated into ten other Indian languages.

Recent work by (Thirumala and Ferracane, 2022) has investigated the application of transformer models pre-trained on multiple languages, specifically focusing on Hindi and Tamil question-answering, demonstrating enhanced performance in extractive QA tasks.

Additionally, the Extended Chaii dataset has been developed containing Tamil translations from the SQuAD dataset, designed specifically for

[1] https://github.com/l3cube-pune/indic-nlp/tree/main/L3Cube-IndicSQUAD

question-answering tasks in low-resource Indic languages. The dataset consists of 2,855 training instances, 460 validation instances, and 250 test instances, making it a valuable resource for Tamil language processing.

The MMQA dataset (Gupta et al., 2018) contains 5,495 question-answer pairs in English and Hindi, covering factoid and short descriptive questions across multiple domains. This dataset is specifically designed to evaluate both bilingual and cross-lingual question answering that processes queries in either Hindi or English and retrieves answers in either language from documents in Hindi or English. The MMQA framework represents an important contribution toward multilingual information access, particularly beneficial in the Indian context.

2.3 Natively Annotated Datasets

While translation-based approaches have been instrumental in creating resources for low-resource languages, natively constructed datasets offer unique advantages in preserving linguistic authenticity.

BanglaQuAD (Rony et al., 2024) represents a significant contribution to Bengali language processing, containing 30,808 question-answer pairs constructed directly from Bengali Wikipedia articles by native speakers. Unlike translation-based approaches, this methodology avoids potential pitfalls associated with translated datasets, including loss of linguistic authenticity and contextual accuracy. The authors provide a detailed analysis of question types and answer distributions, along with baseline performance metrics using both monolingual and multilingual models.

The INDIC QA Benchmark (Singh et al., 2025) represents one of the most recent and comprehensive efforts in this domain, covering 11 major Indian languages and addressing both extractive and abstractive QA tasks. This benchmark aims to standardize evaluation across multiple Indic languages, enabling more direct comparisons of model performance across linguistic boundaries. The benchmark incorporates various question types and difficulty levels, providing a nuanced understanding of model capabilities across different linguistic structures found in Indic languages.

L3Cube-IndicQuest (Rohera et al., 2024) takes a more comprehensive approach by covering 19 Indic languages, making it one of the most linguistically diverse QA datasets available. Unlike many existing datasets that focus primarily on question-answering capabilities in general domains, L3Cube-IndicQuest specifically addresses five domains of particular relevance to the Indian context: Literature, History, Geography, Politics, and Economics. Each language subset contains carefully curated question-answer pairs designed to evaluate a model's ability to represent and process knowledge specific to Indian cultural and regional contexts.

The ChAII Dataset (Singh et al., 2025) features context-question-answer triples in Hindi and Tamil gathered directly without translation. Created by native speaker annotators, this dataset presents realistic information-seeking tasks focused on Wikipedia articles. The dataset includes 1,104 questions with the Hindi portion translated into ten other Indian languages.

3 Experimental Setup

3.1 Data Collection

The data collection process utilized the Stanford Question Answering Dataset (SQuAD 2.0), originally in English, which comprises over 150,000 question-answer pairs. Notably, about 34% of these questions are unanswerable, challenging models to handle ambiguity and non-definitive answers effectively. Each row in SQuAD 2.0 includes essential components such as title, context, question, answer start index, and answer text.

To create datasets, the robust translation and transliteration procedure involved developing a sophisticated algorithm to address inconsistencies that arise during translation, particularly in locating the correct answer index post-translation. By doing so, it was ensured that the translated datasets accurately reflect the nuances of the original English dataset while adapting to the linguistic characteristics of the target language. This approach not only enhances the quality of the translated datasets but also facilitates the development of more accurate question-answering models for low-resource languages.

3.1.1 Translation Methodologies

When creating multilingual QA datasets through translation, several methodologies can be employed, each with its advantages and challenges. Beyond the approach based on MahaSQuAD (Ruturaj et al., 2023), other researchers have explored various techniques for cross-lingual transfer in QA contexts.

Language	Model	EM%	F1%	EM (Has_ans)	F1 (Has_ans)	EM (No_ans)	F1 (No_ans)	BLEU% (Unigram)	BLEU% (Bigram)
Hindi	HindiRoBERTa	56.20	59.67	50.79	57.8	61.50	61.50	61.8	53.5
	MurilBERT	53.21	56.89	53.05	60.49	53.37	53.37	62.8	55.0
Punjabi	PunjabiBERT	51.04	54.53	47.59	54.63	54.42	54.42	54.4	46.1
	MurilBERT	50.80	54.41	47.08	54.38	54.40	54.40	54.5	46.3
Gujarati	GujaratiBERT	49.00	52.91	47.36	55.26	50.61	50.61	52.8	44.2
	MurilBERT	48.09	52.24	47.63	56.00	48.54	48.54	54.7	47.0
Kannada	KannadaBERT	50.97	54.90	48.54	56.49	53.34	53.34	52.3	45.9
	MurilBERT	49.64	53.81	48.27	56.68	50.99	50.99	53.7	44.6
Tamil	TamilBERT	50.97	54.44	46.38	53.39	55.47	55.47	53.03	44.26
	MurilBERT	49.70	53.14	46.40	53.34	52.94	52.94	53.85	44.83
Bengali	BengaliBERT	50.07	54.27	46.93	55.42	53.14	53.14	57.7	49.5
	MurilBERT	49.36	53.70	46.98	55.75	51.70	51.70	56.6	48.2
Telugu	TeluguBERT	52.17	55.34	44.98	51.37	59.24	59.24	54.8	47.5
	MurilBERT	51.11	54.38	44.30	50.90	57.82	57.82	52.9	45.1
Oriya	OdiaBERT	54.33	57.60	44.61	51.22	63.82	63.82	56.8	48.6
	MurilBERT	48.65	52.47	43.75	51.48	53.44	53.44	49.7	41.7
Malayalam	MalayalamBERT	51.02	49.42	42.24	49.26	59.59	59.59	52.0	43.2
	MurilBERT	45.67	49.62	41.89	49.89	49.36	49.36	46.9	38.6
Marathi	MahaBERT	51.28	54.88	51.04	58.31	51.52	51.52	57.9	49.9
	MurilBERT	50.13	53.91	51.26	58.92	49.03	49.03	57.7	49.4

Table 1: Performance of various models on different languages.

(Kumar et al., 2022b) proposed Multilingual Contrastive Training (MuCoT), a three-stage pipeline for question-answering in low-resource languages. This approach utilizes translation and transliteration with contrastive training across language families, showing particular effectiveness when data from the same language family is grouped. Their experiments demonstrated that translations from Indo-Aryan languages (Bengali and Marathi) significantly improved performance on Hindi, while Dravidian language data (Telugu and Malayalam) enhanced Tamil performance.

More recently, Self-Translate-Train (Ri et al., 2024) has emerged as a promising approach that leverages large language models to generate translations without requiring external translation systems. This method generates synthetic training data in the target language by utilizing the model's own translation capabilities, demonstrating substantial performance gains across several non-English languages without intensive additional data collection.

In creating IndicSQuAD, these approaches were built upon while addressing the specific challenges of Indic languages, such as maintaining context and handling linguistic nuances during translation. The methodology focused particularly on ensuring accurate mapping of answer spans in translated passages, a significant challenge when dealing with languages that differ substantially in word order and sentence structure from English.

Language	**Family**	**Script**
Marathi	Indo-Aryan	Devanagari
Hindi	Indo-Aryan	Devanagari
Punjabi	Indo-Aryan	Gurmukhi
Bengali	Indo-Aryan	Bengali
Gujarati	Indo-Aryan	Gujarati
Oriya	Indo-Aryan	Oriya
Tamil	Dravidian	Tamil
Telugu	Dravidian	Telugu
Kannada	Dravidian	Kannada
Malayalam	Dravidian	Malayalam

Table 2: Languages, their Families, and Scripts

3.2 Languages covered

IndicSQuAD includes question-answering datasets for 9 Indic languages, covering a diverse set of Indo-Aryan and Dravidian languages. These languages vary significantly in terms of script, morphology, and linguistic resources, making the dataset a valuable resource for multilingual and low-resource NLP research.

India's linguistic diversity is immense, with the 2011 Census identifying 122 major languages and 1,599 other languages. Among these, the most widely spoken languages are Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Punjabi, Kannada, Odia, and Malayalam. Despite their extensive use, many of these languages are considered low-resource in the field of Natural Language Processing (NLP) due to the limited availability of annotated datasets and linguistic tools. This scarcity poses significant challenges in developing robust NLP applications, as models trained on high-resource languages often fail to generalize effectively to low-resource contexts. By creating comprehensive question-answering datasets for these languages, IndicSQuAD aims to bridge this gap, facilitating the development of more inclusive and effective NLP applications.

3.3 Robust approach

The creation of IndicSQuAD employed a robust translation strategy to preserve linguistic accuracy and contextual integrity in low-resource languages. To address the challenge of aligning translated answers with their corresponding spans in translated passages, the English context was first segmented into sentences. Each sentence and its associated answer were then translated into the target language. Using similarity analysis tools, the most contextually appropriate span in the translated passage was identified to match the translated answer. This approach ensured precise alignment, overcoming the common mismatch between independently translated answers and contexts.

Figure 1 illustrates the methodology employed to accurately map translated answers to their corresponding spans within translated passages. The robust algorithm developed for MahaSQuAD ensures precise alignment of translated answers within their contexts through the following steps:

1. **Sentence Segmentation**: The English context is divided into individual sentences using the NLTK library.

2. **Answer Sentence Identification**: The English sentence containing the answer is identified from the individual sentences.

3. **Translation**: Both the identified sentence and the answer are translated into Marathi (target language) using Google Translate.

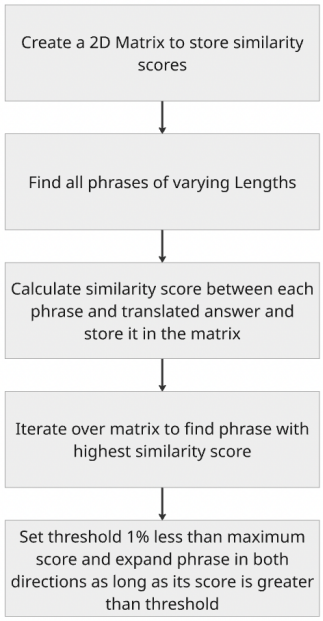


Figure 1: Algorithm for obtaining the answer and the answer span from the context

4. **Similarity Analysis**: Within the translated Marathi sentence, all possible substrings are compared to the translated answer using the SimilarityAnalyzer from the MahaNLP library (Joshi, 2022c; Magdum et al., 2023). The library uses embedding models released in (Deode et al., 2023). A similarity score matrix is generated to identify the substring with the highest similarity to the translated answer.

5. **Answer Span Determination**: The substring with the maximum similarity is selected as the base answer. Adjacent words are appended to this base answer, and the similarity is recalculated. If the new similarity score remains within 1% of the maximum, the extended phrase is accepted. This iterative process ensures that the translated answer accurately reflects the original meaning and context.

6. **Transliteration**: To maintain script consistency, named entities and numerical values are transliterated into the Devanagari script using the AI4Bharat Transliteration Engine.

To further enhance consistency, named entities were transliterated into Devanagari script using the AI4Bharat Transliteration Engine, and numerical values were converted into their counterparts. This meticulous process minimized errors and ensured uniformity across the dataset. The resulting dataset,

comprising 118,516 training samples, 11,873 validation samples, and 11,803 test samples, provides a scalable framework for translating SQuAD into other low-resource languages while maintaining linguistic and cultural nuances.

3.4 Dataset Statistics

Each of the ten languages in IndicSQuAD, including Marathi, Hindi, Bengali, Telugu, Tamil, Gujarati, Punjabi, Kannada, Oriya, and Malayalam, consists of **118,516 entities** in the training set, **11,873 entities** in the validation set, and **11,803 entities** in the test set. This large-scale dataset provides a robust foundation for training and evaluating QA models, addressing the scarcity of high-quality annotated resources for Indic low-resource languages.

Training set	118,516 samples
Validation set	11,873 samples
Test set	11,803 samples

Table 3: Dataset Statistics

4 Benchmarking and Experiments

4.1 Models used

To evaluate IndicSQuAD, monolingual and multilingual models were employed to establish baseline performances across the ten Indic languages.

- **Monolingual Models**
  Monolingual models are language-specific models trained solely on a particular language. These models are optimized for the linguistic characteristics of their respective languages, often leading to improved performance compared to multilingual counterparts. For each language in IndicSQuAD, BERT-based monolingual models for low-resource languages such as HindiBERT (Joshi, 2022a), PunjabiBERT, GujaratiBERT, KannadaBERT, TamilBERT, BengaliBERT, OdiaBERT, and MalayalamBERT were utilized, fine-tuned on the corresponding datasets.

- **Multilingual Models**
  In this research, MuRIL-BERT (Multilingual Representations for Indian Languages) was utilized, a transformer-based language model pre-trained on 17 Indian languages, including Marathi. MuRIL-BERT has demonstrated robust performance in understanding and processing Indian languages, making it a suitable choice for this study. This model's architecture allows it to effectively capture linguistic nuances across multiple Indian languages, facilitating the development of more accurate and efficient natural language processing applications.

4.2 Experimental Setup

Fine-tuning was conducted on the models using a custom dataset spanning three epochs and utilizing A100 GPUs with a consistent batch size of 32. The carefully selected hyperparameters include n_best_size (which refers to the number of predictions provided per question) set to 2, which significantly shaped the training dynamics and influenced the experimental outcomes. The other key hyperparameters employed during fine-tuning included a learning rate of 1e-4 and the AdamW optimizer. These adjustments were crucial in refining the model and enhancing its performance.

4.3 Results

The evaluation of the IndicSQuAD dataset from Table 1 highlights the superior performance of monolingual models over the multilingual MurilBERT across most Indic languages. For example, HindiRoBERTa outperformed MurilBERT for Hindi, achieving higher EM and F1 scores (56.20% and 59.67%, respectively, compared to 53.21% and 56.89%). Similarly, language-specific models like BengaliBERT and TamilBERT demonstrated better results in their respective languages, with BengaliBERT achieving an EM score of 50.07% compared to MurilBERT's 49.36%. These monolingual models consistently showed better contextual understanding and exact match accuracy.

In contrast, MurilBERT exhibited more generalized performance but lagged in capturing language-specific nuances, especially in low-resource languages like Telugu and Malayalam. For instance, TeluguBERT achieved an EM score of 52.17%, outperforming MurilBERT's 51.11% by leveraging its tailored design for Telugu. This trend underscores the importance of monolingual models in improving language-specific performance and highlights the need for further optimization of multilingual models to close the gap in low-resource language processing.

5 Conclusion and Future work

IndicSQuAD represents a significant advancement in addressing the scarcity of high-quality, large-scale question answering datasets for Indic languages. By systematically translating and adapting the widely-used SQuAD dataset into nine major Indic languages, this work not only bridges the resource gap but also establishes robust baselines using both language-specific and multilingual models. The comprehensive evaluation framework highlights the superior performance of monolingual models in capturing linguistic nuances, while also underscoring the challenges faced by multilingual models in low-resource settings. The dataset, along with the accompanying models and evaluation tools, will be made publicly available upon publication, fostering further research and development in multilingual NLP. Moving forward, expanding IndicSQuAD to additional languages, creating domain-specific datasets, and integrating multimodal data will further enhance the accessibility and effectiveness of AI-powered applications for Indic language speakers. This initiative is a crucial step toward reducing digital inequality and ensuring that speakers of low-resource languages can fully benefit from advances in natural language understanding and information retrieval.

While IndicSQuAD provides a strong foundation for question-answering (QA) tasks in Indic languages, there are several directions for future research and development:

1. **Expansion to More Languages**
   Extending the dataset to cover additional low-resource Indic languages, such as Assamese, Manipuri, and Santali, to improve multilingual accessibility and representation.

2. **Domain-Specific QA Datasets**
   Creating specialized datasets for legal, medical, and financial domains to improve real-world applicability in Indic languages.

3. **Multimodal QA for Indic Languages**
   Extending the dataset to incorporate images, videos, and speech, enabling multimodal question-answering for a more inclusive AI ecosystem.

4. **Interactive and Real-World Applications**
   Deploying QA models trained on IndicSQuAD into real-world applications, such as chatbots, voice assistants, and educational tools, to enhance accessibility and usability.

Limitations

A limitation of IndicSQuAD is that it is created through translation of SQuAD rather than native annotation. This reliance on translated data may reduce the linguistic authenticity of the contexts and questions. It can also introduce artifacts such as unnatural phrasing, loss of cultural nuances, or inconsistencies in answer-span alignment. Furthermore, the dataset may not fully reflect the diversity of information-seeking behavior found in native speakers of Indic languages.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Preprint*, arXiv:2003.05002.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Raviraj Joshi. 2022a. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Raviraj Joshi. 2022c. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan,

Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Gokul Karthik Kumar, Abhishek Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022a. Mucot: Multilingual contrastive training for question-answering in low-resource languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 15–24.

Gokul Karthik Kumar, Abhishek Singh Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022b. Mucot: Multilingual contrastive training for question-answering in low-resource languages. *Preprint*, arXiv:2204.05814.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

Vidula Magdum, Omkar Jayant Dhekane, Sharayu Sandeep Hiwarkhedkar, Saloni Sunil Mittal, and Raviraj Joshi. 2023. mahanlp: A marathi natural language processing library. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 34–40.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ryokan Ri, Shun Kiyono, and Sho Takase. 2024. Self-translate-train: Enhancing cross-lingual transfer of large language models via inherent capability. *Preprint*, arXiv:2407.00454.

Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context. *arXiv preprint arXiv:2409.08706*.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan, Sumon Kanti Dey, Amzad Hossain Rafi, Amzad Hossain Rafi, Ashraf Hasan Sirajee, and Jens Lehmann. 2024. Banglaquad: A bengali open-domain question answering dataset. *Preprint*, arXiv:2410.10229.

Ghatage Ruturaj, Kulkarni Aditya Ashutosh, Patil Rajlaxmi, Endait Sharvi, and Joshi Raviraj. 2023. Mahasquad: Bridging linguistic divides in marathi question-answering. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 497–505.

Abhishek Kumar Singh, Vishwajeet kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. *Preprint*, arXiv:2407.13522.

Adhitya Thirumala and Elisa Ferracane. 2022. Extractive question answering on queries in hindi and tamil.

A Appendix

Language	Model link
Marathi	marathi-squad-bert
Hindi	hindi-squad-bert
Bengali	bengali-squad-bert
Telugu	telugu-squad-bert
Tamil	tamil-squad-bert
Gujarati	gujarati-squad-bert
Punjabi	punjabi-squad-bert
Kannada	kannada-squad-bert
Oriya	oriya-squad-bert
Malayalam	malayalam-squad-bert

Table 4: Language-specific SQUAD BERT models on HuggingFace

HILIGAYNER: A Baseline Named Entity Recognition Model for Hiligaynon

James Ald Teves, Ray Daniel Cal, Josh Magdiel Villaluz, Jean Malolos,
Mico Magtira, Ramon Rodriguez, Joseph Marvin Imperial and Mideth Abisado
National University Philippines
jrimperial@national-u.edu.ph, mbabisado@national-u.edu.ph

Abstract

The language of Hiligaynon, spoken predominantly by the people of Panay Island, Negros Occidental, and Soccsksargen in the Philippines, remains underrepresented in language processing research due to the absence of annotated corpora and baseline models. This study introduces **HILIGAYNER**, the first publicly available baseline model for the task of Named Entity Recognition (NER) in Hiligaynon. The dataset used to build HILIGAYNER contains over 8,000 annotated sentences collected from publicly available news articles, social media posts, and literary texts. Two Transformer-based models, mBERT and XLM-RoBERTa, were fine-tuned on this collected corpus to build versions of HILIGAYNER. Evaluation results show strong performance, with both models achieving over 80% in precision, recall, and F1-score across entity types. Furthermore, cross-lingual evaluation with Cebuano and Tagalog demonstrates promising transferability, suggesting the broader applicability of HILIGAYNER for multilingual NLP in low-resource settings. This work aims to contribute to language technology development for underrepresented Philippine languages, specifically for Hiligaynon, and support future research in regional language processing.[1]

1 Introduction

The coverage and representation of diverse regional languages play a key role in the widespread adoption of any AI-based technology across the globe. While English remains the most highly researched and high-resourced language, initiatives from the research community, such as the SEACrowd (Cahyawijaya et al., 2025; Lovenia et al., 2024) for Southeast Asian languages, Masakhane (Adelani et al., 2023, 2021) for African languages, and Aya Project (Üstün et al., 2024; Singh et al., 2024) for global participation, have effectively made its impact to close the AI language gap (Bassignana et al., 2025; Pava et al., 2025).

A recent survey of digital support levels of languages showed that regional Philippine languages are among the lowest representations worldwide (Simons et al., 2022). One particular language is **Hiligaynon**[2], which is an Austronesian regional language spoken by over 10 million people in Western Visayas, particularly Panay Island, Negros Occidental, and Soccsksargen (McFarland, 2008; Robles, 2012). To initiate a step towards progress in Hiligaynon representation, researchers are encouraged to build resources and corpora for fundamental natural language processing tasks. One of these fundamental tasks is Named Entity Recognition (NER) or the task of automatic identification of textual mentions of persons, organizations, locations, and related categories (Nadeau and Sekine, 2007; Tjong Kim Sang and De Meulder, 2003a; Yadav and Bethard, 2018).

In this work, we present HILIGAYNER, the first publicly available NER corpus and finetuned models for Hiligaynon. Specifically, our contributions towards democratizing language resources for Hiligaynon are as follows:

1. A compilation of cleaned sentence-level Hiligaynon dataset of over 8,000 entries from online publicly accessible news articles, social media posts, and translated texts.

2. A compilation of span-level BIO-encoded annotations of the Hiligaynon dataset for the named entity recognition task (NER), specifically covering four entity categories (PER, ORG, LOC).

3. Two finetuned multilingual Transformer-based models, mBERT and XLM-RoBERTa,

[1] Code and data: https://github.com/jvlzloons/HiligayNER

[2] https://www.ethnologue.com/language/hil/

for token-level sequence labeling of Hiligaynon texts.

By releasing the dataset, model checkpoints, and evaluation scripts under an open license, we aim to supply the foundational tools required for broader NLP development in Western Visayas and the wider Philippine research community.

2 Related Works

Robust NER systems enable downstream applications such as knowledge-graph construction, information retrieval, and domain-specific analytics (Zhou et al., 2019). State-of-the-art performance is now achieved by combining lexicon-based gazetteers (Rijhwani et al., 2020), data-augmentation techniques (Yaseen and Langer, 2021), and deep neural architectures ranging from BiLSTM-CRF (Chiu and Nichols, 2016) to multilingual transformer encoders (Cotterell and Duh, 2017; Tan et al., 2024). Early Philippine NER studies concentrated almost exclusively on Tagalog, the national language. Statistical sequence models dominated. (Alfonso et al., 2013) applied Conditional Random Fields (CRF) to biographical texts and reported an F1 of 83%, while (Ebona et al., 2014) achieved 80.5% with a maximum-entropy classifier on short-story data. Subsequent CRF experiments on a larger newswire corpus produced a lower but still respectable 75.7% overall F1 (Cruz et al., 2016).

Cebuano, the second most widely spoken native tongue in the country, received attention slightly later. Maynard's rule-based adaptation of the ANNIE system yielded 69.1% F1 on a modest test set (Maynard et al., 2003). Cross-lingual neural CRFs, transferring knowledge from Tagalog, pushed performance to 81.8% (Cotterell and Duh, 2017). More recently, (Gonzales et al., 2022) introduced a hybrid CNN–BiLSTM pipeline that surpassed 95% precision and recall, albeit on only 200 manually annotated news articles. The largest Cebuano-based research to date is CebuaNER (Pilar et al., 2023a), which released a 4,258 article gold-standard corpus and baseline CRF/BiLSTM models that exceeded 70% F1 across entity classes. These milestones underscore both the feasibility and the demand for regional-language NER resources in the Philippines.

In contrast, Hiligaynon still lacks a public NER corpus or baseline model. Computational work has been limited to tokenization heuristics and the compilation of morphosyntactic lexicons (McFarland, 2008); no peer-reviewed study has tackled entity annotation or sequence labelling. This shortfall hampers information-extraction pipelines for regional journalism, public administration, and social-media analytics in Western Visayas, where Hiligaynon is the dominant medium.

3 Building HILIGAYNER: A Baseline NER Model for Hiligaynon

3.1 Dataset Collection

HILIGAYNER was assembled in three sequential phases: data collection, expert annotation, and reliability testing. Five online platforms hosting publicly available content were crawled to capture a sizeable representation of contemporary Hiligaynon texts as reported in Table 1. Each row in Table 1 refers to a single sentence extracted from the respective source. The dataset was segmented at the sentence level to facilitate BIO tagging and sentence-level NER annotations. The initial, raw collection comprised 17,647 sentences, but was reduced to 8,082 after preprocessing to remove malformed strings, empty lines, and non-Hiligaynon texts.

Source	Original	Cleaned
Ang Pulong Sang Dios	11,000	5,500
Ilonggo News Live	3,925	1,877
Hiligaynon News and Features	2,281	276
Bombo Radyo Bacolod	286	276
Ilonggo Balita sa Uma	155	153

Table 1: Statistics of publicly available data sources used in building HILIGAYNER.

3.2 Annotation Process and Reliability Testing

Three (3) undergraduate linguistics students who are also native speakers of Hiligaynon were tasked to annotate the corpus using Label Studio (Tjong Kim Sang and De Meulder, 2003a). The guidelines for annotating follow the CoNLL-2003 BIO convention (Tjong Kim Sang and De Meulder, 2003b) with four entity categories: Person (B-PER and I-PER), Organization (B-ORG and I-ORG), Location (B-LOC and I-LOC), and Other (OTH). For reference, in BIO tagging for NER, the B-prefix represents the first token of a named entity, while the I-prefix represents subsequent terms of a named entity. Refer to an example of a tagged sentence below using the BIO convention:

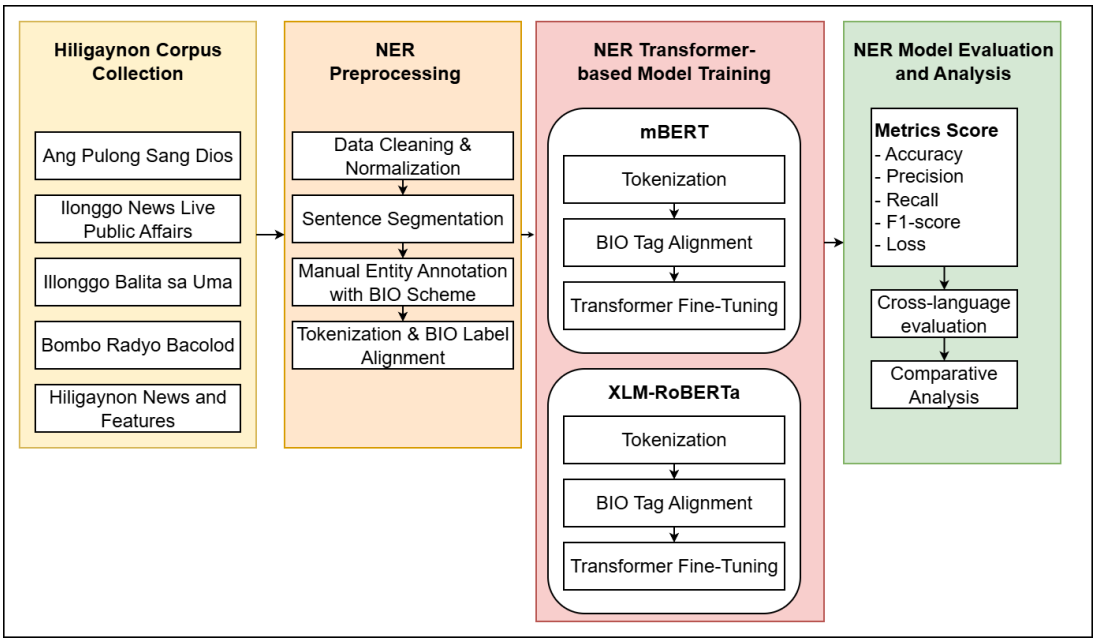


Figure 1: The overall methodology of developing HILIGAYNER using annotated news articles, social media posts, and literary text datasets in Hiligaynon using Transformer architectures mBERT and XLM-RoBERTa.

B-PER Aling **I-PER** Myrna **O** went **O** to **B-LOC** Iloilo **I-LOC** City .

The annotators received ten hours of joint training, including pilot rounds on 250 sentences with adjudication by a supervising linguist. Disagreements were resolved through consensus meetings and the final labels were exported in CoNLL format. To assess the reliability of the annotations, a stratified 10% subset of the corpus was selected and annotated independently by all three annotators. Cohen's κ was then computed based on pairwise comparisons within this overlapping subset to measure annotation consistency. Cohen's κ, a statistical metric widely adopted in NER studies (Artstein and Poesio, 2008; Tjong Kim Sang, 2002). The remaining portion of the dataset was divided among annotators for individual annotation. Table 2 reports on the scores showing an observed agreement = 0.9493, expected agreement = 0.7273, and yielding $\kappa = 0.8141$. According to conventional interpretation, a $\kappa \geq 0.80$ equates to substantial agreement, which denotes that the annotations of the HILIGAYNER dataset are of high quality and suitable for reproducible model training.

3.3 Finetuning

To establish strong baselines for HILIGAYNER, we fine-tuned two multilingual transformer encoders

Metric	Value
Observed Agreement	0.9493
Agreement by Chance	0.7273
Cohen's κ	0.8141

Table 2: Cohen's κ agreement results from annotations.

Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) using the standard token-classification pipeline in Hugging Face Transformers (Wolf et al., 2020). Both models are pretrained on large cross-lingual corpora and have shown competitive zero-shot and few-shot performance on sequence-labelling tasks (Conneau et al., 2020; Nakayama, 2019).

Multilingual BERT (mBERT). mBERT is a 12-layer, 768-hidden, 12-head encoder trained on Wikipedia dumps from 104 languages (Devlin et al., 2019). For NER, we attach a softmax-classifier head that maps each contextual token representation ht to a probability distribution over the four entity tags (PER, ORG, LOC, OTH):

$$P(\hat{y}|x = \prod_{t=1}^{T} softmax(Wh_t + b) \quad (1)$$

XLM-RoBERTa. XLM-RoBERTa extends the vanilla RoBERTa architecture (Pires et al., 2019) to

100 languages, pretrained on 2.5 TB of CommonCrawl with a SentencePiece tokenizer and larger capacity (24 layers, 1024 hidden, 16 heads) (Conneau et al., 2020). We replicated the mBERT fine-tuning recipe but lowered the learning rate to 3×10^{-5}, following XLM-R recommendations. Empirically, XLM-R attains higher recall on low-frequency tags, confirming earlier cross-lingual findings (Conneau et al., 2020).

4 Result and Discussion

4.1 Training mBERT and XLM-RoBERTa

Figures 2 and 3 plot the optimization trajectories for mBERT and XLM-RoBERTa, respectively. In both cases, the training loss decays monotonically during the first 100 batches and flattens thereafter, signaling rapid convergence under the chosen hyperparameters. Validation loss closely tracks the training curve and stabilizes at <0.05, indicating an absence of over-fitting.

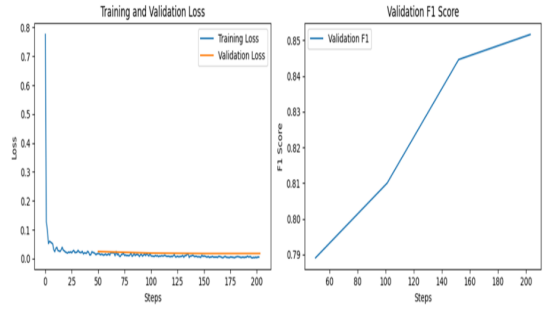


Figure 2: Training loss, validation loss, and F1 score per training step for the finetuned mBERT model.

Figure 3: Training loss, validation loss, and F1 score per training step for the finetuned XLM-RoBERTa model.

Figures 2 and 3 reveal rapid, stable convergence. Training loss drops sharply and levels off; validation loss mirrors this trajectory, remaining below 0.05. F1 improves in tandem—mBERT from 0.79 to 0.87, XLM-R to 0.88—without divergence between training and validation curves. The results confirm that the three-epoch, AdamW fine-tuning regimen achieves generalisation without over-fitting.

4.2 Model Evaluation

Tables 3 and 4 report token-level precision, recall, and F1 for the two Transformer-based models. In the case of mBERT, the model attains a macro F1 of 0.86, with near-perfect recognition of Person-based named entities at 0.96 and 0.94 for B-PER and I-PER. Location-based entities follow as the second-most correctly recognized at 0.83 and 0.82 for B-LOC and I-LOC. At the same time, Organization remains the most challenging entity to recognize for mBERT at 0.82 and 0.79. Nonetheless, these values are all relatively decent performances given that they exceed the 0.80 benchmark.

In the case of XLM-RoBERTa, we see a comparable high performance where Person-based entities are the most correctly recognized span, giving 0.96 and 0.94 for B-PER and I-PER. Location entities scored moderately, with B-LOC of 0.82 and I-LOC of 0.84 for F1, while organization entities remained the most challenging, yielding 0.81 for B-ORG and 0.79 for I-ORG.

For both models, we observe a general pattern where performance metrics correlate with entity tag frequency, with higher scores in categories with larger support counts (e.g., I-PER with 2,181 instances) compared to less frequent categories such as B-ORG (505 cases). These findings are consistent with prior multilingual-NER evaluations showing that pretrained transformers handle person names best and struggle with organization boundary cues (Conneau et al., 2020; Pilar et al., 2023b).

The study reports token-level precision, recall, and F1 scores as the primary evaluation metrics. Entity-level evaluation was not conducted, as the scope of this work is to establish a baseline for Hiligaynon NER using token-level annotation and modeling. The evaluation approach follows the convention used in the recently published CebuaNER study (Pilar et al., 2023a), which also adopted token-level reporting as a standard for establishing baselines in low-resource Philippine languages. The researchers recognize that span-level evaluation provides a stricter measure of system performance and leave this as an important direction for future work.

4.3 Error Analysis

Figures 4 and 5 expose the distribution of residual errors after fine-tuning for XLM-RoBERTa and

Tagset	Precision	Recall	F1-Score	Support
B-PER	0.95	0.97	0.96	1,754
I-PER	0.93	0.94	0.94	2,181
B-LOC	0.79	0.86	0.83	565
I-LOC	0.82	0.83	0.82	1,237
B-ORG	0.77	0.87	0.82	505
I-ORG	0.77	0.82	0.79	944

Table 3: Performance of the finetuned mBERT model using HILIGAYNER across NER categories.

Tagset	Precision	Recall	F1-Score	Support
B-PER	0.95	0.97	0.96	1,777
I-PER	0.93	0.95	0.94	2,268
B-LOC	0.79	0.86	0.82	577
I-LOC	0.83	0.85	0.84	1,228
B-ORG	0.76	0.87	0.81	514
I-ORG	0.74	0.84	0.79	910

Table 4: Performance of the finetuned XLM-RoBERTa model using HILIGAYNER across NER categories.

mBERT, respectively. In both matrices, person entities dominate the main diagonal B-PER and I-PER account for > 96% of their respective instances, confirming that multilingual transformers consistently capture personal-name cues. Both models maintain negligible cross-category bleed between person and non-person tags (< 0.5%), and false positives for rare classes remain below 1% of total predictions. The matrices, therefore, corroborate the aggregate metrics where the entity segmentation is reliable for PER, adequate for LOC, and bottlenecked by ORG boundary precision. Targeted gazetteer augmentation or span-level objectives should prioritize the ORG–LOC boundary to yield substantive gains.

4.4 Crosslingual Performance with Cebuano and Tagalog

Table 5 presents the crosslingual performance of both the mBERT and XLM-RoBERTa models finetuned on HILIGAYNER. Results from zero-shot evaluation on Cebuano and Tagalog yield macro F1 scores of ≈ 0.46 (0.44 to 0.46) for both languages, which are comparable to earlier Philippine crosslingual results (Cotterell and Duh, 2017; Pires et al., 2019). Precision, on the other hand, is marginally higher on Cebuano, reflecting closer lexical affinity within the Central Philippine subgroup (Imperial and Kochmar, 2023a,b). Although lower than in-language scores, the outcome demonstrates that the

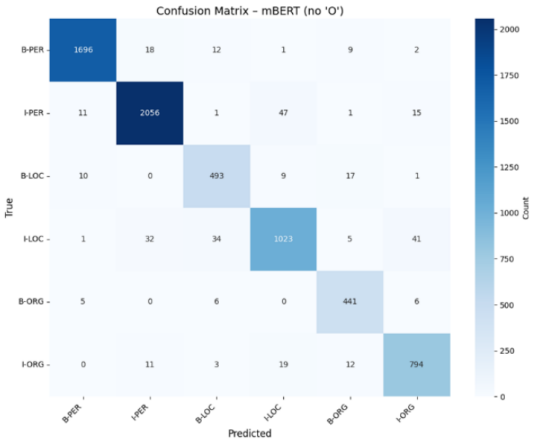


Figure 4: Confusion matrix of the finetuned mBERT model using HILIGAYNER across NER categories, omitting the OTH tag for brevity.

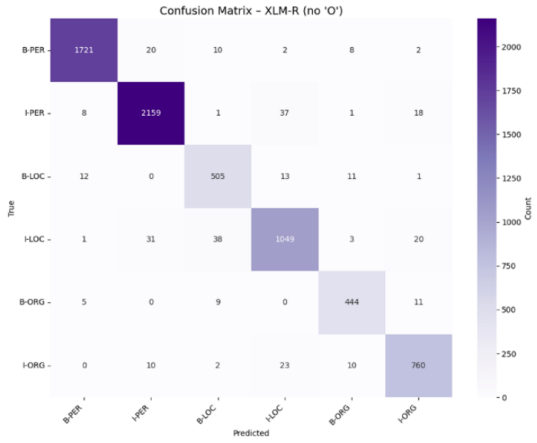


Figure 5: Confusion matrix of the finetuned XLM-RoBERTa model using HILIGAYNER across NER categories, omitting the OTH tag for brevity.

released model checkpoints offer a viable starting point for rapid adaptation to neighboring languages. The higher performance of Cebuano over Tagalog may be attributed to its lexical and syntactic proximity to Hiligaynon, as both belong to the Central Philippine language subgroup and share similar morphological patterns and word order. In contrast, Tagalog, while still within the same Austronesian family, exhibits more divergent lexical structures. It is also worth mentioning that Cebuano, Tagalog, and Hiligaynon are written using the Latin script, which may have contributed to their crosslingual generalization.

5 Conclusion

This study presents HILIGAYNER, the first publicly available baseline NER model and dataset

Metrics	mBERT		XLM-RoBERTa	
	CEB	TAG	CEB	TAG
Precision	0.4402	0.3998	0.4340	0.3894
Recall	0.4773	0.4991	0.4984	0.5221
F1-Score	0.4580	0.4439	0.4640	0.4461
Accuracy	0.9727	0.9639	0.9736	0.9633

Table 5: Cross-lingual performance of the finetuned mBERT and XLM-RoBERTa models using HILIGAYNER with Cebuano and Tagalog languages.

for Hiligaynon, a digitally underrepresented regional language in Western Visayas, Philippines. The HILIGAYNER dataset was systematically annotated under CoNLL BIO guidelines by native speakers and validated with strong inter-annotator agreement ($\kappa = 0.81$). Finetuning experiments on two multilingual models mBERT and XLM-RoBERTa yielded macro F1 ≈ 0.86, surpassing the 0.80 threshold on all primary tags (Person, Location, and Organization), which presents a high-quality baseline performance. Additional error analysis showed that residual confusion is concentrated in organization–location boundaries, while zero-shot transfer to Cebuano and Tagalog achieved competitive F1 ≈ 0.46, confirming cross-lingual utility.

By releasing the corpus, annotation protocol, training scripts, and model checkpoints under a permissive license, we provide a reproducible foundation for downstream Hiligaynon NLP and rapid adaptation to related Central Philippine languages (Imperial and Kochmar, 2023a,b). For future work, we recommend further efforts on increasing and diversifying the content of HILIGAYNER, such as adding finer-grained tags (e.g., Event, Date), exploring domain-adaptive pre-training on regional news, and incorporating gazetteer-augmented span objectives to improve organization recognition. These directions will further advance language technology for Hiligaynon and other low-resource languages.

Acknowledgments

All datasets collected for this study are publicly available and are used for non-commercial research purposes. We acknowledge the sources of the Hiligaynon data from Ang Pulong Sang Dios, Ilonggo News Live, Hiligaynon News and Features, Bombo Radyo Bacolod, and Ilonggo Balita sa Uma. We acknowledge the financial support provided by the National University Philippines and the Department of Science and Technology for the General Access Multilingual Online Tool for Public Health Drug-Reporting (GamotPH) Project. JAT, RDC, and JMV completed this work as part of their internship at National University Philippines.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, and 46 others. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

R. Alfonso, J. Cheng, and E. Bautista. 2013. Named-entity recognition in tagalog using conditional random fields. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27)*, Taipei, Taiwan.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI gap: How socioeconomic status affects language technology interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18647–18664, Vienna, Austria. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino,

Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

R. Cruz, C. Cheng, and M. Roxas. 2016. Tagalog named-entity recognition using conditional random fields. In *Proceedings of the 8th Workshop on Asian Language Resources (ALR)*, pages 52–59.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Ebona, J. Golla, and M. Sison. 2014. Named-entity recognition on tagalog short stories using maximum entropy. *Philippine Computing Journal*, 9(2).

Joshua Andre Huertas Gonzales, J-Adrielle Enriquez Gustilo, Glenn Michael Vequilla Nituda, and Kristine Mae Monteza Adlaon. 2022. Developing a hybrid neural network for part-of-speech tagging and named entity recognition. In *Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference*, pages 7–13.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023a. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2003. Ne recognition in resource-poor languages using rule-based approaches: The case of cebuano. In *Proceedings of the LREC Workshop on Minority Languages*.

R. D. McFarland. 2008. *The Philippine Languages*. SIL International, Dallas, TX.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Hiroki Nakayama. 2019. seqeval: A python framework for sequence-labeling evaluation. https://github.com/chakki-works/seqeval. GitHub repository.

Juan Pava, Haifa Badi Uz Zaman, Caroline Meinhardt, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of llm development in low-resource language contexts. White paper, Stanford Institute for Human-Centered Artificial Intelligence.

Ma. Beatrice Emanuela Pilar, Dane Dedoroy, Ellyza Mari Papas, Mary Loise Buenaventura, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023a. CebuaNER: A new baseline Cebuano named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 792–800, Hong Kong, China. Association for Computational Linguistics.

Ma. Beatrice Emanuela N. Pilar, Ellyza Mari J. Papas, Mary Loise Buenaventura, Dane C. Dedoroy, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023b. Cebuaner: A new baseline cebuano

named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 37)*, pages 792–800, Hong Kong, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online. Association for Computational Linguistics.

C Robles. 2012. Hiligaynon: An endangered language. In *Multilingual Philippines [Author]. 2nd Philippine Conference Workshop on Mother Mother Tongue-Based Multilingual Education (MTBMLE 2), Iloilo*, volume 6.

Gary F. Simons, Paul Lewis, and Charles Fennig. 2022. Assessing digital support for the world's languages. Technical Report SIL International Working Paper, SIL International.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Gian Carlos Tan, Jhan Kyle Canlas, Ren Joseph Ayangco, Daeschan Blane Gador, Mico Magtira, Jean Malolos, Ramon Rodriguez, Joseph Marvin Imperial, and Mideth Abisado. 2024. CebBERT: A lightweight data-transparent DistilBERT model for Cebuano language processing. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 904–913, Tokyo, Japan. Tokyo University of Foreign Studies.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, pages 155–158.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003a. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003b. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Taha Yaseen and Philipp Langer. 2021. Data-augmentation strategies for low-resource named-entity recognition. In *Proceedings of ICON 2021: 18th International Conference on Natural Language Processing*, pages 280–292, Pune, India. ICON 2021.

Peng Zhou, Wei Shi, Jin Tian, and 1 others. 2019. Position-aware attention and memory for knowledge-graph construction. *Information Processing & Management*, 56(3).

An Investigation of Sentiment Polarity in Chinese VO Idioms Based on NLP Models

Xueyi Wen
School of Humanities
Nanyang Technological Univ.
XUEYI001@e.ntu.edu.sg

Hongzhi Xu
Inst. of Language Sciences
Shanghai Intl. Studies Univ.
hxu@shisu.edu.cn

Lili Yang
Sch. of Chinese Lang. and Lit.
Soochow University
llyang@suda.edu.cn

Abstract

Construction expressions pose significant challenges to Natural Language Processing (NLP) tasks such as natural language understanding and sentiment analysis due to their semantic opacity, where the overall meaning cannot be directly inferred from the literal meanings of their components. Chinese Verb-Object (VO) constructions are full of such special constructions and idioms, presenting specific challenges in sentiment analysis. This paper employs NLP and statistical methods to investigate the sentiment polarity of Chinese VO idioms and the correlations between sentiment polarity and literal and contextual meaning. The result of our analysis shows that Chinese VO idioms in general exhibit a negative sentiment bias, and their polarity is more closely tied to contextual factors than literal meaning. This finding aligns with the contextualized nature and semantic opacity of constructions.

Keywords: Chinese VO idioms; NLP; Sentiment Analysis

1 Introduction

Sentiment analysis, a key task in Natural Language Processing (NLP), aims to extract opinions, sentiments, evaluations, attitudes, and emotions from text (Liu, 2017). One major challenge of sentimental analysis is semantic opacity, particularly in expressions such as sarcasm, metaphor, and constructions, where traditional sentiment analysis struggles to capture the underlying meaning and emotional tendency.

Chinese, being a complex language and a major world language, displays frequent semantic opacity phenomena. The study of Chinese constructions' sentimental tendency is of significant importance for improving the performance of Chinese sentiment analysis systems. Attitudes and scalar are important factors impacting semantic opacity. Constructions, as non-recursive, non-trivial phrasal structures (Zhan, 2017), carry both attitudinal (sentiment polarity) and scalar meanings. Specifically, the attitudinal meaning, also termed emotional polarity, includes positive, negative, and neutral. However, due to the non-compositionality of form and meaning (Goldberg, 1995), it's often hard to induce construction meanings, which makes automatic parsing hard for them. Research shows that many Chinese constructions bear sentiment polarity (Zhan and Wang, 2020), with a tendency towards negative attitudes (Fang, 2017). This study focuses on Chinese Verb-Object (VO) idiomatic constructions, using large language models and statistical methods to explore sentiment polarity and its correlations behind its semantic opacity.

We address the following research questions:

1. Compared to literal meaning, do Chinese VO idioms tend to exhibit negative sentiment polarity?
2. Is the negative sentiment polarity of VO idioms related to their literal meaning?
3. Is the negative sentiment polarity of VO idioms related to context?

2 Related Work

Idiomatic expressions can, to some extent, reflect the sentiment polarity of their users. Many studies have used the sentiment-bearing properties of idioms to construct sentiment classification models. Xie and Wang (2014) utilized Chinese idiom resources, primarily four-character idioms, to build a novel unsupervised framework for training general-purpose sentiment classifiers. Williams et al. (2015) demonstrated that using English idioms as features can improve the performance of traditional sentiment analysis models. Wu and

Li (2022) constructed a Chinese construction corpus for sentiment analysis and found that although constructions are not the primary semantic units, they still carry a certain proportion of attitudinal semantic information. Similarly, Tahayna et al. (2022) showed that English idiom corpora annotated with sentiment polarity can enhance the performance of sentiment classifiers. These findings suggest a correlation between idioms and sentiment polarity, and several scholars have conducted research in this area. However, studies focusing specifically on Chinese VO idioms remain scarce, and there is a lack of established sentiment-annotated corpora for this category.

In Modern Chinese, there exists a large number of idiomatic constructions whose emotional meanings have been conventionalized through pragmatic usage. The classification and definition of constructions are diverse. Previous research on the sentiment of constructions has mostly been limited to individual cases or small categories. For example, Du (2005) investigated the "V/A 个 P" (*V/A gè P*, 'V/A one P') and "Q 才 VP" (*Q cái VP*, 'Q only then VP') structures in Chinese idioms, arguing that both function as exclamatory sentences expressing negation. Deng and Huang (2002) and Gan (2008) examined the "不 A 不 B" construction (*bù A bù B*, 'not A not B') from the perspectives of internal structural relations, constraints on negation, and processes of constructionalization. Li (2008) studied the pragmatic marker "问题是" (*wèntí shì*, 'the problem is') and its negative evaluative function. Li (2011, 2014) explored the constructions "X 真是(的)" (*X zhēnshì(de)*, 'X really is / what X really is') and "好你个 + X" (*hǎo nǐ ge X*, 'what a X / how X!') in terms of negative evaluation and underlying causes.

Nonetheless, research on the VO idiom category is limited, and large-scale statistical analyses beyond individual cases are lacking. Therefore, the present study focuses on the major category of VO idioms in Chinese, applying NLP methods to automatically assign sentiment values and employing statistical analysis to investigate sentiment tendencies within this relatively fixed category.

3 Construction of Chinese VO idiom sentimental corpus

The data for this study were drawn from the Chinese portion of the OpenSubtitles corpus in the Open Parallel Corpus (OPUS) (Tiedemann, 2012), the Chinese Linguistics Corpus (CCL) (Center for Chinese Linguistics, Peking University, 2003), and the *Xinhua Dictionary* (Chinese Language Editorial Committee, 2016). OPUS is a comprehensive database of parallel corpora widely used in machine translation research. It is compiled from aligned online translation texts collected from the internet. The Chinese portion of OpenSubtitles contains over 8 million lines and approximately 150 million characters.

The CCL Chinese corpus is developed by the Center for Chinese Linguistics at Peking University, comprising texts from the 11th century BCE to the present. It contains over 500 million characters of Modern Chinese and more than 200 million characters of Classical Chinese, with a total exceeding 700 million characters. In this study, the CCL corpus served as an additional large-scale source for retrieving sentences containing VO idioms, ensuring broader temporal coverage and linguistic variety.

We first summarized 540 Chinese VO-structured idioms trisyllabic form based on literature and dictionaries. The *Xinhua Dictionary* was used to provide both the literal meaning explanation and the idiomatic meaning explanation for each idiom, enabling analysis of the relationship between sentiment polarity and literal semantics.

Using regular expression matching, we retrieved sentences containing these VO idioms from the OPUS and CCL corpora. Considering the factors of idiom *usage frequency* and *familiarity*, idioms with fewer than 50 example sentences were excluded, resulting in 168 VO idioms. From these, we randomly sampled 25 idioms, totaling approximately 210,000 characters, for manual annotation. Each example sentence was assessed for semantic completeness and fluency to determine its validity. For valid sentences, we further annotated whether the VO instance was used in its literal sense or idiomatic sense. Accordingly, the annotation categories were: *invalid data*, *literal-meaning data*, and *idiomatic-meaning data*. The resulting manually annotated VO idiom corpus contains approximately 208,000 characters of valid data. The data information is shown in Appendix A.

4 Sentimental polarity analysis of Chinese VO idioms

4.1 Calculation and correlation model used

Our analysis addresses three related questions: whether a VO construction used idiomatically exhibits a systematically different sentiment polarity from its literal usage, and whether idiomatic polarity is predictable from the surrounding sentential context or its constituents. To answer these questions we use the sentence-level sentiment score as the fundamental observational unit and compute, for each annotated instance, up to three scalar values: the sentiment of the sentence with the VO used idiomatically ("idiom"), the sentiment of a sentence in which the same VO is used literally ("literal"), and the sentiment of the surrounding context obtained by masking the VO ("context"). The overall workflow of the annotation and sentiment scoring pipeline is illustrated in Figure 1. The values for each sentence are then aggregated within constructions to permit paired within-construction comparisons and pooled across constructions to assess overall tendencies.

As for statistical test, for each comparison, we first evaluate distributional assumptions using the Shapiro–Wilk test (α =0.05) together with graphical diagnostics. Where the paired differences or residuals approximate normality, we employ the paired Student's t-test for difference-in-means and Pearson's r for linear association; where normality or linearity is violated, we replace these with the Wilcoxon signed-rank test and Spearman's ρ, respectively. To mitigate issues arising from multiple per-construction tests, we control the false discovery rate via the Benjamini–Hochberg procedure and report effect sizes with confidence intervals. Figure 2 illustrates the logic of our first set of comparisons, which test whether idiomatic and literal sentiment values differ systematically both within and across VO constructions.

The second analysis examines whether idiomatic sentiment polarity is correlated with its surrounding context. For each idiomatic sentence, we remove the VO and compute a sentiment score for the remaining context. We then assess whether these contextual values covary with idiomatic sentiment, which would suggest that the local usage environment contributes to idiom polarity. Figure 3 depicts this procedure schematically. Finally, to account for unbalanced instance counts and construction-specific baselines, we also fit mixed-effects models with idiomatic sentiment as the dependent variable, contextual sentiment as a fixed effect, and random intercepts for construction.

Automated sentiment scoring and preprocessing are performed with standard Chinese NLP tools: Jieba is used for segmentation and SnowNLP for sentence-level sentiment estimation (Feng, 2012; Zhang, 2012). To ensure our conclusions are not artifacts of a particular scorer or small-sample idiosyncrasies, we report robustness checks including confidence intervals for key statistics, repetition of principal tests with corresponding nonparametric measures, and a manual sanity check on a held-out subset.

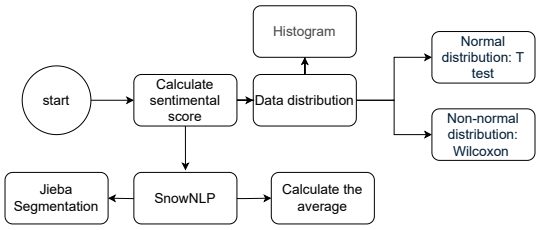


Figure 1: Research methodology workflow.

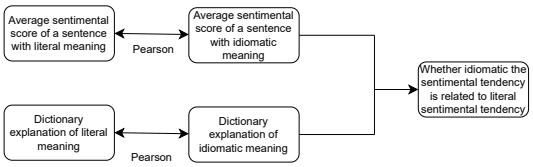


Figure 2: Correlation between idiomatic and literal sentiment polarity.

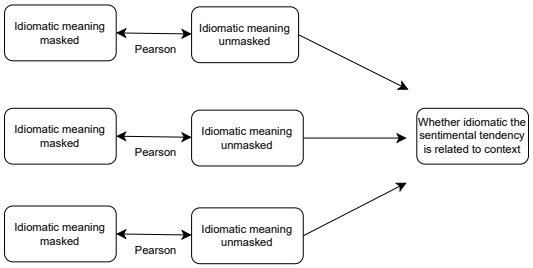


Figure 3: Correlation between idiomatic sentiment polarity and context.

4.2 Data distribution

We used Python's visualization tools to plot normal probability plots for all literal and idiomatic sentences to examine the data distribution, as shown in Figure 4. It can be observed that the

data roughly follow a normal distribution. Therefore, a Student's t-test was applied to assess correlations across the overall dataset. However, due to variations in the frequency of VO construction usage, familiarity, and degree of idiomatization in Chinese, some literal usage examples for certain idioms were limited in the corpus. In cases where the sentence distribution of individual VO constructions deviated from normality, the Wilcoxon signed-rank test was employed for correlation analysis.

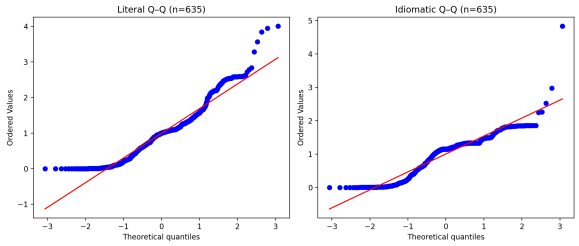


Figure 4: Overall data distribution: normal probability plots

4.3 Negative sentimental polarity of Chinese VO construction

As shown in Table 1, for most Chinese VO constructions, the average sentiment values of idiomatic usages are lower than those of their literal counterparts. The overall mean sentiment values across all VO constructions are 0.547 for literal meanings and 0.530 for idiomatic meanings, indicating a general tendency for idiomatic usages to convey slightly more negative sentiment.

To evaluate the statistical significance of these differences, paired t-tests were conducted for each VO construction as well as for the overall dataset; results for all constructions are provided in Appendix B1. The overall p-value is less than 0.01, confirming that the difference is statistically significant.

Table 1 lists several representative examples (a mix of idioms showing idiomatic negativity and idioms with positive idiomatic senses). The full set of 25 constructions and the complete significance test results are available in Appendix B2. In addition, the full table of all 25 constructions in Appendix C1, which reports sentiment scores based on dictionary definitions, also reveals a marked tendency for idiomatic meanings to be more negative than their literal counterparts (0.3414 vs. 0.63), further corroborating this finding.

Idiom	Literal Avg	Idiom Avg
拖下水 tuō xiàshuǐ ‘drag someone into trouble‘	0.506	0.468
有一手 yǒu yī shǒu ‘be skillful / have a trick‘	0.531	0.545
露头角 lù tóu jiǎo ‘begin to show one's talent‘	0.530	0.553
戴帽子 dài màozi ‘be cuckolded‘	0.592	0.588
吃闲饭 chī xiánfàn ‘do nothing productive‘	0.606	0.533
开绿灯 kāi l ǜ dēng ‘give the green light / permi	0.489	0.510
Overall	**0.547**	**0.530**

Table 1: Representative examples of average sentiment values (literal vs. idiomatic) for selected VO constructions. Full table of all 25 constructions and full significance test results are in Appendix B2.

Overall, Chinese VO idiomatic constructions tend to convey more negative sentiment compared to their literal meanings. However, in some cases, the literal meaning may exhibit more positive sentiment than the idiomatic meaning. This phenomenon can be partly explained by certain idioms such as “露头角” (*lù tóu jiǎo*, ‘begin to show one's talent’) and “有一手” (*yǒu yī shǒu*, ‘be skillful / have a trick’), whose idiomatic meanings are inherently positive and carry high sentiment values. In other cases, discrepancies may be due to limitations of the SnowNLP model in capturing idiomatic sentiment. For example, in the corpus, the sentence “她给他戴绿帽子” (*tā gěi tā dài l ǜ màozi*) is automatically assigned a sentiment score of 0.662 (positive) by SnowNLP, whereas its actual idiomatic meaning, referring to infidelity in a romantic relationship, is negative.

4.4 The difference of Chinese VO construction sentimental polarity between idiomatic and literal meaning

Based on idiomatic and literal meanings extracted from the *Xinhua Dictionary* and scored using SnowNLP, the comparison is shown in Table 2. Note that for the idioms “讨生活” (*tǎo shēnghuó*) and “吃闲饭” (*chī xiánfàn*), literal explanations were not listed in the dictionary and are therefore excluded from the comparison. Idioms such as “露头角” (*lù tóu jiǎo*) and “有一手” (*yǒu yī shǒu*) have positive sentiment values in their idiomatic meanings (0.619 and 0.894, respectively). For the remaining VO idioms, the sentiment val-

ues of the idiomatic meanings are lower than those of the literal meanings, consistent with our expectations. For instance, the literal meaning of "乱弹琴" (*luàn tán qín*) is "to play music in a disorderly manner without clear melody or rhythm, sounding disharmonious," which has a sentiment score of 0.891. Its idiomatic meaning, "to act recklessly or talk nonsense," has a sentiment score of 0.017. Both human intuition and the machine-assigned sentiment score clearly indicate a negative sentiment tendency when this VO construction is used idiomatically.

Idiom	Literal Sentiment	Idiom Sentiment
乱弹琴 luàn tán qín 'act recklessly / talk nonsense'	0.891	0.017
有一手 yǒu yī shǒu 'be skillful / have a trick'	0.999	0.894
露头角 lù tóu jiǎo 'begin to show one's talent'	0.797	0.719
拖下水 tuō xiàshuǐ 'drag someone into trouble'	0.070	0.216
Overall	0.634	0.341

Table 2: Representative examples of dictionary-based sentiment values for selected VO constructions. Full table of all 25 constructions is in Appendix C1.

The Pearson correlation coefficients for the two comparison methods are presented in Table 3. Both coefficients are close to 0, indicating that there is almost no linear relationship between the idiomatic sentiment and its literal meaning. Once an expression becomes idiomatic, its usage often diverges substantially from the original literal meaning. Some expressions carry negative connotations idiomatically, but their literal meanings may not convey strong negative sentiment, or may even exhibit opposite sentiment polarity.

Method	p-value	Pearson r
Sentence examples	0.929	-0.002
Dictionary explanations	0.567	0.121

Table 3: Correlation test between literal and idiomatic sentiment values.

4.5 Chinese VO construction sentimental polarity correlation with context

In this study, we assessed the sentiment of idiomatic usages within context by masking the idiom in sentences using the [SEP] token in Python. For example, the sentence:

> Eg. 4.5 你可以跳楼，但你会把我拖下水，那就让你成为杀人凶手。
> *nǒ1d0 kǒ113yǒ131 tiǒ0e0olǒ0f3u, dǒ0e0n nǒ1d0 huǒ0ec bǒ101 wǒ1d0 [SEP][SEP], nǒ0e0 jiǒ0f9 rǒ0e0ng nǒ1d0 chǒ0e9ngwǒ0e9i shǒ101rǒ0e9n xiǒ101ngshǒ1d0.*
> 'You may jump off the building, but you will [drag me into trouble], and then you will become a murderer.'

was processed as:

> 你 可 以 跳 楼 ， 但 你 会 把 我[SEP][SEP]， 那 就 让 你 成 为 杀 人凶手。
> *nǐ kěyǐ tiàolóu, dàn nǐ huì bǎ wǒ [SEP][SEP], nà jiù ràng nǐ chéngwéi shārén xiōngshǒu.*
> 'You may jump off the building, but if you [SEP][SEP], then you will become a murderer.'

The masked sentence was then tokenized and analyzed for sentiment. Differences between the idiomatic sentiment of VO constructions and the corresponding contextual sentiment were evaluated for statistical significance. Representative results are shown in Table 4. The full table of all constructions is provided in Appendix C2. For clarity and data hygiene, we removed two constructions ("生活" and "吃") since their literal usages are rare. We also calculated the overall effect sizes to quantify their practical magnitude. Aggregating over all 25 idiomatic sentences (N = 2489) yields Pearson $r = 0.634$ ($p < 0.001$), which corresponds to $r^2 \approx 0.402$. This shows that contextual embeddings explain roughly 40.2% of the variance in idiomatic sentiment scores. In regression terms, this implies Cohen's $f^2 = r^2/(1 - r^2) \approx 0.672$, a large effect by conventional benchmarks.

5 Discussion on sentiment tendencies and correlations of VO constructions

From the above data analysis, the feasibility and results of using NLP large models to investigate sentiment tendencies and correlations of Chinese constructions can be observed. This provides a foundation for future research, offering data-driven analysis and empirical support.

Starting from the research questions and based on the data analysis, it can be seen that Chinese

Idiom	Pearson r	p-value
拖下水 tuō xiàshuǐ 'drag someone into trouble'	0.955	0.001
有一手 yǒu yī shǒu 'be skillful'	0.978	0.000
开绿灯 kāi l ǜ dēng 'permit'	0.990	0.000
咬耳朵 yǎo ěrduo 'whisper in ears'	0.989	0.000
乱弹琴 luàn tán qín ' talk nonsense'	0.992	0.000
Overall	0.634	0.000

Table 4: Representative Pearson correlation between idiomatic sentiment and contextual embedding sentiment for selected VO constructions. Full results are in Appendix C2.

VO constructions exhibit a negative sentiment tendency compared to their literal meanings. This is consistent with the findings of Wu et al. (2017), who observed that negative-attitude meanings in Chinese idioms dominate both frequency and sentiment values in idiomatic corpora. This indicates that negative-attitude meanings in Chinese are largely conveyed through constructional meanings (Fang, 2017). At the same time, it reflects the significant asymmetry in the distribution of constructional meanings, the asymmetry of positive and negative attitudes at the constructional level in Chinese, and the manifestation of the negativity bias at the cognitive level (Rozin and Royzman, 2001).

Regarding the correlation analysis of idiomatic sentiment, two approaches were considered: dictionary-based sentiment correlations and literal-to-idiomatic sentiment correlations. The results show that the sentiment tendency of VO idioms is not correlated with their literal meanings. As a constructional structure, VO idioms cannot have their meaning or form predicted from their component parts once they have become a conventionalized expression (Goldberg, 1995). This is consistent with cognitive construction grammar perspectives. Meanwhile, the sentiment of the context surrounding idioms is correlated with the idiomatic sentiment but not with the literal meaning. This also highlights the implicit nature of idiomatic sentiment: negative evaluations in Chinese VO idioms are associated with contextual sentiment.

For example, consider the sentence processed in the study (see Example in 4.5):

Here, the idiom 拖下水 (*tuō xiàshuǐ* 'drag someone into trouble') was masked with `[SEP]` tokens for sentiment analysis. This illustrates that the contextual sentiment is correlated with the idiomatic sentiment, while it is not correlated with the literal meaning of the construction. This aligns with the notion of "constructionalization context" proposed by Traugott and Trousdale (2013), referring to the multifaceted linguistic environment influencing the formation of constructions, including discourse or textual context. The occurrence of a unit, partially or entirely, depends on its context, which can be described through various relations, including syntactic, morphological, phonological, semantic, and pragmatic functions.

6 Limitations

Despite the contributions of this study, several limitations should be acknowledged. First, the data scale could be further expanded. In the current study, only 25 commonly used VO idioms with at least 50 occurrences in the corpus were randomly selected. Therefore, the present study should be viewed as an exploratory case study illustrating the feasibility of combining NLP-based sentiment analysis with statistical testing. Future work will expand the coverage to a larger set of idiomatic expressions, allowing us to test whether the negative polarity tendency observed here generalizes across the lexicon. Second, the NLP model employed may have limitations in sentiment analysis accuracy, as automatic scoring may not fully capture nuanced idiomatic meanings. Third, the contextual information used in this study is limited to the single sentence containing the construction, which may omit discourse-level context and affect sentiment interpretation.

Future research directions include enlarging the database and data selection, collecting more examples of VO idioms, and improving the sentiment analysis models by incorporating pre-trained embeddings or additional sentiment features. Moreover, extending the context to include preceding and following sentences will better capture discourse-level sentiment effects for VO idioms.

7 Conclusion

This study investigated sentiment polarity and correlations of Chinese VO idioms using NLP-based sentiment analysis and statistical methods. The results show that VO idioms generally convey more

negative sentiment than their literal meanings, and that idiomatic sentiment is largely independent of the literal meaning but correlated with the contextual sentiment. These findings contribute to a better understanding of constructional semantics and sentiment in Chinese, providing a data-driven foundation for future research on idiomatic expressions. Furthermore, this study demonstrates the feasibility of leveraging NLP models for large-scale analysis of sentiment in constructions, offering empirical support for research in cognitive construction grammar and the pragmatics of idiomatic expressions.

References

Center for Chinese Linguistics, Peking University. 2003. Chinese linguistics corpus (ccl). http://ccl.pku.edu.cn/. Accessed: 2025-08-10.

Chinese Language Editorial Committee. 2016. *Xinhua Dictionary*, 12th edition. Commercial Press, Beijing.

Yingshu Deng and Gu Huang. 2002. On the negation meaning and constraints of "no a no b". *Chinese Language Learning*, (4):16–20.

Daoliu Du. 2005. A study on the structures "v/a个p" and "q才vp" in chinese idioms. *Journal of Chinese Language Learning*, (1):48–55.

Mei Fang. 2017. Conventionalization of negative evaluation expressions. *Zhongguo Yuwen*, (2):131–147.

Shuo Feng. 2012. Jieba: Chinese word segmentation. Accessed: 2025-08-10.

Lihao Gan. 2008. The construction meaning and negative tendency of "no a no b": A cognitive and pragmatic analysis. *Rhetoric Learning*, (2):56–60.

Adele E. Goldberg. 1995. *A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Xiaojun Li. 2011. The negative evaluative function of the construction "x真是(的)". *Contemporary Rhetoric*, (4):23–30.

Xiaojun Li. 2014. A study on the negative evaluation of the construction "好你个+x". *Journal of Pragmatics Research*, (3):45–53.

Zongjiang Li. 2008. Negative evaluation in the pragmatic marker "the problem is". *Modern Rhetoric*, (2):34–39.

Bing Liu. 2017. Sentiment analysis: Mining opinions, sentiments, and emotions.

Paul Rozin and Edward B. Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320.

B. M. Tahayna, R. K. Ayyasamy, and R. Akbar. 2022. Automatic sentiment annotation of idiomatic expressions for sentiment analysis task. In *IEEE Access*, volume 10, pages 122234–122242.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Elizabeth C. Traugott and Graeme Trousdale. 2013. *Constructionalization and Constructional Change*. Oxford University Press, Oxford.

L. Williams, C. Bannister, M. Arribas-Ayllon, A. Preece, and I. Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Yinqing Wu and Dejun Li. 2022. A study of chinese construction corpus compilation and application for sentiment analysis. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 475–484. Chinese Information Processing Society of China.

Yinqing Wu and 1 others. 2017. Sentiment distribution in chinese idioms. *Journal of Chinese Linguistics*, 45(2):123–145.

Songxian Xie and Ting Wang. 2014. Construction of unsupervised sentiment classifier on idioms resources. *Journal of Central South University*, 21(4):1376–1384.

Weidong Zhan. 2017. From phrase to construction: Theoretical issues in building a construction knowledge base. *Journal of Chinese Information Processing*, 31(1):230–238.

Weidong Zhan and Jiajun Wang. 2020. Representation and annotation of constructions in modern chinese. In *Proceedings of the 21st International Workshop on Chinese Lexical Semantics*.

Shuo Zhang. 2012. Snownlp: A python library for chinese text processing. Accessed: 2025-08-14.

A Data distribution

B Average sentiment values and significance tests for 25 Chinese VO constructions and their idiomatic meanings

C Full dictionary- and context-based sentiment tables

Table A1: Frequencies of literal and idiomatic uses of 25 Chinese VO idioms in the annotated corpus.

Idiom	**Literal sentences**	**Idiomatic sentences**
擦屁股 (cā pìgu) — wipe someone's butt	132	137
戴帽子 (dài màozi) — wear a hat	677	56
照镜子 (zhào jìngzi) — look in the mirror	180	59
回老家 (huí lǎojiā) — return to old home	206	45
拖下水 (tuō xiàshuǐ) — drag into water	7	186
开后门 (kāi hòumén) — open back door	54	15
说梦话 (shuō mènghuà) — talk in sleep	48	25
啃骨头 (kěn gǔtou) — gnaw bone	16	12
做文章 (zuò wénzhāng) — write article / exploit for gain	3	70
开绿灯 (kāi l ǜ dēng) — give the green light / permit	7	31
剃光头 (tì guāngtóu) — shave bald head	38	3
浇冷水 (jiāo lěngshuǐ) — pour cold water	30	19
有一手 (yǒu yī shǒu) — be skillful / have a trick	7	168
踢皮球 (tī píqiú) — kick the ball / pass responsibility	6	53
咬耳朵 (yǎo ěrduo) — whisper in ears	9	74
摇尾巴 (yáo wěibā) — wag tail / fawn on someone	66	8
讨生活 (tǎo shēnghuó) — make a living	0	94
乱弹琴 (luàn tánqín) — act recklessly / talk nonsense	8	47
露头角 (lù tóujiǎo) — begin to show one's talent	0	491
露马脚 (lù mǎjiǎo) — reveal one's hidden flaw	0	48
开夜车 (kāi yèchē) — work/study late at night	5	60
打哈哈 (dǎ hāhā) — laugh off / pretend to agree	1	89
出风头 (chū fēngtóu) — show off / attract attention	1	496
吃闲饭 (chī xiánfàn) — live idly / do nothing productive	0	86
睡大觉 (shuì dàjiào) — sleep heavily / nap	111	117
Total	1612	2489

Table B1: Average sentiment values for 25 Chinese VO constructions and their idiomatic meanings.

Idiom	**Literal Avg**	**Idiom Avg**
擦屁股 (cā pìgu) — wipe someone's butt	0.481	0.487
戴帽子 (dài màozi) — wear a hat	0.592	0.588
照镜子 (zhào jìngzi) — look in the mirror	0.528	0.518
回老家 (huí lǎojiā) — return to hometown	0.537	0.528
拖下水 (tuō xiàshuǐ) — drag someone into trouble	0.506	0.468
开后门 (kāi hòumén) — use backdoor	0.465	0.488
说梦话 (shuō mènghuà) — talk in one's sleep	0.499	0.498
啃骨头 (kěn gǔtou) — gnaw a bone	0.539	0.547
做文章 (zuò wénzhāng) — make an issue / exploit for gain	0.552	0.533
开绿灯 (kāi l ǜ dēng) — give the green light / permit	0.489	0.510
剃光头 (tì guāngtóu) — shave head	0.581	0.568
浇冷水 (jiāo lěngshuǐ) — pour cold water	0.480	0.483
有一手 (yǒu yī shǒu) — be skillful / have a trick	0.531	0.545
咬耳朵 (yǎo ěrduo) — whisper in ears	0.530	0.522
摇尾巴 (yáo wěibā) — wag tail / fawn on someone	0.514	0.505
踢皮球 (tī píqiú) — pass the buck / pass responsibility	0.520	0.498
讨生活 (tǎo shēnghuó) — make a living	0.558	0.545
乱弹琴 (luàn tánqín) — act recklessly / talk nonsense	0.541	0.535
露头角 (lù tóujiǎo) — begin to show one's talent	0.530	0.553
露马脚 (lù mǎjiǎo) — reveal one's hidden flaw	0.504	0.520
开夜车 (kāi yèchē) — work late / study late at night	0.505	0.523
打哈哈 (dǎ hāhā) — joke around / laugh off	0.511	0.528
出风头 (chū fēngtou) — show off / be in limelight	0.543	0.542
吃闲饭 (chī xiánfàn) — be idle / live idly	0.606	0.533
睡大觉 (shuì dàjiào) — sleep heavily / nap	0.516	0.521
Overall	0.547	0.530

Table B2: Significance tests for 25 Chinese VO constructions (paired t-test and Wilcoxon signed-rank test).

Idiom	t**-statistic**	t**-p-value**	w**-statistic**	w**-p-value**
擦屁股 (cā pìgu) — wipe someone's butt	-1.424	0.156	4470.000	0.333
戴帽子 (dài màozi) — wear a hat	0.480	0.631	740.000	0.371
照镜子 (zhào jìngzi) — look in the mirror	1.962	0.051	992.000	0.101
回老家 (huí lǎojiā) — return to hometown	1.628	0.105	472.000	0.615
拖下水 (tuō xiàshuǐ) — drag someone into trouble	2.638	0.009	4.000	0.109
开后门 (kāi hòumén) — use backdoor	-1.784	0.079	60.000	0.284
说梦话 (shuō mènghuà) — talk in one's sleep	0.087	0.931	134.000	0.443
啃骨头 (kěn gǔtou) — gnaw a bone	-0.468	0.644	29.000	0.470
做文章 (zuò wénzhāng) — make an issue / exploit for gain	0.797	0.428	0.000	0.250
开绿灯 (kāi l ǜ dēng) — give the green light / permit	-1.029	0.310	11.000	0.688
剃光头 (tì guāngtóu) — shave head	0.561	0.578	4.000	0.875
浇冷水 (jiāo lěngshuǐ) — pour cold water	-0.232	0.817	77.000	0.490
有一手 (yǒu yī shǒu) — be skillful / have a trick	-2.741	0.007	131.000	0.021
咬耳朵 (yǎo ěrduo) — whisper in ears	0.902	0.370	7.000	0.297
摇尾巴 (yáo wěibā) — wag tail / fawn on someone	0.853	0.397	10.000	0.578
踢皮球 (tī píqiú) — pass the buck	1.952	0.056	3.000	0.313
讨生活 (tǎo shēnghuó) — make a living	0.731	0.466	0.000	0.500
乱弹琴 (luàn tánqín) — act recklessly / talk nonsense	0.576	0.567	9.000	0.844
露头角 (lù tóujiǎo) — begin to show one's talent	-1.143	0.253	0.000	0.500
露马脚 (lù mǎjiǎo) — reveal one's hidden flaw	-0.785	0.436	1.000	1.000
开夜车 (kāi yèchē) — work late / study late at night	-1.471	0.146	3.000	0.156
打哈哈 (dǎ hāhā) — joke around / laugh off	-1.098	0.275	0.000	0.500
出风头 (chū fēngtou) — show off / be in limelight	0.093	0.926	1.000	1.000
吃闲饭 (chī xiánfàn) — be idle / live idly	4.272	0.000	0.000	0.500
睡大觉 (shuì dàjiào) — sleep heavily / nap	-1.295	0.196	2745.000	0.286
Overall	10.046	0.000	512790.000	0.000

Table C1: Full dictionary-based sentiment values for all 25 Chinese VO constructions. '–' indicates that a literal explanation was not listed in the dictionary.

Idiom	**Literal Sentiment**	**Idiom Sentiment**
擦屁股 (cā pìgu) — wipe someone's butt	0.116	0.031
戴帽子 (dài màozi) — wear a hat; (idiom) be cuckolded	0.950	0.451
照镜子 (zhào jìngzi) — look in the mirror	0.923	0.661
回老家 (huí lǎojiā) — return to hometown	0.975	0.433
拖下水 (tuō xiàshuǐ) — drag someone into trouble	0.070	0.216
开后门 (kāi hòumén) — use backdoor / bribe	0.012	0.190
说梦话 (shuō mènghuà) — talk in one's sleep	0.508	0.465
啃骨头 (kěn gǔtou) — gnaw a bone	0.367	0.355
做文章 (zuò wénzhāng) — make an article / exploit for gain	0.999	0.227
开绿灯 (kāi l ǜ dēng) — give the green light / permit	0.084	0.008
剃光头 (tì guāngtóu) — shave head	0.406	0.541
浇冷水 (jiāo lěngshuǐ) — pour cold water	0.077	0.342
有一手 (yǒu yī shǒu) — be skillful / have a trick	0.999	0.894
咬耳朵 (yǎo ěrduo) — whisper in ears	0.626	0.545
摇尾巴 (yáo wěibā) — wag tail / fawn on someone	0.999	0.073
踢皮球 (tī píqiú) — kick the ball / pass responsibility	0.990	0.215
讨生活 (tǎo shēnghuó) — make a living	–	0.618
乱弹琴 (luàn tán qín) — act recklessly / talk nonsense	0.891	0.018
露头角 (lù tóu jiǎo) — begin to show one's talent	0.797	0.719
露马脚 (lù mǎjiǎo) — expose one's hidden problem	0.614	0.270
开夜车 (kāi yèchē) — work/study late at night	0.224	0.189
打哈哈 (dǎ hāhā) — laugh off / pretend to agree	0.940	0.400
出风头 (chū fēngtóu) — show off / attract attention	0.493	0.094
吃闲饭 (chī xiánfàn) — live idly / do nothing productive	–	0.157
睡大觉 (shuì dàjiào) — sleep heavily / nap	0.784	0.406
Overall	0.634	0.341

Table C2: Full Pearson correlation between idiomatic sentiment and contextual embedding sentiment for all Chinese VO constructions.

Idiom	**Pearson r**	**p-value**
擦屁股 (cā pìgu) — wipe someone's butt	0.842	0.000
戴帽子 (dài màozi) — wear a hat; (idiom) be cuckolded	0.034	0.800
照镜子 (zhào jìngzi) — look in the mirror	0.153	0.201
回老家 (huí lǎojiā) — return to hometown	-0.061	0.691
拖下水 (tuō xiàshuǐ) — drag someone into trouble	0.955	0.001
开后门 (kāi hòumén) — use backdoor / bribe	-0.081	0.748
说梦话 (shuō mènghuà) — talk in one's sleep	-0.470	0.015
啃骨头 (kěn gǔtou) — gnaw a bone	-0.474	0.119
做文章 (zuò wénzhāng) — make an article / exploit for gain	0.828	0.379
开绿灯 (kāi l ǜ dēng) — give the green light / permit	0.990	0.000
剃光头 (tì guāngtóu) — shave head	0.381	0.619
浇冷水 (jiāo lěngshuǐ) — pour cold water	0.020	0.936
有一手 (yǒu yī shǒu) — be skillful / have a trick	0.978	0.000
咬耳朵 (yǎo ěrduo) — whisper in ears	0.989	0.000
摇尾巴 (yáo wěibā) — wag tail / fawn on someone	0.489	0.266
踢皮球 (tī píqiú) — kick the ball / pass responsibility	0.907	0.034
乱弹琴 (luàn tán qín) — act recklessly / talk nonsense	0.992	0.000
露头角 (lù tóu jiǎo) — begin to show one's talent	1.000	1.000
露马脚 (lù mǎjiǎo) — expose one's hidden problem	1.000	1.000
开夜车 (kāi yèchē) — work/study late at night	0.997	0.000
打哈哈 (dǎ hāhā) — laugh off / pretend to agree	1.000	1.000
出风头 (chū fēngtóu) — show off / attract attention	1.000	1.000
睡大觉 (shuì dàjiào) — sleep heavily / nap	0.993	0.000
Overall	0.689	0.000

Domain Adaptation for Multi-document Summarisation: A Case Study in the Medical Research Domain

Kushan Hewapathirana[1,2] Nisansa de Silva[1] C.D. Athuraliya[2] Piumi Kandanaarachchi[3]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]ConscientAI, Sri Lanka
[3]District General Hospital, Hambantota, Sri Lanka
{kushan.22, nisansa}@cse.mrt.ac.lk
cd@conscient.ai, piumikandanaarachchi@gmail.com

Abstract

Effectively summarising medical research is critical for supporting evidence-based decision-making in healthcare. While fine-tuning task-specific models on domain data is established practice, the comparative advantages over increasingly capable general-purpose LLMs remain an open question. This study systematically evaluates domain-adapted PRIMERA against several open-source large language models (LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, and Gemma 7B) in zero-shot settings using the MS^2 dataset, which includes $20,000$ systematic reviews summarising over $470,000$ medical studies. Fine-tuning leads to notable improvements in ROUGE scores—ROUGE-1 from 12.8 to 33.0, ROUGE-2 from 2.0 to 6.5, and ROUGE-L from 8.1 to 22.6. Comparative evaluation indicates that the fine-tuned model consistently achieves stronger performance across all three ROUGE metrics, human evaluations, and LLM-as-a-judge assessments. These results suggest that domain-adapted models can offer advantages over general-purpose LLMs in specialised settings, particularly where factual accuracy and coverage are critical, though at the cost of reduced flexibility across domains.

1 Introduction

Multi-document summarisation (MDS) is a challenging Natural Language Processing (NLP) task that aims to generate a summary by combining information from multiple sources. MDS involves handling conflicting, duplicate or complementary information to produce a summary that represents the overall content (Hewapathirana et al., 2023). The goal of MDS is to condense a collection of documents into a single, cohesive summary that captures the main points and ideas of the original documents (Ma et al., 2023; Afsharizadeh et al., 2022; Abid, 2022). Automatic summarisation can be classified into two primary categories: extractive and abstractive. *Extractive text summaries* contain keywords, phrases, and sentences that are extracted verbatim from the source documents (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021), whereas *abstractive text summaries* generate summaries that include paraphrased sentences and new terms that may not be found in the original documents (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021; Abid, 2022).

MDS can involve summarising different types of documents, including short sources, long sources, and hybrid sources. Short sources are documents such as tweets, product reviews, or headlines that convey a smaller amount of information. In contrast, long sources are lengthy documents such as news articles or research papers that contain a large amount of information and detail. Hybrid sources contain one or few long documents with several to many short documents, such as a scientific summary from a long paper with several corresponding citations (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021; Yu, 2022; Abid, 2022; Hewapathirana et al., 2023; Wolhandler et al., 2022).

MDS researchers use various techniques to generate abstractive summaries, such as natural language generation, deep learning models, and neural machine translation. These techniques enable the automatic creation of summaries that are coherent, informative, and useful (Ma et al., 2023; Afsharizadeh et al., 2022; Hewapathirana et al., 2024).

This study addresses three key research questions: (1) How does domain-adapted fine-tuning of a task-specific MDS model compare with task-specific baselines and recent open-source LLMs in zero-shot settings? (2) To what extent do fine-tuned models generalise beyond their training dataset, across domains and document types? (3) How do automatic metrics (ROUGE (Lin, 2004)) align with human evaluation and LLM-as-a-judge assessments in the medical summarisation domain?

Our main contributions include: A compara-

tive evaluation of fine-tuned PRIMERA against task-specific models (PEGASUS, LED) and open-source LLMs (LLaMA 3.2, Mistral, Gemma, OpenChat); empirical evidence demonstrating that domain adaptation provides measurable advantages over zero-shot approaches in specialized settings where factual accuracy is critical; and an analysis of generalization capabilities across document types, revealing both the strengths and limitations of domain-specific fine-tuning.

2 Related Work

MDS has evolved significantly with the introduction of transformer-based architectures. Early models such as BigBird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020) addressed the challenge of handling long input sequences via sparse and sliding window attention. These architectures were extended in summarisation tasks by replacing standard self-attention mechanisms in models like BART (Lewis et al., 2020), enabling more efficient long-context encoding. Hierarchical encoders have also been explored to capture inter-document structure more effectively.

Pre-trained transformer models such as BERTSUM (Liu and Lapata, 2019), BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and T5 (Raffel et al., 2020) have established strong baselines for abstractive summarisation. These models leverage large-scale pre-training to capture rich contextual information and have demonstrated high-quality generation across a variety of tasks. PRIMERA (Xiao et al., 2022), a Longformer Encoder-Decoder (LED)-based model trained with an entity-based pyramid pretraining strategy, has shown state-of-the-art performance in MDS benchmarks.

Domain-specific approaches such as CGSUM (Chen et al., 2022), which uses citation-guided selection for summarising scientific papers, and DAMEN (Moro et al., 2022), which incorporates indexing and discriminative filtering for medical MDS, illustrate the benefit of incorporating structural or domain-aware features into summarisation pipelines.

Recent work has also examined the challenge of synthesising sentiment or conflicting perspectives across multiple documents. DeYoung et al. (2024) proposed using Diverse Beam Search (Vijayakumar et al., 2016) to generate a range of candidate summaries, selecting the one most representative of the aggregate view. This improves robustness to variations in input structure and composition.

For training and evaluation, benchmark datasets such as DUC and TAC [1] [2] have historically been used, although they suffer from size and positional bias. More recent alternatives include Multi-News (Fabbri et al., 2019), WikiSum (Liu et al., 2018), and WikiHow (Koupaee and Wang, 2018), which offer larger and more diverse summary corpora. Additionally, datasets such as Rotten Tomatoes (Leon, 2020) have supported the evaluation of aggregation quality across subjective inputs.

In parallel, the rise of large language models (LLMs) has significantly impacted summarisation. While proprietary models such as GPT-4 and Claude have shown strong results, their closed nature and resource demands limit reproducibility (Laskar et al., 2023). Open-access models such as LLaMA 3.2 3B (Meta AI, 2024), Mistral 7B (Jiang et al., 2023), Gemma 7B (Team et al., 2024), and OpenChat 7B (Wang et al., 2023) offer competitive performance in summarisation while remaining lightweight enough for deployment in constrained academic or clinical environments. These models enable researchers to explore LLM-based summarisation in low-resource settings without sacrificing modern capabilities.

3 Model Selection

In this study, we selected models based on a combination of empirical performance in MDS, architectural diversity, domain relevance, and computational feasibility. Our selection process was guided by a thorough review of recent literature and benchmarking studies, with ROUGE scores (Lin, 2004) and adoption in the MDS community serving as key criteria.

After careful evaluation, we chose to assess the performance of three summarisation models: PRIMERA (Xiao et al., 2022), PEGASUS (Zhang et al., 2020), and Longformer Encoder-Decoder (LED) (Beltagy et al., 2020). PRIMERA is a state-of-the-art MDS model that leverages the Longformer architecture and an entity pyramid masking strategy to enhance content selection during pre-training. It has consistently outperformed earlier methods in benchmark evaluations (Afsharizadeh et al., 2022; Ma et al., 2023; DeYoung et al., 2024). PEGASUS, with its Gap Sentence Gen-

[1] https://duc.nist.gov/
[2] https://tac.nist.gov/

eration (GSG) objective, is particularly effective in generating summary-worthy sentences and has demonstrated strong performance on abstractive summarisation tasks. LED, with its sparse attention mechanism, serves as a strong baseline due to its efficiency in handling long input sequences and its use in earlier MDS studies.

In addition to these task-specific models, we evaluated a set of recent open-access LLMs with strong general-purpose summarisation capabilities: LLaMA 3.2 3B (Meta AI, 2024), Mistral 7B (Jiang et al., 2023), Gemma 7B (Team et al., 2024), and OpenChat 7B (Wang et al., 2023). These models represent lightweight alternatives to proprietary commercial LLMs and are increasingly being adopted in low-resource and open research settings. Though not explicitly fine-tuned for MDS, their instruction-following abilities allow for effective few-shot or zero-shot summarisation, making them valuable for comparative evaluation against domain-specific models.

These LLMs were selected based on availability, performance in recent summarisation benchmarks, and their suitability for deployment under academic resource constraints. All experiments involving LLMs were conducted on a server with a 4-core CPU, 64 GB RAM, and a single A2 GPU with 16 GB VRAM, which limited the feasibility of larger commercial models and motivated the use of accessible open-weight alternatives.

Together, these model selections enable a comprehensive comparison across specialised, pretrained summarisation models and general-purpose LLMs in the medical MDS setting.

3.1 MS^2 Dataset

Existing summarisation datasets often lack biomedical specificity, limiting their effectiveness for domain-specific summarisation tasks (Ma et al., 2023; Afsharizadeh et al., 2022; Abid, 2022). To address this, DeYoung et al. (2021) introduced the MS^2 dataset, specifically curated for biomedical document summarisation using systematic literature reviews. These reviews synthesize evidence from multiple studies, providing concise and clinically relevant summaries. e.g., a review on Vitamin B12 supplementation in older adults may aggregate diverse findings (Andrès et al., 2010).

The dataset was constructed by filtering the Semantic Scholar Open Research Corpus (Lo et al., 2020) using a multi-stage pipeline: keyword heuristics to identify systematic reviews (220K), a PubMed filter for biomedical relevance (170K), and a SciBERT-based classifier (Beltagy et al., 2019) for final selection, yielding 20K high-quality review-summary pairs. This makes MS^2 a robust benchmark for training and evaluating biomedical summarisation models with strong relevance to evidence-based healthcare applications.

Dataset		PRIMERA	PEGASUS	LED
Multi-News	R-1	**42.0***	32.0*	17.3*
	R-2	**13.6***	10.1*	3.7*
	R-L	**20.8***	16.7*	10.4*
Multi-Xscience	R-1	**29.1***	27.6*	14.6*
	R-2	**4.6***	4.6*	1.9*
	R-L	**15.7***	15.3*	9.9*
WikiSum	R-1	**28.0***	24.6*	10.5*
	R-2	**8.0***	5.5*	2.4*
	R-L	**18.0***	15.0*	8.6*
BigSurvey-MDS	R-1	23.9$^{\diamond}$	**38.9**†	39.8†
	R-2	4.1$^{\diamond}$	**9.0**†	9.4†
	R-L	11.7$^{\diamond}$	**16.2**†	16.1†
MS^2	R-1	12.8$^{\diamond}$	12.7$^{\diamond}$	**25.8**‡
	R-2	2.0$^{\diamond}$	1.5$^{\diamond}$	**8.4**‡
	R-L	8.1$^{\diamond}$	8.3$^{\diamond}$	**19.3**‡
Rotten Tomatoes	R-1	25.4$^{\bullet}$	**27.4**$^{\bullet}$	25.6$^{\bullet}$
	R-2	8.4$^{\bullet}$	**9.5**$^{\bullet}$	8.0$^{\bullet}$
	R-L	19.8$^{\bullet}$	**21.1**$^{\bullet}$	19.6$^{\bullet}$

Table 1: **ROUGE scores of selected models across different domains.** Datasets include *Multi-News* (Fabbri et al., 2019), *Multi-Xscience* (Lu et al., 2020), *WikiSum* (Liu et al., 2018), *BigSurvey-MDS* (Liu et al., 2023), *MS^2* (DeYoung et al., 2021), and *Rotten Tomatoes* (Leon, 2020). Sources: * Xiao et al. (2022), † Liu et al. (2023), ‡ DeYoung et al. (2021), $^{\bullet}$ Wang et al. (2022), $^{\diamond}$ Hewapathirana et al. (2023).

3.2 Evaluation Metrics

We evaluated the fine-tuned PRIMERA model using a combination of automated metrics, human evaluation, and LLM-based judgments to comprehensively assess its performance on the MS^2 dataset.

Automated Metrics. The primary automated metric used was ROUGE (Lin, 2004), a standard for MDS evaluation (Ma et al., 2023). ROUGE measures the overlap between the generated summary and the reference summary, focusing on precision and recall. We employed two key variants: ROUGE-N, which calculates *n*-gram overlap, and ROUGE-L, which assesses sentence-level similarity based on the longest common subsequence (Lin, 2004). These metrics enabled objective comparison across baseline and state-of-the-art models.

Human Evaluation. To assess qualitative performance, we conducted human evaluations focusing on five criteria: *Relevance*, *Coherence*, *Coverage*, *Conciseness*, and *Accuracy*. Three expert annotators from the medical domain independently rated 50 randomly sampled summaries. Inter-annotator agreement was measured using Krippendorff's Alpha (α) (Krippendorff, 1989), which is robust to multiple raters and missing data. This human assessment provided critical insight into the linguistic and domain-specific fidelity of the summaries.

LLM-as-a-Judge Evaluation To further assess summary quality from a model-based perspective, we employed the DeepEval framework[3] using Meta's LLaMA 3 90B Instruct model (`us.meta.llama3-2-90b-instruct-v1:0`) hosted via AWS Bedrock. Evaluation was conducted on a sample of 50 summaries due to cost constraints. By adopting DeepEval's standardized summarisation evaluation protocol[4], we ensured reproducibility and methodological consistency. This LLM-based evaluation complemented ROUGE and human assessments by providing judgments on factual consistency, coherence, and overall quality.

Domain-Specific Fine-Tuning. The medical domain represents a unique challenge due to its complexity and specialized vocabulary. To address this, we fine-tuned PRIMERA using the MS^2 dataset, comprising medical research papers. Our objective was to improve PRIMERA's performance in generating accurate and concise summaries tailored for biomedical literature.

Training Configuration. Fine-tuning was performed with carefully chosen hyperparameters: a learning rate of $5e$-05, batch size of 4, and 3 training epochs. We used the Adam optimizer with betas (0.9, 0.999) and epsilon $1e$-08, along with a linear learning rate scheduler. A random seed of 42 ensured reproducibility. The Hugging Face `trainer` API[5] was used to manage the training process efficiently on our available infrastructure.

[3] https://github.com/confident-ai/deepeval

[4] https://deepeval.com/docs/metrics-summarization

[5] https://huggingface.co/docs/transformers/main_classes/trainer

4 Results

4.1 Fine-tuned Model Performance

The fine-tuning of the PRIMERA model on the MS^2 dataset resulted in significant improvements in performance. Prior to fine-tuning, the ROUGE-1, ROUGE-2, and ROUGE-L scores were observed to be approximately 12.8, 2.0, and 8.1, as shown in Table 1. However, after fine-tuning the model, these scores significantly increased to 33.0, 6.5, and 22.6, as presented in Table 2. Notably, the fine-tuned PRIMERA model outperformed the state-of-the-art LED model which achieved ROUGE-1 and ROUGE-L scores of 25.8 and 19.3 respectively.

Table 2: Performance of our fine-tuned PRIMERA model on various domain-specific datasets

Dataset		PRIMERA
Multi-News	R-1	39.1
	R-2	11.7
	R-L	18.0
BigSurvey-MDS	R-1	33.0
	R-2	7.1
	R-L	12.7
MS^2	R-1	33.0
	R-2	6.5
	R-L	22.6

In order to assess the generalization capabilities of the fine-tuned model, we also evaluated its performance on a different domain dataset, the Multi-news dataset (Fabbri et al., 2019). This dataset primarily consists of news articles and their corresponding human-written summaries from the website `newser.com`. It encompasses a diverse range of news sources, making it more representative of real-world scenarios compared to previous datasets such as DUC and Newsroom (Fabbri et al., 2019). The fine-tuned model exhibited slightly lower ROUGE scores when tested on the Multi-news dataset. Although there was a slight drop in performance compared to the initial PRIMERA model performances, the results remained reasonable. Furthermore, we evaluated the fine-tuned model on the BigSurvey dataset (Liu et al., 2023), which consists of survey papers and their corresponding summaries. The dataset includes two levels of target summaries: a comprehensive long summary and a concise short summary. The fine-tuned PRIMERA model demonstrated improved performance on the BigSurvey dataset as well.

These findings demonstrate mixed generalization patterns. While the fine-tuned model shows strong performance on MS^2 and maintains im-

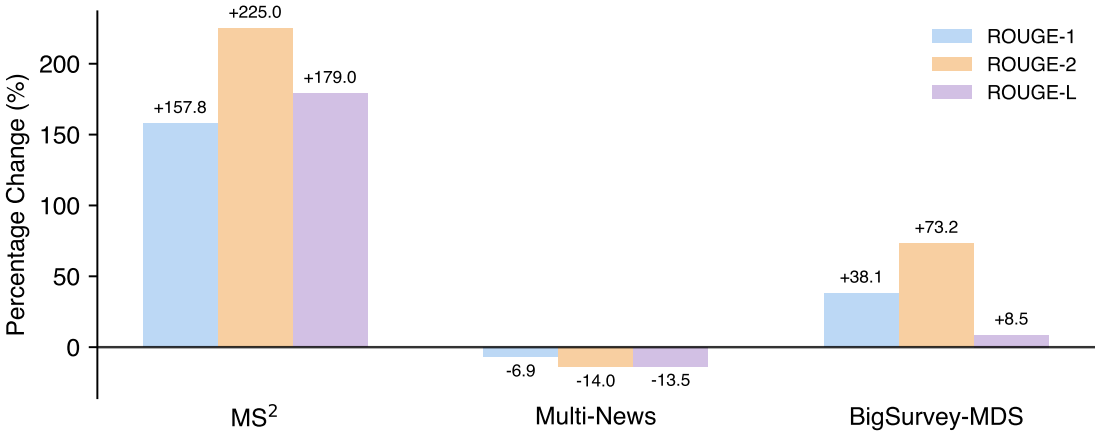


Figure 1: Percentage Improvement/Reduction of Fine-Tuned PRIMERA Model

provements on the structurally similar BigSurvey dataset (+38.1% R-1, +73.2% R-2, +8.5% R-L), it experiences modest performance degradation on Multi-News (−6.9% R-1, −14.0% R-2, −13.5% R-L). This suggests the model has adapted to the specific characteristics of medical research papers rather than learning a fully generalizable MDS strategy. The improved performance on BigSurvey likely reflects similarity in document structure (both are research papers with systematic organization) rather than pure domain transfer. These results indicate that domain-specific fine-tuning trades broad generalization for targeted performance gains, which may be acceptable or even desirable when the deployment context matches the training domain.

A comparison on the percentage change on all three datasets is given in Figure 1.

4.2 Zero-Shot Evaluation of Open LLMs

To understand how well recent open-source LLMs perform in the medical domain without task-specific fine-tuning, we evaluated several zero-shot models on the MS^2 dataset. Specifically, we selected four high-performing and resource-accessible models: LLaMA 3.2 3B, Mistral 7B, Gemma 7B, and OpenChat 7B. These models were chosen based on their availability, instruction-following capabilities, and competitive performance in prior evaluations on general summarisation tasks.

The evaluation was performed using the ROUGE metric suite to maintain consistency with our fine-tuned PRIMERA results. Table 3 presents the ROUGE-1, ROUGE-2, and ROUGE-L F1 scores obtained for each model.

These results demonstrate that while open LLMs exhibit a basic ability to generate summaries, their performance lags behind the fine-tuned PRIMERA model on this domain-specific dataset. For instance, the best-performing zero-shot model (LLaMA 3.2 3B) achieves only 18.73 in ROUGE-1, compared to 33.0 for fine-tuned PRIMERA. Notably, all zero-shot models struggled with ROUGE-2, indicating limited ability to capture fine-grained bi-gram-level details common in dense, technical summaries.

Table 3: Zero-shot ROUGE performance of open LLMs on MS^2.

Model	ROUGE-1	ROUGE-2	ROUGE-L
LLaMA 3.2 3B	18.73	3.07	10.97
Mistral 7B	18.40	3.56	11.80
OpenChat 3.5	17.80	3.43	11.20
Gemma 7B	15.60	2.90	9.85

These findings highlight the value of domain adaptation in specialised tasks such as medical summarisation, where zero-shot summarisation remains limited by lack of domain-specific knowledge and training alignment.

4.3 Qualitative Analysis in the Context of Medical Domain

In addition to showcasing the effectiveness of fine-tuned models for medical research summarisation, it is important to highlight the significant benefits that this model brings to the field of healthcare and medical research. Medical research papers are often dense and lengthy, containing a wealth of information that is crucial for healthcare professionals, researchers, and policymakers. However, the sheer volume and complexity of these papers can make it challenging to extract the key findings and insights efficiently, resulting in either a poor analysis of a

research paper or complete disregard of the publication due to the difficulty in summarising and coming to a conclusion on its findings. Moreover, focusing solely on the abstract and conclusion for the sake of efficiency may cause the reader to overlook important concepts as well. Hence, having a system in place to efficiently review the entire paper and generate a summary in a fraction of the time would be highly beneficial to any medical personnel.

Generated

The results of this meta-analysis suggest that surgical ablation of the left atrial fibrillation (LAF) is associated with a significantly lower incidence of recurrent atrial arrhythmias compared with catheter ablation. However, there was no significant difference in the incidence of major adverse events between the two groups. Surgical ablation appears to be a safe and effective treatment strategy for the treatment of LAF.

Ground Truth

Subgroup analysis demonstrated similar trends, with higher freedom from AF in the surgical ablation group for paroxysmal AF patients. The incidence of pacemaker implantation was higher, while no difference in stroke or cardiac tamponade was demonstrated for the surgical versus catheter ablation groups. Current evidence suggests that epicardial ablative strategies are associated with higher freedom from AF, higher pacemaker implantation rates and comparable neurological complications and cardiac tamponade incidence to catheter ablative treatment.

Figure 2: Comparison of Generated Summary from Our Model and Ground Truth.

Figure 2 showcases a comparison between the generated summary produced by our fine-tuned PRIMERA model and the corresponding ground truth summary. To aid the visualization, we have distinguished sentences in both summaries using different colors to demonstrate similarities and differences between them. To ensure a reliable evaluation of the summary's quality, considering the challenging task of human evaluation for all generated summaries, we conducted a sample analysis on a small subset of generated summaries and ground truth pairs using domain experts. In the figure, we have highlighted certain sentences in different colors to represent specific findings. Yellow - Express an agreement that surgical ablation yields better results than catheter ablation for AF patients, Green - Indicate an agreement that the occurrence of adverse events does not significantly differ between the two groups, namely surgical ablation and catheter ablation groups, Blue - Facts that are different between the two statements. The first is the usage of the term "LAF" which is medically inaccurate, and the second is epicardial ablation being mentioned in the ground truth in place of surgical ablation but entirely missing from the generated summary.

An analysis of the summaries generated by the proposed model revealed both advantages and disadvantages when compared to the ground truth. The generated statement fully aligned with two out of the three points of the ground truth, with only a few minor errors. Nevertheless, the word "epicardial ablation" has been used in place of "surgical ablation" in the ground truth, which is conceptually correct and would be easily understood by a healthcare professional who is familiar with the subject. However, deciphering this has been difficult for the algorithm given the complexity of the content. On the other hand, the term "Left Atrial Fibrillation" or "LAF" is used which is medically incorrect with regards to this research. Therefore it is evident that developing a flawless system for summarising complex content remains a significant challenge.

To further examine how the fine-tuned model compares to recent open LLMs, we performed a qualitative comparison using the LLaMA 3.2 3B model as an illustrative example. Although the LLaMA model generated syntactically fluent outputs, it frequently omitted critical findings and occasionally introduced factual inconsistencies. For instance, as shown in Table 4, summaries from LLaMA often presented overly generic conclusions or contradicted key results stated in the source documents.

As shown in Table 4, PRIMERA outputs closely aligned with the factual content and structure of the ground truth summaries, capturing nuanced medical relationships such as treatment hierarchies and combined efficacy. In contrast, LLaMA summaries either ignored critical qualifiers or introduced misleading claims; such as denying evidence where it actually exists. These issues are particularly concerning in clinical contexts where accuracy is non-negotiable.

This comparison highlights that while open LLMs can offer a strong baseline in zero-shot setups, they fall short in medical domain-specific summarisation without adaptation. Despite the resource-intensive nature of LLMs, our findings demonstrate that smaller, fine-tuned models like PRIMERA can deliver superior factual accuracy and domain relevance. These qualitative differences further support the value of targeted domain

Table 4: Sample summaries comparing ground truth, PRIMERA, and LLaMA 3.2 3B outputs.

Ground Truth	PRIMERA Output	LLaMA 3.2 3B Output
PDE5Is were significantly more effective than placebo or SSRIs for treating PE, while SSRIs were better than placebo. Combined treatment had better efficacy than PDE5Is alone.	PDE5Is were more effective than SSRIs or placebo. Combined treatments were more effective than PDE5Is alone.	There is no evidence to support the use of SSRIs for treating PE. Insufficient data to determine their effectiveness.
Silver-level evidence concerning the beneficial effects of mineral baths compared to no treatment. No clear effects for other balneological treatments were found.	Mineral baths showed beneficial effects over no treatment. Effects of other treatments were unclear.	Balneotherapy is effective for osteoarthritis of the knee in adults.
Oral cobalamin improves serum vitamin B12 and hematological parameters. Avoids discomfort and cost of injections. Supported for clinical use.	Oral cobalamin is effective in improving serum B12 and blood parameters and avoids injections.	Oral cobalamin is effective for vitamin B12 deficiency, but there is limited evidence supporting either oral or intramuscular use.

adaptation to ensure safe and reliable summarisation in high-stakes fields like healthcare.

4.4 Inter-annotator Agreement

To evaluate the subjectivity and consistency among human annotators, we used Krippendorff's Alpha (α) (Krippendorff, 1989), a robust statistical measure for inter-rater reliability that is well-suited to ordinal-scale annotations and missing data scenarios. Unlike Cohen's Kappa (Cohen, 1960), Krippendorff's Alpha supports multiple annotators and provides a more generalizable reliability estimate. Cohen's Kappa, while commonly used, assumes only two raters and is less robust to missing data, making Krippendorff's Alpha a more appropriate choice for our setting. Three medical experts independently evaluated 50 randomly selected samples—both ground truth and generated summaries—against the original documents. Each summary was scored on five key criteria: Relevance, Coherence, Coverage, Conciseness, and Accuracy—using a rubric co-developed with a medical expert (See Appendix A). The results, shown in Table 5, reveal consistently moderate inter-annotator agreement, with generated summaries demonstrating slightly higher agreement across all criteria.

Notably, criteria such as Coverage ($\alpha = 0.5541$) and Coherence ($\alpha = 0.5038$) exhibited the highest reliability for generated outputs, indicating that the model produces content that aligns more consistently with expert expectations. Interestingly, annotator agreement was higher for generated summaries than for ground truth in all five evaluation dimensions. This suggests that, while subjectivity remains inherent to human judgment, the generated summaries (PRIMERA) offer a more stable and interpretable baseline, especially when paired with a domain-aligned rubric.

Evaluation Criterion	Ground Truth	Generated Summary
Relevance	0.3012	0.4512
Coherence	0.4531	0.5038
Coverage	0.4215	0.5541
Conciseness	0.2536	0.3534
Accuracy	0.3824	0.4817

Table 5: Krippendorff's Alpha scores for ground truth and generated summaries across five evaluation criteria.

While Krippendorff's Alpha provides a robust measure of agreement, the relatively small sample size (sample of 50 summaries, compared to the full dataset of $20,000$ reviews) limits the strength of claims we can make about population-level properties. The observed patterns suggest interesting trends—particularly the higher agreement on generated summaries—but should be interpreted as preliminary findings that warrant validation on larger samples in future work. Statistical significance testing was not performed due to the small sample size and the ordinal nature of the data.

4.5 LLM-as-a-Judge Results

As introduced in the Evaluation Metrics section, we employed DeepEval's LLM-as-a-judge framework to compare 50 summaries generated by the fine-tuned PRIMERA model and four open-source LLMs—LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, and Gemma 7B—using Meta's LLaMA 3 90B Instruct model (`us.meta.llama3-2-90b-instruct-v1:0`) as the judge.

The comparative G-Eval scores for all models are summarised in Table 6.

The results indicate that PRIMERA, a domain-adapted model, consistently outperformed the

Table 6: Scores for PRIMERA and Open LLMs (Sample size: 50)

Metric	PRIMERA	LLaMA 3.2 3B	Mistral 7B	OpenChat 7B	Gemma 7B
Avg. Score	0.3452	0.2732	0.2850	0.2780	0.2650
Std. Deviation	0.1274	0.0788	0.0821	0.0773	0.0755
Median Score	0.3461	0.2655	0.2750	0.2700	0.2600
Max Score	0.6273	0.4434	0.4650	0.4520	0.4310
Min Score	0.1254	0.1309	0.1420	0.1355	0.1282

evaluated open-source LLMs across all metrics. Among the zero-shot baselines, Mistral 7B achieved the highest G-Eval scores, followed by OpenChat 7B, LLaMA 3.2 3B, and Gemma 7B.

Although the open LLMs demonstrated general summarisation capability in zero-shot settings, their performance declined on criteria such as factual consistency and coverage when applied to medical texts. These results suggest that, in specialised domains such as healthcare, domain-specific fine-tuning remains necessary to achieve reliable and contextually accurate summarisation.

5 Discussion

This study provides a systematic comparison of domain-adapted task-specific models versus general-purpose LLMs for medical MDS. While fine-tuning on domain-specific data is now standard practice, and improved performance on that domain is expected, the key question we addressed is whether such adaptation offers meaningful advantages over increasingly capable zero-shot LLMs, particularly in resource-constrained settings where deploying large models may be impractical.

Our comparative evaluation reveals that domain adaptation continues to provide substantial benefits for specialized summarisation tasks. The fine-tuned PRIMERA model consistently outperformed all evaluated open-source LLMs (LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, Gemma 7B) across automated metrics, human judgments, and LLM-as-a-judge assessments. Critically, this performance advantage came with significantly lower computational requirements—PRIMERA can run efficiently on modest hardware, while larger LLMs demand substantial resources even for inference.

To further examine generalisation after fine-tuning, we evaluated the model on out-of-domain datasets. On Multi-News, which predominantly consists of news articles, performance declined modestly relative to the pre-fine-tuned PRIMERA baseline (-6.9% to -14.0% across ROUGE metrics), consistent with mild catastrophic forgetting. By contrast, on the BigSurvey dataset of survey papers, the fine-tuned model showed clear gains ($+8.5\%$ to $+73.2\%$), indicating stronger adaptability within research-style domains that share structural similarity with biomedical papers.

Beyond ROUGE-based performance trends, we examined qualitative aspects of summarisation through human evaluation. Expert ratings offered complementary insights into coherence, factual accuracy, and domain relevance, providing a human-centred view of fine-tuning effects. Summaries from the fine-tuned PRIMERA model were rated higher across relevance, coherence, coverage, conciseness, and accuracy, and were generally more structured and focused than some human-written references, though minor factual inconsistencies remained. These findings motivated further reliability analysis using inter-annotator agreement and automated overlap metrics.

In this study, we evaluated the quality of machine-generated summaries for MDS in the medical domain using two key methodologies: inter-annotator agreement analysis via Krippendorff's Alpha (Krippendorff, 1989) and content overlap comparison using ROUGE metrics. Krippendorff's Alpha (α), chosen for its robustness to multiple raters and missing data, provided a more reliable estimate of human agreement than traditional methods like Cohen's Kappa. The aim was to assess both the consistency of expert annotations and the relative quality of generated summaries compared to human-written ones.

To complement traditional metrics, we also compared PRIMERA against several zero-shot open LLMs using LLM-as-a-judge evaluations. Despite reasonable performance by open models (e.g., Mistral in G-Eval), PRIMERA consistently outperformed them across all metrics. As shown in Table 6, the G-Eval scores further reinforce these findings. PRIMERA attained the highest average score (0.3452), outperforming all zero-shot baselines, with Mistral 7B (0.2850) and LLaMA 3.2 3B (0.2732) following. Scores closer to 1 indicate stronger factual consistency and coherence; thus, the ≈ 0.07 - 0.08 margin highlights a clear qualitative advantage. PRIMERA also showed slightly higher variance (SD = 0.1274), reflecting diverse yet consistently strong outputs.

This reinforces the conclusion that fine-tuned domain-specific models continue to offer critical advantages, especially in specialised, high-stakes domains such as biomedical summarisation. While

open LLMs offer flexibility and broad applicability, their outputs often lack the factual consistency and specificity demanded by domain-expert tasks. The qualitative error analysis also supported this finding—highlighting key factual inaccuracies and omissions in LLM-generated outputs.

Although large open LLMs are powerful, their resource requirements for inference and deployment can be prohibitive in many real-world settings. Pre-trained models like PRIMERA, when fine-tuned on task-specific datasets such as MS^2, demonstrate competitive performance with significantly lower computational demands. This trade-off highlights a practical benefit of domain-adapted summarisation models: they offer a scalable, cost-effective alternative while still maintaining high-quality performance.

Therefore, while larger open models may continue to improve, our findings suggest that fine-tuned pre-trained models remain a highly valuable and robust solution for domain-specific summarisation, particularly in scenarios where resource efficiency and reliability are paramount.

6 Conclusion

This study explored the domain adaptation of the PRIMERA model for MDS, with a focus on the biomedical domain using the MS^2 dataset. Through systematic fine-tuning, we demonstrated that the adapted model significantly outperforms the pre-trained baseline and competitive models such as LED, particularly in ROUGE metrics. Furthermore, our evaluations across multiple domains, including news and survey articles, indicate that the fine-tuned PRIMERA model retains strong generalisation capabilities beyond its training domain.

To provide a more nuanced perspective on summary quality, we incorporated LLM-as-a-judge evaluations using the DeepEval framework. These results further confirmed that the fine-tuned PRIMERA model surpasses leading open-source LLMs (e.g., LLaMA 3.2 3B, Mistral 7B) in terms of factual accuracy, coherence, and overall summarisation quality. While open LLMs showed promise in zero-shot settings, they underperformed in capturing domain-specific nuances critical for biomedical content.

Qualitative analyses echoed this gap, highlighting factual inconsistencies in zero-shot summaries, whereas PRIMERA more reliably retained core evidence and reasoning. Importantly, the resource efficiency of PRIMERA—compared to computationally intensive LLMs—positions it as a practical solution for real-world deployment in specialised domains.

Our findings demonstrate that domain-adapted task-specific models remain valuable for specialized summarisation, particularly when factual accuracy, computational efficiency, and deployment constraints are considerations. While the performance advantages of fine-tuning on domain data are expected, this study provides empirical evidence that such adaptation offers meaningful benefits over zero-shot LLMs in medical contexts—though at the cost of reduced generalization across domains.

Limitations

While this study provides convergent evidence across automated, human, and LLM-based evaluations, several limitations should be acknowledged. First, both the human and LLM-as-a-judge assessments were conducted on a small qualitative subset of 50 summaries (approximately 0.25% of the test set) due to computational constraints. Although this sample offers indicative comparative trends, it limits the statistical strength of our conclusions and precludes significance testing.

Second, open-source LLMs were evaluated only in zero-shot settings. Fine-tuning these models on the MS^2 dataset or employing few-shot prompting could substantially improve their performance, potentially narrowing the observed performance gap with PRIMERA.

Third, our evaluation was restricted to the biomedical domain. While cross-domain tests on news and survey papers provided preliminary evidence of generalisation, further investigation across other specialised domains—such as legal, financial, and technical documentation—is needed to assess the broader applicability of domain adaptation.

Future work should therefore incorporate larger-scale human and LLM-based evaluations, compare fine-tuned general-purpose LLMs against domain-specific models, and explore domain-aware accuracy metrics tailored for medical summarisation. Ultimately, the choice between fine-tuned task-specific models and general-purpose LLMs will depend on deployment context, balancing computational resources, domain sensitivity, and generalisation needs.

References

Azal Minshed Abid. 2022. Multi-Document Text Summarization Using Deep Belief Network.

Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri, and Grzegorz Chrupała. 2022. A Survey on Multi-document Summarization and Domain-Oriented Approaches. *Journal of Information Systems and Telecommunication (JIST)*, 1(37):68.

Emmanuel Andrès, Helen Fothergill, and Mustapha Mecili. 2010. Efficacy of oral cobalamin (vitamin b12) therapy. *Expert opinion on pharmacotherapy*, 11(2):249–256.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Jingqiang Chen, Chaoxiang Cai, Xiaorui Jiang, and Kejia Chen. 2022. Comparative graph-based summarization of scientific papers guided by comparative citations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5978–5988. International Committee on Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513. Association for Computational Linguistics.

Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2024. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084. Association for Computational Linguistics.

Kushan Hewapathirana, Nisansa De Silva, and C D Athuraliya. 2023. Multi-Document Summarization: A Comparative Evaluation. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 19–24. IEEE.

Kushan Hewapathirana, Nisansa de Silva, and C D Athuraliya. 2024. M2DS: Multilingual Dataset for Multi-document Summarisation. In *International Conference on Computational Collective Intelligence*, pages 219–231. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mahnaz Koupaee and William Yang Wang. 2018. WikiHow: A Large Scale Text Summarization Dataset. *arXiv preprint arXiv:1810.09305*.

Klaus Krippendorff. 1989. *Content analysis: an introduction to its methodology*. Sage Publications.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352. Association for Computational Linguistics.

Stefano Leon. 2020. Rotten tomatoes movies and critic reviews dataset. `https://bit.ly/RTdataset`. (Accessed on 06/24/2023).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES. In *International Conference on Learning Representations*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Generating a Structured Summary of Numerous Academic Papers: Dataset and Method. In *THE 31ST INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 4259–4265.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074. Association for Computational Linguistics.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2023. Multi-document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys*, 55(5):1–37.

Meta AI. 2024. Llama 3.2 model card. https://huggingface.co/meta-llama/Llama-3.2-1B. Accessed: 2025-08-16.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 21(140):1–67.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 19 others. 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. DIVERSE BEAM SEARCH: DECODING DIVERSE SOLUTIONS FROM NEURAL SEQUENCE MODELS. *arXiv preprint arXiv:1610.02424*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180. Association for Computational Linguistics.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761–5769. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263. Association for Computational Linguistics.

Benjamin Yu. 2022. Evaluating pre-trained language models on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 188–192. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. *Advances in neural information processing systems*, 33:17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

A Human Evaluation Rubric for Summary Assessment

This rubric was used by medical domain experts to evaluate both ground truth and model-generated summaries. Each summary was assessed independently across five criteria using a five-point Likert scale.

A.1 Evaluation Scale

Criterion-Specific Guidelines

Relevance Does the summary capture key clinical findings?

Score	Interpretation
1	Very Poor — fails completely in this aspect
2	Poor — significant issues are present
3	Fair — partially meets expectations, with flaws
4	Good — mostly meets expectations, minor issues
5	Excellent — fully meets expectations

Table 7: Likert scale used across all evaluation criteria.

1: Largely irrelevant or misleading.

2: Omits several key points.

3: Covers some relevant points; others diluted.

4: Includes most important findings.

5: Captures all clinically essential points.

Coherence Is the summary logically structured and easy to follow?

1: Confusing or disjointed.

2: Poor logical flow.

3: Basic structure with inconsistencies.

4: Mostly well-structured, minor issues.

5: Clear, fluent, and logically organized.

Coverage Does the summary reflect the breadth of the source?

1: Misses most critical content.

2: Narrow focus; limited scope.

3: Some key elements included.

4: Broadly representative.

5: Fully covers main findings.

Conciseness Is the summary free from redundancy or unnecessary detail?

1: Excessively verbose.

2: Frequent redundancy.

3: Some inefficiencies.

4: Mostly succinct.

5: Highly concise and focused.

Accuracy Are the statements factually correct?

1: Contains major factual errors.

2: Several inaccuracies.

3: Some vague or incorrect content.

4: Mostly accurate, minor issues.

5: Fully accurate and aligned with source.

Instructions to Annotators

- Read all source documents before scoring a summary.
- Evaluate each criterion independently.
- Use the rubric definitions to ensure consistency.
- When uncertain, assign the most justifiable score.

Speaking to the Inner Child: A Discourse Analysis of Healing Narratives in Digital Therapeutic Culture

Romeo Jr. E. Tejada[1], Jose Marie E. Ocdenaria[2] and Kenneth C. Lisbo[3]
[1,2]Ateneo de Davao University
[3]Mindanao State University - General Santos

Abstract

Inner-child healing has become a significant genre of self-help discourse on social media, where users share personal narratives of emotional recovery. While this narrative genre is growing, the linguistic strategies through which healing identities are constructed in these non-clinical, digital spaces remain underexplored. This study bridges this gap through a multi-method Discourse Analysis integrating corpus analysis, Systemic Functional Linguistics, and Narrative Analysis, applied to 100 Facebook and Reddit posts from 2024-2025. The analysis reveals that narrators consistently deploy agentive first-person positioning, mental, material, and relational processes, temporal framing, and redemptive narrative structures to frame healing as a moral, intentional, and transformative identity practice. The study argues that inner-child healing discourse performs identity work that aligns with broader therapeutic and participatory digital cultures such as commodification and reflective resistance. These findings contribute to understanding how everyday digital discourse shapes expressions of self and well-being online. They offer insights for scholars of discourse, digital and popular culture, and inclusive language in positive-psychology communities.

Keywords: Discourse Analysis, Healing, Inner-child, Trauma, Digital Discourse, Systemic Functional Linguistics (SFL), Narrative Analysis

1 Introduction

The development of digital spaces transformed how people discuss, process, and construct mental health narratives. Social media platforms become interactive spaces where users publicly share struggles, recovery experiences, and self-healing practices. This has given rise to a new genre of self-healing discourse (Hayvon, 2024). Among these discussions, inner-child healing has gained prominence as a therapeutic approach that encourages individuals to reconnect with their childhood selves to address emotional wounds and cultivate self-compassion (Bradshaw, 1990). Traditionally, inner-child healing has been explored within clinical psychology, self-help literature, and mindfulness-based interventions (MBIs). Its emergence in digital spaces has introduced a new dimension to healing discourse, where everyday users participate in therapeutic storytelling and engage in self-directed healing practices.

Users frequently share highly engaging healing narratives on platforms such as Facebook and Reddit. In the study of Gibbs and Franks (2002), cancer patients often incorporate metaphorical storytelling to frame their experiences. These narratives are expressions of personal healing and discursive acts that shape identity, construct meaning, and generate social validation. While the linguistic and narrative aspects of mental health discourse have been explored in clinical settings (Pennebaker, 2011; White & Epston, 1990), little research has examined how language is used to construct healing identities in non-clinical, digital environments.

In relation to that, social media influencers appeared as prominent actors in shaping public discourse, including in self-healing and wellness narratives. They nurture emotional connections with their followers, shaping their perceptions while navigating commercial interests (Li & Feng, 2022; Zhang & Mac, 2023). Although some of them provide valuable insights, concerns were raised regarding potential public opinion manipulation, particularly when commercial interests are not transparent (Goanta & Ranchordás, 2020). This issue is vital in sensitive areas like mental health and self-care, where some influencers may promote self-healing products, services, or lifestyle changes for profit rather than genuine well-being (Arriagada & Bishop, 2021). In 2024-2025, inner-child healing went viral in Philippine social media, sparking debates on its links to consumerism (Sarza, 2024).

The rise of social commerce has further blurred these lines, with influencers using live broadcasting and interactive engagement to market wellness-related products, thereby influencing how self-healing discourse is shaped and consumed (Alam et al., 2022). While this commercial aspect can provide an avenue and resources to individuals seeking support, it also advances ethical concerns about the exploitation of personal healing narratives for digital marketing strategies (Abidin, 2016).

2 Purpose of the Study

This study examines how inner-child healing is expressed, constructed, and framed in social media narratives. Specifically, it identifies the dominant linguistic features being used in these narratives; examines how narrators linguistically construct their healing identities; and analyzes how these narratives reflect broader cultural ideologies. This study provides insights into how digital platforms serve as spaces for emotional storytelling, healing, and community affirmation. The findings contribute to ongoing conversations on trauma discourse, therapeutic culture, and the role of language in shaping emotional and social recovery in digital spaces.

3 Theoretical Underpinnings

This study is anchored in a multi-theoretical framework integrating Systemic Functional Linguistics (SFL), Discourse Analysis (DA), Narrative theory (Labov & Waletzky, 1967), and Trauma Theory. The ideational and interpersonal, and metafunctions of SFL (Halliday & Matthiessen, 2014) provide the foundation for analyzing how language represents experience, performs identity, and organizes meaning to reflect ideology. In SFL, the recognition of the link between language and society has existed as early as 1964, where Halliday, McIntosh, and Strevens (1964) stated that language is "a form of activity of human beings in societies".

CDA (Fairclough, 2013) complements this by uncovering how healing discourse reproduces or resists dominant ideologies, and Trauma Theory (van der Kolk, 2014; Herman, 2015) emphasizes the long-term impact of traumatic experiences on physical and psychological well-being that can lead to difficulties in self-regulation, emotional processing, and relational capacity. Collectively, these frameworks present a comprehensive understanding of how narrators use language not only to articulate pain and healing but to engage in discursive healing that resonates with collective cultural ideologies.

4 Research Design

This study implemented a multi-methods approach to address the research questions. It integrates Corpus Linguistics for empirical pattern detection and Narrative and Thematic Analysis (Braun & Clarke, 2006) for identity construction. SFL provided the linguistic lens for identifying process types, and thematic structure, while CDA revealed the sociocultural implications behind seemingly personal healing discourse. This integration allows the study to balance empirical linguistic evidence with interpretative depth, making it suitable for exploring both how language is used and what it reveals about identity and healing in digital contexts.

5 Corpora and Data Collection

A corpus of 100 social media posts, along with their comments, was collected from Facebook and Reddit with a total of 20,112 tokens. Posts were selected using purposive sampling, based on the following inclusion criteria: (a) Posts must contain personal reflections on inner-child healing; (b) posts must be user-generated healing narratives; (c) posts must be published between the years 2024 and 2025. Promotional content, clinical discussions, and unrelated mental health posts were excluded to maintain focus on organic healing discourse. Keyword searches such as inner child healing, healing journey, and self-recovery were used to locate relevant data and were translated into English. All collected texts were compiled into a plain-text corpus format for analysis in AntConc, a freeware corpus analysis toolkit.

6 Data Analysis

The inner child healing narratives were analyzed in three main phases based on the study's research objectives. Phase one focuses on identifying common lexical patterns in the healing narratives through corpus linguistic analysis. Using AntConc, the study generated word frequency lists and collocation outputs to identify frequently used words and phrases and analyzed them through the lens of ideational metafunction. The second part involves SFL's interpersonal metafunction to examine how narrators position themselves as healing subjects and narrative analysis (Labov & Waletzky, 1967) to examine the structure of the texts. Together, these

phases offer a multi-layered perspective on how language mediates trauma, constructs identity, and enacts cultural values in digital healing discourse.

7 Trustworthiness

Multiple strategies are employed to ensure the reliability of this study. The research upholds anonymity and ethical considerations to protect the narrators' identities and ensure objective analysis. While public social media data may be accessible without consent, researchers still have ethical obligations to protect users' privacy and anonymity (Ford et al., 2021). No personally identifiable information was recorded, and usernames were omitted in reporting findings, avoiding selection and confirmation bias.

8 Results and Discussion

Research indicates that various forms of childhood trauma (e.g., abuse, neglect, and adverse life events) can lead to long-term effects on brain function and mental health (Cai et al., 2023; Hovens et al., 2010). These traumatic experiences can create an "inner child" with unresolved emotional wounds, fears, and maladaptive coping mechanisms into adulthood. People construct their identity through narratives in complex and multifaceted ways, drawing on cultural resources, personal experiences, and social contexts. Research shows that individuals use narratives to make sense of their lives, integrate past experiences with future aspirations, and position themselves within broader societal frameworks (McAdams & McLean, 2013; McLean & Syed, 2016).

8.1 Frequency and Collocations

The collocation analysis shows that inner-child healing narratives are highly self-referential and affective, dominated by first-person pronouns (I, my) paired with verbs of emotion and action (feel, want, buy, heal). In terms of agency, narrators present as both experiencers of emotion and agents of change, framing healing as an intentional, self-directed process. Frequent collocates such as safe, enough, happy, and past reveal a movement from vulnerability to empowerment, while inner and child anchor the discourse in the metaphor of emotional rebirth.

8.1.1 Dominant Linguistic Features

This section presents the dominant linguistic features used in inner-child healing narratives on social media, drawing on Systemic Functional Linguistics (SFL), particularly the ideational metafunction (Halliday & Matthiessen, 2014) and supported by Corpus Linguistics (Baker, 2006), the analysis investigates how narrators linguistically encode their experiences of trauma, emotional restoration, and identity formation. Ideational Metafunction represents experience. This is how people use language to talk about the world through actions, events, people, thoughts, and feelings. This involves examining how experiential meaning is realized through the transitivity system, which classifies processes into types (e.g., mental, material, relational), identifies participants in those processes (e.g., Actor, Senser, Carrier), and specifies the circumstances surrounding them (e.g., time, place, reason). The ideational metafunction uncovers what kind of experience is being presented, who is

Table 1: Collocates computed in AntConc

Word	Freq.	Collocates	Word	Freq.	Collocates
I	864	proud, happy, grateful, try, buy, afraid, learn, cry, learn, want, love	feel	63	safe, happy, enough, abandoned, emotional, exhausted, pressured, unsafe
You	395	Incredible, amazing, healing	want	41	to try, become, break cycle, explore, heal, start, buy, feel safe
my	386	inner child, mom, parents, siblings, partner, family, siblings, anxiety, kids, self, boyfriend, body, loved ones	buy	40	games, shoes, things, toys, myself
child	324	trauma, inner child, dreams, abandonment, buy, experience	old	37	toy, hobby, beliefs, wounds, self
inner	240	child, healing	back	31	then (historical contrastive structure)
love	70	Compassion, family, peace	past	31	Trauma, pain, self

Linguistic Feature	Description	SFL Elements	Narratives
High Use of First-Person Pronouns	Frequent use of I, me, and my positions the narrator as a central participant, emphasizing personal introspection.	Participant: Senser or Actor (in mental or material processes); Experiential Meaning	"I know for a fact that the little kid in me is so proud..." (Post 18) "I booked a solo trip to the aquarium". (Post 43)
Mental and Relational Processes	Used to express feelings, identity, and internal states; reflects emotional recognition.	Process Types: Mental (Senser, Phenomenon); Relational (Carrier, Attribute); Experiential Meaning	"I feel overwhelmed by how much I missed..." (Post 84) "I believe she deserved better. I deserved better." (Post 71)
Temporal Deixis and Contrastive Tense	Narratives are often structured through time markers (e.g., back then, now), highlighting transformation.	Circumstances of Time: Clause Complexes showing Logical Meaning	"Back then, my dream was simple... Now, I can buy..." (Post 4) "Now, I can buy the things I need and afford my own groceries". (Post 26)
Material Processes and Symbolic Reenactment	Healing actions (e.g., buying, giving) are encoded as material processes to represent agency and self-care.	Process Type: Material (Actor, Goal); Experiential Meaning	"Healing my inner child by buying the shoes..." (Post 40) "I gave them P1,000 each... a part of me is also healing." (Post 6)
Lexical Fields of Emotion, Family, Nostalgia	Use of emotionally charged and culturally specific words to evoke sensory memory and social belonging.	Mental Processes: Lexical Cohesion, Experiential Meaning	"I still remember how my toes scrunched up..." (Post 24) "As the breadwinner and eldest among five siblings, my happiness comes from surprising my family." (Post 22)

Table 2: Dominant Linguistic Features used in Inner-Child Healing Narratives

involved, and how it unfolds. From the corpus of 100 social media narratives, six dominant linguistic features emerged.

8.1.2 High Use of First-Person Pronouns

One of the dominant features across the corpus is the dominance of first-person pronouns (I, me, my), which reflect the intensely personal and self-referential nature of healing narratives. In SFL terms, the speaker frequently occupies the role of Senser (in mental processes) or Actor (in material processes), reinforcing discourse of individual agency and introspective labor. According to Eggins (2004), material processes are processes of "doing", usually concrete, tangible actions. On the other hand, mental processes involve what the senser think, feel, or perceive.

I	know for a fact	that the little kid in me is so proud...
Senser	Mental (Cognition)	Phenomenon (P18)

I	booked	a solo trip	to the aquarium (P43)
Actor	Material (Action)	Goal	Circumstance (Location)

These examples highlight how narrators position themselves not only as experiencers of pain but as reflective agents. In the first sentence, "know" is a mental process, and the speaker ("I") is the Senser, evidencing internal validation and emotional maturity. P18 stated that, "Some of us didn't get the chance to enjoy life when we were young, only when we became adults". The speaker recognized the experienced healing and affirms that her inner child will be so proud of how things changed with her experience.

The second example, on the other hand, is a material process type of transitivity. The actor in this context is checking their childhood goals as an adult when they visit Korea. The consistent use of first-person reference aligns with Pennebaker's (2011) claim that trauma narratives often involve heightened self-focus and introspection, serving as a vehicle for self-disclosure and identity negotiation. This pattern also reflects the individualistic framing of emotional healing in contemporary digital discourse, where one's healing journey is not only internal but performed for a witnessing audience.

8.1.3 Dominance of Mental and Relational Processes

Narratives are rich in mental processes (feel, remember, realize, wish) and relational processes (is, was, becomes), which represent emotional states, personal realizations, and identity definitions.

I	feel	overwhelmed by how much	I	missed	as a child (P84)
Senser	Mental (Affect)	Phenomenon	Actor	Material (Action)	Circumstance (Time)

I	believe	she	deserved	better. (P71)
Senser	Mental (Cognition)	Carrier	Relational (Attributive)	Attribute/ Phenomenon

The first example exemplifies a mental process. In Systemic Functional Linguistics (SFL), mental processes involve sensing. In this case, the verb feel functions as a mental process of affect, where the narrator (Senser) experiences an emotional state (Phenomenon) described as overwhelmed. The embedded clause "how much I missed as a child" further elaborates on the cause of this emotional state and contains a material process, with I as the Actor and missed as the process of "doing" (or in this case, not experiencing something in the past). The phrase "as a child" functions as a circumstance of time, situating the emotional experience temporally. Mental processes are highly prevalent in the narratives as the inner child is a mental concept

perceived by the narrators.

The second example combines a mental process with relational attributive processes to convey both cognitive realization and emotional affirmation. In SFL, relational process expresses states of being, identification, or possession, and it functions to link participants through meanings of classification, attribution, or equivalence (Halliday & Matthiessen, 2014). The first clause contains a mental process of cognition with "I" as the Senser and the projected clause "she deserved better" functioning as the Phenomenon. Within that projected clause, "she" serves as the Carrier, "deserved" as the relational process, and "better" as the Attribute, expressing a judgment about another person's worth. The second clause, "I deserved better," follows this structure.

Such process types are essential to experiential meaning because they allow speakers to articulate inner psychological states and relational self-perceptions, both of which are central to trauma processing and therapeutic reauthoring (White & Epston, 1990). The frequency of these clauses across the corpus affirms that healing is presented less as an external event and more as a felt, thought, and narrated process.

8.1.4 Temporal Deixis and Contrastive Tense Structures

From the narratives gathered, narrators often structure healing as a temporal journey, using deictic markers (e.g., back then, now, before, today) and tense shifts to frame their experiences as movement through time such as the examples in Post 4 and 26.

Back then	*my dream*	*was*	*simple*	*to be able to eat*	*at Jollibee*
Circum -stance: Time	*Carrier: Possessor*	*Relational (Process)*	*Attribute*	*Material (Embedded Process)*	*Circum -stance (Location)*

Now	I	can buy	the things	I need	and afford	my own groceries.
Circum -stance: Time	Actor	Material (Pro-cess)	Goal	Embedded Clause	Material (Process)	Goal

These examples illustrate how the narrator contrasts past deprivation with present empowerment, establishing logical meaning through a before-and-after narrative arc. In P4, the main clause is structured as a relational attributive process. "My" is the carrier or the Possessor, "was" functions as the relational process, and "simple dream" is the Attribute being assigned. The phrase "Back then" operates as a Circumstance of Time, situating the experience in the past. The adverbial infinitive clause functions as an embedded material process, expressing a desired but unrealized action, while "at Jollibee" specifies the location. Similarly, P26 presents a material transitivity process.

These types of transitivity processes emerged significantly in the narratives. "I" remains the central Actor or "My" describing possession, affirming agency gained over time. In SFL, the usage of Circumstance of Time following the contrastive tense structure highlights the transformation of the narrators. P4 narrates being able to eat at Jollibee at present, while P26 can now buy things and afford groceries. These highlight the temporal shift from past lack to present empowerment. The narrators use these transitivity structures to linguistically encode the healing journey through agency, time, and material progress. Such contrasts support McAdams and McLean's (2013) theory of narrative identity, where personal development is organized through time-based storytelling.

Moreover, the statements illustrate clause complexes to construct logical meaning. Clause complexes involve two or more clauses to express relationships such as cause-effect, compare-contrast, and past-future-present (Halliday & Matthiessen, 2014). These statements form a paratactic clause complex that expresses contrast between the present capabilities and past aspirations of the narrators. This exemplifies the growth of the Actors and Carriers over time of healing.

8.1.5 Material Processes for Symbolic Restorative Actions

Another dominant feature in the narratives of healing inner child is the use of material process verbs (buy, give, treat, recreate), often to describe symbolic or restorative actions aimed at the inner child.

Healing	my inner child	by buying	the shoes	I	wanted	in high school.
Material (Process)	Goal	Material (Process)	Goal	Senser	Mental (Cogni-tion)	Circum -stance: Time

I	gave	them	P1,000 each...	a part of me	is	also healing.
Actor	Material (Process)	Recipient	Goal	Carrier	Relational (Process)	Attribute / Material (Result)

These narratives above show how narrators actively repair past emotional wounds through material process. For example, P40 involves healing and buying with inner child and shoes as the Goals of the sentences. This presents a symbolic act of purchasing a long-desired item as an act of healing or emotional reparation. The embedded clause

introduces the desire of the senser in the past, reinforcing the nostalgic reparative framing of the action. In the same way, P6 provided a direct material process to the recipient.

These actions are not merely transactional; they are symbolic reenactments of unmet emotional or material needs. This resonates with Bourdieu's (1984) concept of symbolic capital, where acts of consumption or giving acquire emotional and social significance. In the context of inner-child healing, material acts linguistically perform emotional restoration, situating physical action as a substitute for emotional closure.

8.1.6 Lexical Fields of Emotion, Family, and Nostalgia

Finally, the analysis reveals recurring semantic domains, emotionally charged and culturally specific lexical fields, that ground the healing narrative in memory, sensory detail, and relational longing.

> I still remember how my toes scrunched up in those tight, white ukay-ukay shoes. (Post 24)

In P24, "remember" is a mental process, and the vivid sensory detail (scrunched, tight, ukay-ukay) enhances emotional realism. The choice of words reflects cultural specificity (e.g., Jollibee, Hello Kitty, ukay-ukay) that situates healing within shared Filipino socio-economic experiences. Lexical fields frequently include emotion (safe, happy, proud, overwhelmed), family (mom, dad, daughter, nephew) nostalgia (birthday, toys, cartoons, Jollibee). Lexical items related to emotion frequently appear in evaluative constructions and affirmations, such as:

> You are worthy of love and healing. (Post 16)

This statement uses relational process (is, are) to construct identity as evolving, affirmed, and emotionally whole. Words like worthy, enough, and deserve serve as linguistic tools of emotional correction, directly confronting long-held feelings of shame, neglect, or rejection. As seen in the KWIC analysis, the word love is a central affective term, used not only to soothe the self but to forge solidarity within the community. This aligns with Herman's (2015) view that affirmation and relational support are integral to trauma recovery, and with White and Epston's (1990) framework for "reauthoring" the self through emotionally charged language.

Family terms such as mom, dad, daughter, and child are often used to situate trauma within early relational dynamics. The family is frequently invoked not just as a source of care but also of pain and emotional neglect. Narratives like:

> My early childhood felt fuzzy, lonely, and distant. My parents were just trying to survive. (Post 61)

> As a child, my mom would give away my favorite toys without asking me. (Post 97)

These reflect the redemptive logic of reparenting, where healing is performed through caregiving to others (especially one's own children) or symbolically to oneself. The frequent appearance of these terms underscores how intergenerational memory is woven into healing discourse, supporting the idea that healing is as much about relational revision as it is about individual repair.

The nostalgic lexicon, including birthday, toys, cartoons, Jollibee, and Hello Kitty, serves as a portal to lost or unrealized childhood joy. These references frequently appear in material process clauses where narrators perform symbolic reenactments of their unmet needs:

> Solo trip to the aquarium-healing my inner child with fishy vibes. (Post 43)

> I wasn't able to fully enjoy my childhood, but now, even something as simple as buying a Hello Kitty item brings joy to my inner child (Post 38)

These examples demonstrate what Boym (2001) calls reflective nostalgia, not an attempt to recreate the past as it was, but to grieve and reinterpret it through symbolic action.

The findings on dominant linguistic features reveal that inner-child healing narratives on social media are shaped by consistent patterns. Using the framework of Systemic Functional Linguistics (SFL), particularly the ideational metafunction and transitivity system, the study found that narrators primarily employ first-person pronouns (e.g., I, my) alongside mental processes (feel, remember), relational processes (is, was), and material processes

Identity-Building Strategy	Description	Linguistic Feature	Narratives
Narrative Structure and the Healing Arc	Narratives follow Labovian structure (orientation to coda) to reframe pain into growth.	Chronological sequencing, use of past and present tense, evaluative coda	"My healing journey started 6 years ago... I now send love to every past version of me." (Post 13)
Pronoun Shifts and Solidarity Building	Shifts from I to you to we to express inclusion and build collective identity.	Pronoun variation; audience address	"I'm so proud of the work we're doing, and together, we will heal" (Post 16)
Healing as Mentorship	Narrators position themselves as guides or mentors. Healing identity evolves from personal recovery to a source of help and influence for others, reflecting empowerment and purpose.	Directive language, community outreach phrases	"Click the link in my bio to book a Soul Contract Reading, or message me today!" (Post 27) "If you want to heal your inner child with me, reach out!" (Post 19)

Table 3: Healing Identity Strategy used in Inner-Child Healing Narratives

(buy, give) to frame healing as both an inner experience and a set of symbolic actions. Temporal deixis (now, back then) adds narrative coherence by charting change over time. Corpus analysis highlights frequent use of emotionally rich and culturally specific lexis (love, safe, trauma, mom, Jollibee), which contribute to lexical cohesion and situate personal healing within recognizable emotional and social contexts.

This section explores how narrators use language to construct their healing identities in inner-child narratives on social media, drawing from the interpersonal metafunction of SFL (Halliday & Matthiessen, 2014) and narrative theory (Labov & Waletzky, 1967; De Fina & Georgakopoulou, 2008). In SFL, the interpersonal metafunction focuses on how language is used to negotiate social roles, express attitudes, and manage interpersonal relationships. This includes systems of mood (statements, questions, commands), modality (possibility, obligation, certainty), and evaluation (judgment, appreciation, affect). These features allow narrators to calibrate their emotional stance, claim credibility, and establish solidarity. Complementing this, narrative theory emphasizes sequencing, framing, and audience interaction as mechanisms through which identity is constructed, particularly in contexts where storytelling functions both as personal expression and social engagement.

8.1.7 Narrative Structure and the Healing Arc

Many narrators adopt a redemptive storytelling arc aligned with the Labovian narrative structure-Orientation, Complication, Evaluation, Resolution, and Coda. This form allows speakers to frame their trauma and healing as part of a chronologically coherent and emotionally purposeful journey.

- Orientation: "My healing journey started 6 years ago..."
- Complication: "I kept attracting emotionally unavailable men."
- Evaluation: "I thought if I were better, I'd be good enough."
- Resolution: "I discovered and healed my inner child abandonment wound."
- Coda: "I now send love to every past version of me."

Such structure from Post 13 allows the narrator to claim insight and transformation, positioning themselves as emotionally mature and self-aware. Similarly, Post 20 frames healing through contrastive time expressions:

> Before stepping into my inner child healing journey, I was trapped in self-doubt... I set healthy boundaries and communicated openly."

These temporal contrasts align with McAdams and McLean's (2013) notion of narrative identity, wherein the past is reinterpreted to validate present growth. However, scholars like Frank (1995) and Smith & Sparkes (2008) remind narrators that trauma is often nonlinear and resistant to closure. Even so, in social media contexts, the expectation for hopeful narratives often compels speakers to perform recovery in ways that are inspirational, coherent, and socially desirable.

8.1.8 Pronoun Shifts, Mentorship, and Solidarity Building

Narrators often shift from I to you to we, creating a discursive bridge between self-reflection and community engagement. In Post 16, the narrator stated, "I'm so proud of the work we're doing, and together, we will heal." This signifies a change of perspective to address a wider audience or the community, which is a common theme in social media accounts. This is further exhibited by Post 23, where he acknowledged the "childish" healing by adults. From his personal perspective, he shifted into a collective stance where he stated, "Who are we to make fun of

them?" This shift not only broadens the narrative's reach but positions the speaker as empathetic, inclusive, and emotionally evolved. According to De Fina and Georgakopoulou (2008), such shifts are indicative of narrators who are aware of their audience and actively perform alignment, transforming the individual journey into a collective one.

Some narrators move beyond personal storytelling to adopt the voice of a mentor, coach, or entrepreneur, turning their pain into a public offering as exhibited in Posts 27, 19, and 30. This shift is supported by Frank's (1995) "wounded healer" figure and reflects how emotional authenticity becomes a source of authority. At the same time, it raises concerns identified by Baker and Greenhill (2021) about the commercialization of healing, where personal growth is stylized into a marketable identity. The use of imperative mood and promotional tone signifies that healing, once private, is now branded and broadcast for audience consumption.

The analysis of how narrators construct their healing identities presented patterns of interpersonal linguistic strategies and narrative structures. Drawing from interpersonal metafunction, narrators employ systems of mood, modality, evaluation, and pronoun shifts. Pronoun shifts from I to you to we, and the use of directive language further empowers narrators to transition from personal testimony toward collective perspective, mentorship, and marketable healing personas.

8.2 Ideologies Emerging from the Narratives

8.2.1 Healing through Play and Nostalgia as Reflective Resistance

Narratives such as "Watching cartoons in my 30s" (P76) and "Back then, my dream was Jollibee with my parents" (P4) highlight how speakers use nostalgia and play to reclaim joy. These posts often begin with circumstances of time or setting (e.g., "Solo trip to the aquarium"), suggesting that emotional reconnection is staged as ritualized action. However, Boym (2001) differentiates between restorative nostalgia, which seeks to reconstruct the past, and reflective nostalgia, a form of longing that dwells in the pain and longing itself, rather than seeking a return to a past that is impossible to recreate. While nostalgia provides validation, it must be balanced with present-day emotional work to avoid sentimentality replacing genuine healing.

These narratives challenge cultural ideals of rational adulthood and reclaim childlike pleasure as a valid healing tool. The textual metafunction emphasizes setting and agency, making symbolic gestures (e.g., eating at Jollibee) central to the healing narrative. Humor and whimsy also function as protective discursive devices, softening emotional disclosure and making trauma more socially acceptable (Goffman, 1959).

8.2.2 Consumerism in Healing and the Commodification of Trauma

Social media influencers have emerged as powerful forces in shaping consumer behavior and public discourse. Their ability to affect purchase behavior and raise awareness on various topics stems from the emotional bonds and perceived authenticity they cultivate with their followers (Li & Feng, 2022; Zhang & Mac, 2023).

Posts such as "I bought the shoes I wanted in high school" (P32) use material processes (buy, own, treat) to present healing as acquirable through consumption. CDA exposes this as a byproduct of capitalist wellness discourse, where emotional needs are addressed through retail therapy and self-gifting. These actions become emotionally loaded forms of symbolic capital (Bourdieu, 1984), where ownership equals recovery.

Some narrators also express discomfort with this trend, asking, "Why is it that when someone says they're 'healing their inner child,' it has to involve consumerism?" This internal contradiction points to the ethical dilemma in commodifying healing. While material purchases can offer a sense of empowerment and self-worth, they also raise questions about accessibility and the commodification of trauma. Hooks (2000) argues that when self-help and therapeutic practices become commodified, healing becomes less about internal transformation and more about participating in consumer rituals. This critique underscores the tension between consumer-driven healing and authentic, internal processes of recovery.

In consumer-driven societies according to psychology, inner child healing is often commodified, with products and services marketed as quick fixes. While acquiring childhood desires may offer nostalgia or brief fulfillment, material possessions cannot address deeper wounds. Such commercialization risks reducing healing to a transaction, distracting from the introspection and emotional work necessary for genuine recovery (Feinstein, 2023; Morris & Barrera, 2024).

9 Conclusion and Implications

The findings of this study reveal that inner-child healing narratives on social media are shaped by patterned linguistic strategies and broader cultural ideologies that collectively construct healing identities as emotionally expressive, morally responsible, and socially visible. Drawing from Systemic Functional Linguistics, the corpus analysis shows that narrators frequently use first-person pronouns, mental and material processes, temporal deixis, and evaluative language to frame healing as a deeply personal yet symbolically performative journey. These narratives follow a redemptive arc, often contrasting past deprivation with present agency, and use modality to project varying degrees of certainty, vulnerability, and empowerment. Pronoun shifts and imperative moods signal the transformation of personal healing into mentorship and solidarity, while lexical choices rooted in emotion, nostalgia, and familial memory reinforce therapeutic coherence and relational alignment. Thus, inner-child healing narratives are not merely introspective expressions but socially and ideologically shaped emotional storytelling. Narrators reclaim agency while also negotiating pressures of performative authenticity, commodification, and social conformity. These linguistic patterns, narrative strategies, and ideological framings highlight how language is used in inner child healing narratives, building a ground for discussing broader complex therapeutic culture and consumerism in digital spaces.

References

Abidin, C. (2016). "Aren't these just young, rich women doing vain things online?": Influencer selfies as subversive frivolity. Social Media + Society, 2(2), 2056305116641342. https://doi.org/10.1177/2056305116641342

Abidin, C. (2016). Visibility labour: Engaging with Influencers' fashion brands and #OOTD advertorial campaigns on Instagram. Media International Australia, 161(1), 86–100. https://doi.org/10.1177/1329878x16665177

Alam, F., Lahuerta-Otero, E., Tao, M. & Feifei, Z. (2022). Let's Buy With Social Commerce Platforms Through Social Media Influencers: An Indian Consumer Perspective. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.853168

Arriagada, A. & Bishop, S. (2021). Between Commerciality and Authenticity: The Imaginary of Social Media Influencers in the Platform Economy. Communication, Culture and Critique, 14(4), 568–586. https://doi.org/10.1093/ccc/tcab050

Baker, P. (2006) Using Corpora in Discourse Analysis. London: Continuum.

Baker, S. A., & Greenhill, A. (2021). The commodification of self-help: Wellness influencers and the problem of authenticity. Journal of Consumer Culture, 21(2), 314–331. https://doi.org/10.1177/1469540519826305

Bourdieu, P. (1984). Distinction: A social critique of the judgement of taste. Harvard University Press.

Boym, S. (2001). The future of nostalgia. Basic Books. https://romancesphere.fas.harvard.edu/news/future-nostalgia

Bradshaw, J. (1990). Homecoming: Reclaiming and Championing Your Inner Child. Bantam Books. https://www.johnbradshaw.com/books/homecoming-reclaiming-and-healing-your-inner-child

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3, 77-101.

Cai, J., Zhao, Y., Li, J., Zhang, J., Liu, D., Liu, Q. & Gao, S. (2023). Long-term effects of childhood trauma subtypes on adult brain function. Brain and Behavior, 13(5). https://doi.org/10.1002/brb3.2981

De Fina, A. & Georgakopoulou, A. (2008). Analysing narratives as practices. Qualitative Research, 8(3), 379–387. https://doi.org/10.1177/1468794106093634

Derewianka, B. (2022). A new grammar companion. PETAA. ISBN 978-1-925132-69-4

Eggins, S. (2004). An Introduction to Systemic Functional Linguistics (2nd ed.), Continuum International Publishing Group.

Fairclough, N. (2013). Critical Discourse Analysis: The Critical Study of Language. Routledge.

Ford, E., Shepherd, S., Jones, K. & Hassan, L. (2021). Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. Frontiers in Digital Health, 2. https://doi.org/10.3389/fd

gth.2020.592237

Frank, A. W. (1995). The wounded storyteller: Body, illness, and ethics. University of Chicago Press.

Feinstein, D. (2023). Using energy psychology to remediate emotional wounds rooted in childhood trauma: preliminary clinical guidelines. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1277555

Gibbs, R. W., & Franks, H. (2002). Embodied Metaphor In Women's Narratives About Their Experiences With Cancer. Health Communication, 14(2), 139–165. https://doi.org/10.1207/S15327027HC1402_1

Goanta, C. & Ranchordás, S. (2020). The regulation of social media influencers: an introduction (pp. 1–20). edward elgar. https://doi.org/10.4337/9781788978286.00008

Halliday, M. A. K. (1964). Syntax and the consumer. In C. I. J. M. Stuart (Ed.), Report of the Fifteenth Annual (First International) Round Table Meeting on Linguistics and Language (pp. 11-24). Washington, DC: Georgetown University Press.

Halliday, M. A. K., & Matthiessen, C. M. (2014). Halliday's introduction to functional grammar. Routledge.

Halliday. M. A. K., A. McIntosh & P. Strevens (1964). Linguistic Sciences and Language Teaching. London: Longmans.

Hayvon, J. C. (2024). Digital Media to Support Healing from Trauma: A Conceptual Framework Based on Mindfulness. Issues in Mental Health Nursing, 45(12), 1258–1267. https://doi.org/10.1080/01612840.2024.2398649

Herman, J. (2015). Trauma and recovery: The aftermath of violence—from domestic abuse to political terror. Basic Books/Hachette Book Group.

Hochschild, A.R. (1983). The managed heart: Commercialization of human feeling. Berkeley, CA: University of California Press. https://doi.org/10.1002/pam.4050030365

Hooks, B. (2000). Where we stand: Class matters. Routledge. https://doi.org/10.4324/9780203905104

Hovens, J. G. F. M., Spinhoven, P., Zitman, F. G., Wiersma, J. E., Penninx, B. W. J. H., Giltay, E. J. & Van Oppen, P. (2010). Childhood life events and childhood trauma in adult patients with depressive, anxiety and comorbid disorders vs. controls. Acta Psychiatrica Scandinavica, 122(1), 66–74. https://doi.org/10.1111/j.1600-0447.2009.01491.x

Labov, W. and Waletsky, J. (1967) Narrative Analysis: Oral Versions of Personal Experience. In: Helm, J., Ed., Essays on the Verbal and the Visual Arts, University of Washington Press, Seattle and London, 3-38. https://doi.org/10.1075/jnlh.7.02nar

Li, X. (Leah) & Feng, J. (2022). Influenced or to be influenced: Engaging social media influencers in nation branding through the lens of authenticity. Global Media and China, 7(2), 219–240. https://doi.org/10.1177/20594364221094668

Mcadams, D. P. & Mclean, K. C. (2013). Narrative Identity. Current Directions in Psychological Science, 22(3), 233–238. https://doi.org/10.1177/0963721413475622

Mclean, K. C. & Syed, M. (2016). Personal, Master, and Alternative Narratives: An Integrative Framework for Understanding Identity Development in Context. Human Development, 58(6), 318–349. https://doi.org/10.1159/000445817

Morris, S. Y., & Barrera, A. Z. (2024). A decolonized mental health framework for black women and birthing people. Journal of Lesbian Studies, 28(4), 642–655. https://doi.org/10.1080/10894160.2024.2356994

Pennebaker, J. W. (2011). The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press. https://doi.org/10.1016/S0262-4079(11)62167-2

Sarza, R. (2024). Healing your inner child. Philippine Daily Inquirer. https://opinion.inquirer.net/178935/healing-your-inner-child

Smith, B., & Sparkes, A. C. (2008). Narrative and its potential contribution to disability studies. Disability & Society, 23(1), 17–28. https://doi.org/10.1080/09687590701725542

Dual-mode N-gram Similarity Detection for Forensic Authorship Analysis

John Blake
University of Aizu
Aizuwakamatsu
Japan
jblake@u-aizu.ac.jp

Kazuma Tamura
University of Aizu
Aizuwakamatsu
Japan
m5281051@u-aizu.ac.jp

Krzysztof Kredens
Aston University
Birmingham
United Kingdom
k.j.kredens@aston.ac.uk

Abstract

This paper introduces a dual-mode n-gram similarity detection tool specifically designed for corpus-based forensic authorship analysis. Intra-corpus mode is used to verify consistency within a dataset while inter-corpus mode is for comparison to a questioned dataset. Preliminary accuracy evaluation of shared n-gram detection is perfect at 100%.

1 Introduction

1.1 Background

Authorship analysis involves the attribution or exclusion of authorship based on linguistic evidence (Grant, 2013), a task that relies heavily on identifying patterns of lexical similarity and dissimilarity across documents. Forensic authorship analysis supports criminal and civil investigations by providing evidence-based attribution of anonymous or disputed texts (Coulthard and Johnson, 2000).

The datasets in forensic contexts tend to be rather small (Carter, 2022), frequently focussing on discovering the authorship of one text by comparing with known texts written by candidate authors. Linguistic similarity between questioned and known documents can offer probative value, especially when reinforced by recurrent and relatively rare (i.e. distinctive) lexical or syntactic patterns. Current approaches, however, rely on corpus query tools such as AntConc (Anthony, 2024), WordSmith Tools (Scott, 2008) and LancsBox (Brezina, 2025), which were developed for other purposes.

In a typical forensic authorship analysis workflow, linguists read and annotate texts to identify potentially distinctive n-grams (Wright, 2017). These n-grams may be examined in context using the keyword-in-context (KWIC) display in a corpus query tool to determine whether their usage is habitual or anomalous (Johnson and Wright, 2014).

When working with multiple documents assigned to either questioned (Q) or known (K) categories, it is essential to compare the n-gram distribution within each category. This intra-group or intra-corpus analysis helps assess authorial stylistic consistency (Cardaioli et al., 2021; Zhu and Jurgens, 2021). Subsequently, the analysis moves to the comparison of inter-group or inter-corpus shared n-grams, where distinctive n-grams in the Q text are checked for overlap in the K dataset, thus helping to establish or exclude authorship (Nini, 2018).

1.2 Problem

Current corpus tools are designed to prioritise frequency-based analysis of large datasets; they are not optimised for saliency (Boswijk and Coler, 2020) nor do they focus on fine-grained analysis of small datasets comprising one or more short texts. Yet, saliency and nuanced analysis of small datasets are of the utmost importance in forensic investigations, where the focus is on identifying potentially distinctive expressions that may distinguish and disambiguate authorship.

Forensic authorship analysts face two main difficulties. First, one way to confirm stylistic consistency within the texts attributed to one author is to discover how many distinctive n-grams are shared between the texts. To do so, the n-grams need to be identified, counted, ranked by frequency and compared among all the texts.

Second, once the distinctive n-grams have been identified, the distinctive n-grams that occur in both Q and K texts need to be compared, which involves identifying, counting, and ranking them by frequency of shared n-grams.

Presently, neither the consistency nor the comparison functionalities are directly available in any corpus tool. Thus, there is a niche that needs to be addressed to improve the workflow for forensic linguists adopting a shared n-gram approach.

1.3 Research Objectives

Our primary motivation is to bridge this gap by designing a user-friendly tool that consolidates these functions into a single platform. The two primary objectives of this project are as follows:

1. to integrate a consistency-checking function that enables users to assess whether documents attributed to a single author exhibit internal stylistic coherence; and

2. to facilitate comparison of shared n-grams between questioned texts and several candidate author datasets.

Together, these objectives form the foundation of the creation of a practical n-gram similarity tool for forensic authorship analysis.

1.4 Contribution

We present a dual-mode similarity detection function integrated into a web application. Unlike other corpus software, our tool supports cross-corpus alignment through an intuitive interface, enabling forensic linguists to:

- identify unigrams, bigrams, and trigrams occurring in both Q and K texts;

- assess stylistic consistency within a set of texts attributed to a single author; and

- compare shared n-grams across and between datasets.

By combining these capabilities, the tool offers a purpose-built solution for detecting potentially distinctive n-grams as indicators of authorship.

2 Related Work

2.1 Authorship Attribution and Verification

Authorship analysis has historically relied on stylometric features such as function word frequency, character n-grams, and syntactic structure (Ding et al., 2017; Klaussner et al., 2015; Lagutina et al., 2019). Statistical techniques include Burrows' Delta (Evert et al., 2017), support vector machines (Diederich et al., 2003), and nearest-neighbour classifiers (Cunningham and Delany, 2021). These methods have been used in authorship attribution in both literary contexts and forensic investigations.

These techniques involve a degree of familiarity with programming, which may range from simply adapting or running an existing program to creating a tailor-made solution. Added to this technical hurdle, in forensic contexts the use of sophisticated technologies should (or must in some jurisdictions) meet the evidentiary standards set out in the Daubert criteria (DeMatteo et al., 2019), which require that the methodology be scientifically valid, reliably applied, and open to scrutiny.

Explaining statistical models, mathematical reasoning, and computational procedures to a lay audience such as a jury presents a considerable challenge: the concepts are often abstract, highly technical, and removed from everyday experience (Coulthard, 2005). This complexity creates opportunities for opposing legal representatives to question, oversimplify, or misrepresent the underlying methods, potentially undermining the credibility of the expert witness testimony (Brodsky et al., 2012; O'Brien and O'Brien, 2017).

However, despite these technical advances, many forensic linguists still rely on workflows that combine multiple tools and require substantial manual intervention to extract, interpret, and triangulate stylistic patterns. Existing tools may be categorised as software libraries or ready-to-use tools.

2.2 Authorship Analysis Programs

Programs include Signature stylometric system[1], the well-respected stylometric R package Stylo (Eder et al., 2024) and the recently released R package Idiolect (Nini, 2024).

Signature provides a simple interface that is suitable for educational rather than forensic use. Stylo is designed for literary investigations of authorship rather than for forensic contexts and so relies on a stylometric approach. Idiolect builds on Nini's approach to linguistic individuality (Nini, 2023) and draws on the Likelihood Ratio Framework (Ishihara, 2021).

A commercial product NeoNeuro[2] offers authorship analysis via its rather dated proprietary program. There is no available details regarding its algorithm and effectiveness. NeoNeuro identifies 4-grams occurring in each of the K texts and compares them to the Q text. The output generated lists of the shared n-grams and provides percentage similiary score for each K text.

The Java Graphical Authorship Attribution Pro-

[1] https://www.philocomp.net/texts/signature.htm

[2] https://neoneuro.com/products/authorship-attribution

gram (JGAAP)[3] is straightforward to use because of its menu-driven user interface. This tool, however, needs to be set up, which may be difficult for those unfamiliar with GitHub.

Both Stylo and Idiolect require knowledge of R scripting, which is an onerous barrier for users with no programming experience. To set up JGAAP some programming knowledge is required although users can operate it without the need for any scripting. JGAAP, however, is no longer actively maintained, limiting its suitability for forensic work. Both Signature and NeoNeuro are simple to use, but very limited in terms of functionality.

General-purpose corpus analysis tools such as AntConc (Anthony, 2024), WordSmith Tools (Scott, 2008), LancsBox (Brezina, 2025), and Sketch Engine (Kilgarriff et al., 2014) were designed for linguistic research and teaching, not forensic investigation. AntConc excels at KWIC concordancing and keyword analysis but lacks built-in facilities for dataset comparison or intra-author stylistic consistency checks. Sketch Engine provides advanced collocational and profiling functions, but its emphasis on large-scale corpora does not align with the small, sensitive datasets typical of forensic work.

In practice, forensic linguists often adapt these tools through ad hoc workflows, combining outputs from multiple searches and using external software such as Excel or SPSS for comparison. This process is time-consuming, prone to error, and difficult to reproduce.

The absence of integrated cross-corpus and within-corpus comparison functionality in these tools means practitioners must either combine multiple outputs or write custom scripts to meet their needs.

2.3 Gap in Existing Approaches

The reviewed tools illustrate a clear gap: there is no single, forensic-oriented platform that integrates corpus management, exploration, and both intra- and inter-corpus similarity detection in a user-friendly environment. Our prototype addresses this gap by enabling forensic analysts to perform these tasks without programming skills, with an emphasis on clarity, reproducibility, and visual traceability.

3 System Overview

Figure 1 shows the system architecture of the corpus tool, focusing on the dual-mode similarity detection functionality. The system includes a file management module that enables users to perform standard Create, Read, Update and Delete (CRUD) operations, an analysis engine that includes the Similarity detection and comparison functionality, and a graphical user interface (GUI) to visualizer the results. The graphical user interface is divided into clearly labelled, colour-coded tabs, such as Manage, Search, Compare, Consistency, and Admin. Each tab corresponds to a major workflow step, allowing users to switch seamlessly between data upload, search configuration, cross-corpus comparison, intra-author consistency checks, and administrative controls. This modular structure supports intuitive navigation and ensures that forensic analysts can focus on linguistic patterns without being overwhelmed by interface complexity.

The system is implemented using a Django backend for robust server-side logic and data management, combined with a React-based single-page interface to ensure a responsive and intuitive user experience. Preprocessing of textual data is handled through the Natural Language Toolkit (NLTK) (Bird et al., 2009) in the development version, enabling tokenisation, normalisation, and n-gram extraction. The production release will harness SpaCy (Neumann et al., 2019), given its substantially faster processing speed, efficient memory management, and optimised pipeline for large-scale text handling.

Data exchange between the server and client is managed via JSON endpoints, ensuring efficient and lightweight communication. Corpora are stored as structured file sets with rich accompanying metadata, allowing for precise filtering and contextual retrieval. An indexing mechanism accelerates search and lookup operations, facilitating near-instantaneous access to relevant textual segments. The user interface organises functionality into clearly labelled tabs for corpus upload, keyword and pattern search, and similarity analysis. Within the similarity module, results are visually enhanced using colour-coded recurrence bars that indicate, at a glance, the number of documents in which a given n-gram occurs. This design balances technical performance with usability, allowing users to navigate between analytical functions

[3] https://github.com/evllabs/JGAAP

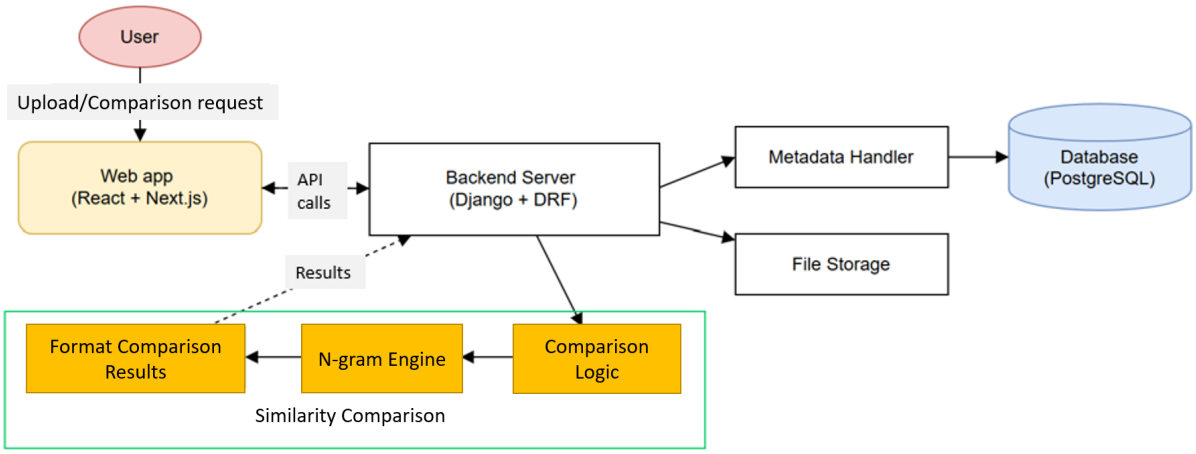


Figure 1: System architecture for similarity detection and comparison

without disrupting workflow.

4 Case Study: Threatening letters

A subset of the Threatening English Language (TEL) Corpus (Gales et al., 2023). was selected to show how the intra-corpus consistency and inter-corpus comparison functions can be harnessed with real-world forensic datasets.

To use the system, plain text files first need to be uploaded. Each file, regardless of its size or content, is treated as a single document. Metadata such as author ID and description can be added, and files are stored with secure identifiers to maintain confidentiality. Tokenization, part-of-speech tagging and basic preprocessing (e.g., lowercasing, punctuation removal) are automatically applied.

The KWIC interface allows users to search for any word, phrase, or regex pattern, retrieving and aligning all instances across the selected corpora as shown in Figure 2. Searches can also be performed using parts-of-speech tags.

Selecting the Compare tab provides access to the intra-corpus consistency and inter-corpus comparison features, both powered by the n-gram engine, which extracts unigrams, bigrams, and trigrams. The system calculates occurrence frequency and distribution across all selected documents, with results used in two modes.

The consistency checker (intra-corpus comparison) sorts results by the number of files in which shared n-grams occur. For any selected author corpus, n-gram recurrence is calculated across all documents, ranked by the number of matching files, and displayed in descending order from the most frequent n-grams (See Figure 3).

The cross-corpus comparison (inter-corpus comparison) sorts results by n-gram frequency in the Q corpus. One document or corpus is designated as Questioned (Q), and the system compares it against any number of Known (K) corpora, identifying overlapping n-grams. These are ranked by frequency in Q, with their frequency in each K corpus highlighted. Figure 4 shows the results of inter-corpus comparison of trigrams on the Compare tab using Searle 005 as the Questioned dataset (Q) and comparing that to two other known datasets, namely Searle 001 and Hickley 001. The first column gives the trigram in order of frequency in Q. The background of shared trigrams occurring in the other datasets is colourized. The raw count and the percentage of each trigram are also given.

5 Evaluation

The system was evaluated using multiple corpora: the Enron Email Corpus (Hussain, 2020), the Blog Authorship Corpus[4], the 100 Idiolects Project[5], and the Threatening English Language (TEL) Corpus (Gales et al., 2023).

In all cases, unigram, bigram, and trigram extraction ran successfully, and the comparison logic produced correct results. The system achieved 100% accuracy in counting, ranking, and comparing n-gram similarities both within and between corpora.

6 Discussion

The tool is transparent, easy to interpret, and requires no programming skills. Its highly visual,

[4] https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus/data

[5] https://fold.aston.ac.uk/handle/123456789/17

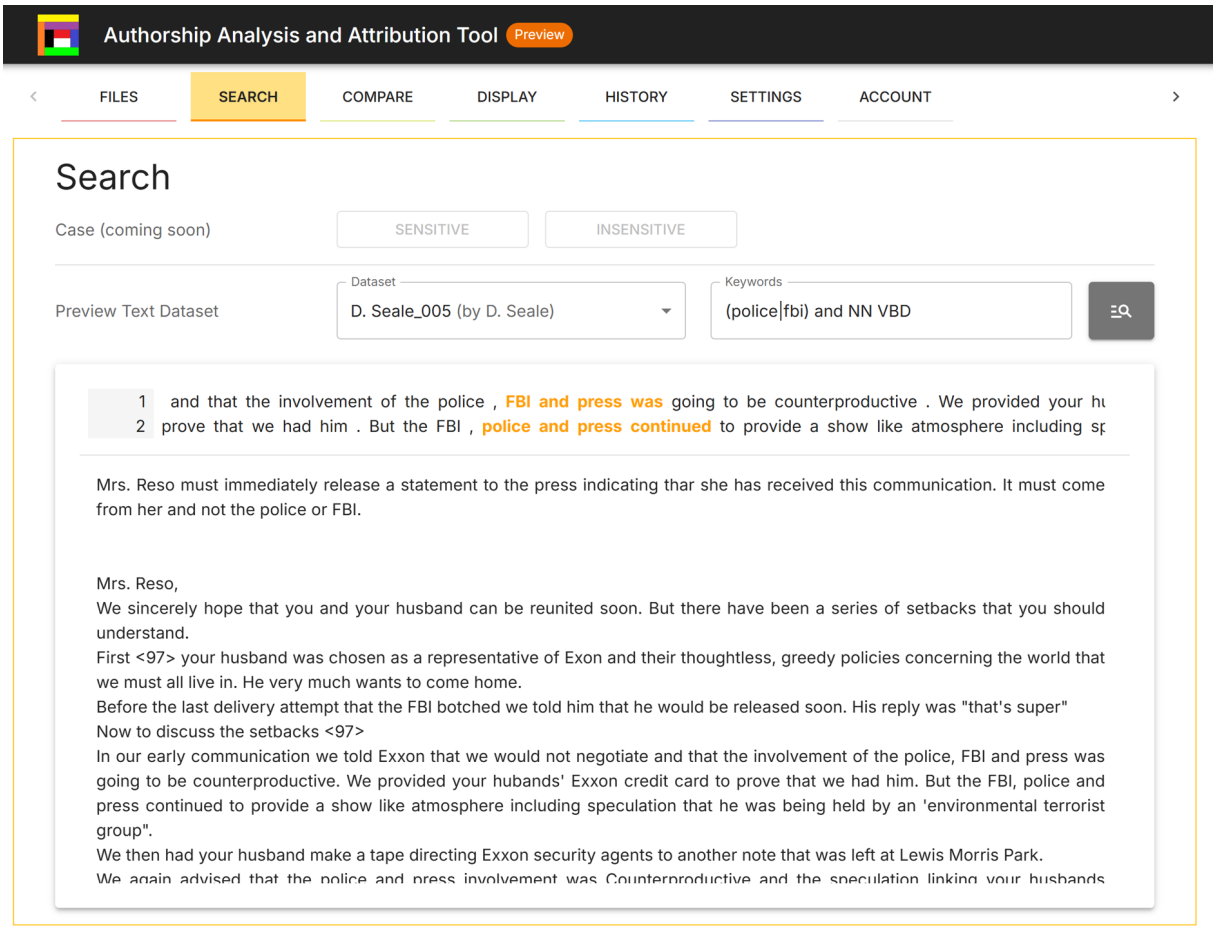


Figure 2: Screenshot of KWIC search.

intuitive interface enables non-technical users to become proficient with minimal training. The dual-mode n-gram similarity detection allows users to assess the consistency of n-gram usage within the texts of a single author (intra-corpus mode) and to evaluate similarity by comparing the questioned dataset with corpora from multiple authors (inter-corpus mode). Together, these capabilities streamline analysis, reduce reliance on multiple external tools, and support reproducible forensic workflows.

While the tool performs well for the intended tasks, several limitations remain. First, the current implementation is optimised for English texts, and performance on morphologically rich or low-resource languages has yet to be validated. Second, the accuracy of results depends on the quality of the input text. Errors introduced during transcription, OCR, or preprocessing may affect n-gram extraction and matching. Third, the system is designed for small to medium-sized datasets; although it can handle larger corpora, response times may increase, particularly during cross-corpus comparisons with multiple large K datasets.

7 Conclusion and Future Work

We have presented a dual-mode n-gram similarity detection tool purpose-built for forensic authorship analysis. The system addresses a niche not met by existing corpus or stylometric tools, enabling both intra-corpus stylistic consistency checks and inter-corpus comparison in a single, user-friendly environment. Its design emphasises transparency, reproducibility, and accessibility, making it suitable for forensic linguists without programming expertise. Evaluation on multiple datasets demonstrated perfect accuracy in shared n-gram detection and ranking.

We plan to release a production version with a two-tier access model and accompanying operational safeguards. The first tier will be a public demo environment offering read-only access to sandboxed corpora, with KWIC, frequency lists, and n-gram overlap on small sample datasets, capped query quotas, and limited file uploads, enabling immediate try-out without compromising data security. The second tier will provide pri-

Actions ANALYZE CONSISTENCY

Rows 50 | 1 2 3 4 5 ... 54

#	N-gram	Topology	Total Count	D. Seale_001 (D. Seale)	D. Seale_002 (D. Seale)	D. Seale_005 (D. Seale)	D. Seale_006 (D. Seale)	John W. Hinckley_001 (John W. Hickley)	Zodiac Killer_001 (Zodiac Killer)
1	warriors of the	4	4	1 (0.344%)	1 (0.202%)	1 (0.147%)	1 (0.248%)	0 (0.000%)	0 (0.000%)
2	of the rainbow	4	4	1 (0.344%)	1 (0.202%)	1 (0.147%)	1 (0.248%)	0 (0.000%)	0 (0.000%)
3	< 97 >	3	4	1 (0.344%)	1 (0.202%)	2 (0.293%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
4	. if you	3	4	1 (0.344%)	0 (0.000%)	2 (0.293%)	0 (0.000%)	0 (0.000%)	1 (0.171%)
5	. we have	3	4	1 (0.344%)	1 (0.202%)	0 (0.000%)	2 (0.496%)	0 (0.000%)	0 (0.000%)
6	if you do	3	3	1 (0.344%)	0 (0.000%)	1 (0.147%)	0 (0.000%)	0 (0.000%)	1 (0.171%)
7	you do not	3	3	1 (0.344%)	0 (0.000%)	1 (0.147%)	0 (0.000%)	0 (0.000%)	1 (0.171%)
8	. warriors of	3	3	0 (0.000%)	1 (0.202%)	1 (0.147%)	1 (0.248%)	0 (0.000%)	0 (0.000%)
9	i 've got	2	4	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (0.260%)	3 (0.514%)
10	place an ad	2	3	1 (0.344%)	0 (0.000%)	2 (0.293%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
11	an ad in	2	3	1 (0.344%)	0 (0.000%)	2 (0.293%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
12	ad in the	2	3	1 (0.344%)	0 (0.000%)	2 (0.293%)	0 (0.000%)	0 (0.000%)	0 (0.000%)

Figure 3: Screenshot of consistency function for trigrams.

vate workspaces for registered accounts with full access to all functionalities, including complete analysis history. The system will be deployed to a productive server using a containerised stack (e.g., Docker) with PostgreSQL, object storage for corpora, and a task queue for long-running jobs, supporting both single-tenant and multi-tenant configurations. Additional features will include single sign-on (SAML/OAuth2), encryption in transit and at rest, rate-limiting, audit logging, and monitoring

References

Laurence Anthony. 2024. Addressing the challenges of data-driven learning through corpus tool design–in conversation with laurence anthony. In *Corpora for language learning: Bridging the research-practice divide*, pages 9–18. Routledge.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Vincent Boswijk and Matt Coler. 2020. What is salience? *Open Linguistics*, 6(1):713–722.

Vaclav Brezina. 2025. Corpus linguistics and ai:# lancsbox x in the context of emerging technologies. *International Journal of Language Studies*, 19(2).

Stanley L Brodsky, Caroline Titcomb, David M Sams, Kara Dickson, and Yves Benda. 2012. Hypothetical constructs, hypothetical questions, and the expert witness. *International Journal of Law and Psychiatry*, 35(5-6):354–361.

Matteo Cardaioli, Mauro Conti, Andrea Di Sorbo, Enrico Fabrizio, Sonia Laudanna, and Corrado A Visaggio. 2021. It's a matter of style: Detecting social bots through writing style consistency. In *2021 in-*

ternational conference on computer communications and networks (ICCCN), pages 1–9. IEEE.

Elisabeth Carter. 2022. Forensic linguistics. In *Handbook of Pragmatics*, pages 572–586. John Benjamins Publishing Company.

Malcolm Coulthard. 2005. The linguist as expert witness. *Linguistics and the Human Sciences*, 1(1):39–58.

Malcolm Coulthard and Alison Johnson. 2000. *Forensic Linguistics: An Introduction to Language in the Justice System*. Routledge.

Padraig Cunningham and Sarah Jane Delany. 2021. K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, 54(6):1–25.

David DeMatteo, Sarah Fishel, and Aislinn Tansey. 2019. Expert evidence: The (unfulfilled) promise of daubert. *Psychological Science in the Public Interest*, 20(3):129–134.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence*, 19(1):109–123.

Steven HH Ding, Benjamin CM Fung, Farkhund Iqbal, and William K Cheung. 2017. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Maciej Eder, Jan Rybicki, Mike Kestemont, Steffen Pielstroem, and Maintainer Maciej Eder. 2024. stylo: R package for stylometric analyses.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

Tammy Gales, Andrea Nini, and Ellen Symonds. 2023. The threatening English language (TEL) corpus. https://research.manchester.ac.uk/en/datasets/the-threatening-english-language-tel-corpus. Accessed April 2025.

Tim Grant. 2013. TXT 4N6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, 21(2):467–494.

Javed Hussain. 2020. Enron email dataset.

Shunichi Ishihara. 2021. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International*, 327:110980.

Alison Johnson and David Wright. 2014. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law/Linguagem e Direito*, 1(1).

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The Sketch Engine. *Lexicography*, 1(1):7–36.

Carmen Klaussner, John Nerbonne, and Çağrı Çöltekin. 2015. Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities*, 30(suppl_1):i114–i129.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Andrea Nini. 2018. An authorship analysis of the jack the ripper letters. *Digital Scholarship in the Humanities*, 33(3):621–636.

Andrea Nini. 2023. *A theory of linguistic individuality for authorship analysis*. Cambridge University Press.

Andrea Nini. 2024. Idiolect: An R package for forensic authorship analysis.

Thomas C O'Brien and David D O'Brien. 2017. Effective strategies for cross-examining an expert witness. *Litigation*, 44(1):26–30.

Mike Scott. 2008. Developing wordsmith. *International Journal of English Studies*, 8(1):95–106.

David Wright. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22(2):212–241.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Actions CHECK COMPARISON

Rows 50 1 2 3 4 5 ... 14

#	Trigram	Q: D. Seale_005 (D. Seale)	K1: D. Seale_001 (D. Seale)	K2: John W. Hinckley_001 (John W. Hickley)
1	a series of	2 (0.293%)	0 (0.000%)	0 (0.000%)
2	< 97 >	2 (0.293%)	1 (0.344%)	0 (0.000%)
3	that the fbi	2 (0.293%)	0 (0.000%)	0 (0.000%)
4	that we would	2 (0.293%)	0 (0.000%)	0 (0.000%)
5	the police ,	2 (0.293%)	1 (0.344%)	0 (0.000%)
6	police , fbi	2 (0.293%)	1 (0.344%)	0 (0.000%)
7	, fbi and	2 (0.293%)	0 (0.000%)	0 (0.000%)
8	fbi and press	2 (0.293%)	0 (0.000%)	0 (0.000%)
9	police and press	2 (0.293%)	0 (0.000%)	0 (0.000%)
10	. we then	2 (0.293%)	0 (0.000%)	0 (0.000%)
11	to a phone	2 (0.293%)	0 (0.000%)	0 (0.000%)
12	they would have	2 (0.293%)	0 (0.000%)	0 (0.000%)
13	tape of your	2 (0.293%)	0 (0.000%)	0 (0.000%)
14	of your husband	2 (0.293%)	0 (0.000%)	0 (0.000%)
15	more concerned with	2 (0.293%)	0 (0.000%)	0 (0.000%)
16	concerned with apprehension	2 (0.293%)	0 (0.000%)	0 (0.000%)
17	with apprehension than	2 (0.293%)	0 (0.000%)	0 (0.000%)
18	apprehension than with	2 (0.293%)	0 (0.000%)	0 (0.000%)
19	than with your	2 (0.293%)	0 (0.000%)	0 (0.000%)
20	with your husbands	2 (0.293%)	0 (0.000%)	0 (0.000%)
21	place an ad	2 (0.293%)	1 (0.344%)	0 (0.000%)
22	an ad in	2 (0.293%)	1 (0.344%)	0 (0.000%)
23	ad in the	2 (0.293%)	1 (0.344%)	0 (0.000%)

Figure 4: Screenshot of comparison function for trigrams.

A Social Listening System for Beauty Products Using Aspect-Based Sentiment Analysis

Thanh-Nhi Nguyen[1,2], **Trong-Hop Do**[1,2]

[1]University of Information Technology,
[2]Vietnam National University Ho Chi Minh City
Corresponding author: Trong-Hop Do (hopdt@uit.edu.vn)

Abstract

In the digital age, the vast amount of unstructured data from online platforms poses significant challenges for businesses aiming to extract actionable insights. Social listening, empowered by big data analytics, emerges as a vital tool to monitor and understand consumer sentiments and trends. This paper introduces a specialized social listening system tailored for the beauty industry, integrating textual and video data sources. Central to our approach is Aspect-Based Sentiment Analysis (ABSA), which dissects consumer feedback into specific product aspects to discern nuanced sentiments. We present novel methodologies for data normalization and subject identification, crucial for enhancing the granularity and accuracy of sentiment analysis in this domain. Normalization here refers to both lexical normalization of noisy social media text and post-ASR transcript normalization, ensuring consistent and comparable linguistic input across modalities. Experimental results demonstrate the effectiveness of our system in extracting and classifying sentiments related to beauty products, surpassing existing benchmark.

1 Introduction

In today's digital age, the vast amounts of unstructured data generated from online sources, such as social media platforms, blogs, forums, and e-commerce reviews, present a significant challenge for businesses and researchers alike. These platforms are replete with valuable insights that can inform strategic decisions, yet the sheer volume and complexity of the data make it difficult to analyze and utilize effectively. One promising solution to this challenge is social listening powered by big data analytics.

Social listening is defined as the process of monitoring digital conversations to understand what customers are saying about a brand, product, or industry online. The term "social listening" refers to the practice of monitoring and analyzing user-generated content on social media platforms (Westermann and Forthmann, 2021). This involves not just tracking mentions and comments but also analyzing the sentiment, context, and trends within these conversations. By leveraging big data technologies, social listening can transform vast amounts of unstructured data into actionable insights, helping businesses to stay ahead of market trends, understand customer needs, and improve their products and services.

Numerous studies have demonstrated the benefits of social listening. For instance, it has been shown to enhance customer relationship management, support product development, and improve marketing strategies (Moe and Schweidel, 2017). By tapping into the collective voice of consumers, companies can gain a deeper understanding of market dynamics and consumer preferences.

In Vietnam, the beauty industry is particularly notable for its dynamic growth and consumer engagement. According to recent metrics, the beauty sector leads the market in sales and has experienced growth rates of nearly 150% during 2021–2023, according to recent e-commerce metrics [1]. This remarkable growth highlights the importance of understanding consumer behavior and trends within this sector. Given the high engagement levels and the substantial market share of beauty products, it becomes imperative to develop sophisticated tools to analyze consumer sentiments and trends effectively.

Aspect-Based Sentiment Analysis (ABSA) plays a crucial role in social listening systems by providing detailed insights into specific aspects of products or services that consumers mention. Unlike traditional sentiment analysis, which provides a general sentiment score for an entire text, ABSA

[1]https://metric.vn/insights/category/metric/ecom-market-research/

breaks down the text into various aspects and assesses the sentiment associated with each aspect. For example, in the beauty industry, a customer review might state: "Màu son đp nhng cht son khô quá" (*The lipstick color is gorgeous, but the texture is too dry*). ABSA would identify two aspects (color and texture) with their respective sentiments (positive for color, negative for texture). This granular level of analysis enables businesses to pinpoint exact areas of strength and weakness in their offerings, leading to more targeted improvements and marketing strategies.

Therefore, in this work, we propose a comprehensive social listening system tailored specifically for beauty products. Our proposed system is designed to leverage data from both text and video sources, incorporating advanced **Normalizer modules** to standardize the information — an effort that marks the first work of its kind. This system performs ABSA to gauge customer sentiment on specific aspects of beauty products, providing granular insights into consumer opinions. Additionally, we attempted to identify mentioned subjects within the data, enabling a more precise understanding of the topics and entities being discussed. We focus on the beauty domain, where product feedback is linguistically diverse and visually driven, making it a challenging and representative case for multimodal social listening.

In light of these developments, we propose a novel social listening system specifically designed for beauty products. Our main contributions can be summarized as follows:

1. We propose a comprehensive social listening system for the beauty industry that integrates both text and video data, with specialized **Normalizer modules** for each data type. To the best of our knowledge, this is the first work to integrate such comprehensive data normalization in a social listening system.

2. We implement an advanced **ABSA module** tailored for the beauty industry, capable of identifying fine-grained aspects and their associated sentiments. This allows for nuanced understanding of consumer opinions on specific product attributes.

3. We pioneer a subject identification feature, aiming to track not only general sentiments but also specific brands being discussed. While the current outputs require further refinement, this represents the first step towards enhancing the granularity of social listening capabilities in the beauty industry.

The rest of this paper is structured as follows: Section 2 explores related work, providing an overview of existing social listening systems and ABSA applications in e-commerce and product reviews. Section 3 outlines our methodology, detailing the design, implementation, data collection, normalization, and analysis processes of our proposed system. Section 4 presents the experimental setup and results, comparing ABSA performance against previous work. Section 5 discusses the limitations of our study, and Section 6 concludes our work.

2 Related Work

Social listening has emerged as a crucial tool for businesses to gain insights from unstructured data on digital platforms. Several studies have demonstrated its effectiveness in various domains. For instance, social listening provides valuable insights into stakeholder perceptions of a company's performance across different dimensions (Westermann and Forthmann, 2021).

Aspect-Based Sentiment Analysis (ABSA) has been widely studied in the context of e-commerce and product reviews. Nasim and Haider (Nasim and Haider, 2017) proposed an ABSA Toolkit for performing aspect-level sentiment analysis on customer reviews. Li et al. (Li et al., 2023) employed aspect-based sentiment analysis of customer-generated content to enhance the prediction of restaurant survival. For Vietnamese, Luc et al. (Luc Phan et al., 2021) built a social listening system based on ABSA for mobile e-commerce.

Recent studies have also addressed normalization in various NLP contexts. For instance, Nguyen et al. (Nguyen et al., 2024) introduced the ViLexNorm corpus for Vietnamese lexical normalization on social media, providing a strong benchmark for transforming informal user-generated text into canonical forms. In parallel, Liao et al. (Liao et al., 2023) investigated post-processing for readability in Automatic Speech Recognition (ASR) transcripts, formulating the task as a sequence-to-sequence generation problem to enhance grammaticality and fluency of spoken-text output. However, to the best of our knowledge, no prior work has integrated both lexical and ASR-based normaliza-

tion within a unified Vietnamese social listening pipeline.

Our work builds upon these foundations, integrating ABSA, text normalization, and subject identification into a comprehensive social listening system specifically designed for the Vietnamese beauty product market. Text normalization is a critical step in processing social media data and ASR output. Subject identification, particularly brand detection, is an important aspect of social media analytics. To the best of our knowledge, we are the first to develop specialized normalization techniques for both social media comments and ASR output, and attempt to detect the mentioned brands from the beauty product reviews in Vietnamese.

3 Methodology

3.1 Pipeline

Our social listening system for beauty products consists of several interconnected modules designed to process and analyze data from both text and video sources. The overall system architecture is illustrated in Figure 1.

The data flow through the system can be described as follows:

1. **Input**: The system accepts two types of input - Comments Data (text) and Videos Data (audio).

2. **Normalization**:

   - Normalization in our system covers lexical normalization (for slang, abbreviations, and misspellings) and transcript normalization (for ASR disfluencies and missing punctuation). Brand names are also standardized to canonical forms where applicable. For text data, the Comment Normalizer processes the input.
   - For audio data, the audio is first passed through the Automatic Speech Recognition (ASR) module to generate a transcript, which is then processed by the Transcript Normalizer.

3. **ABSA**: The normalized text is fed into the Aspect-Based Sentiment Analysis module, which outputs aspect-polarity pairs for each input.

4. **Subject Identification**: Additionally, the normalized text from both sources is passed through the Subject Identifier module, which extracts mentioned brands.

In our system, Apache Kafka (Kreps et al., 2011) serves as a distributed streaming platform, ingesting two types of input from various sources. Kafka allows us to decouple data producers (e.g., social media platforms, video platforms) from our processing pipeline, ensuring high throughput and fault tolerance. PySpark, the Python API for Apache Spark (Zaharia et al., 2016), is utilized to process the data in a distributed manner. PySpark enables us to perform batch and stream processing on the ingested data, allowing for efficient execution of our pipeline modules across a cluster of machines.

3.2 Modules

Our pipeline consists of five key modules:

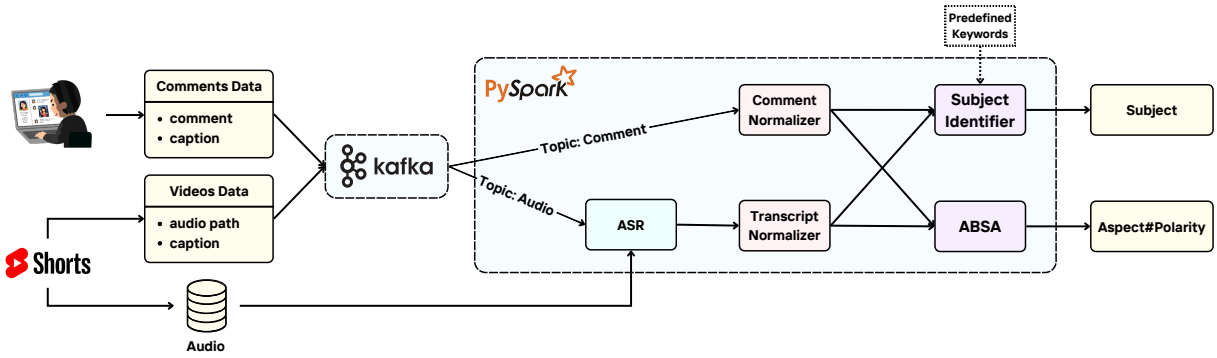
1. **ASR**: We utilize the wav2vec2-base-vietnamese-250h model (Nguyen, 2021) for converting audio to text. This model was pretrained on 13,000 hours of unlabeled Vietnamese YouTube audio and fine-tuned on 250 hours of labeled data from the VLSP 2020 ASR dataset[2]. The ASR challenge associated with this dataset is part of the annual workshop conducted by the Vietnamese Language and Speech Processing (VLSP) community, a prominent event in the field. The transcripts of this dataset were annotated exactly as the audio, retaining all disfluencies and phonetic variations while lacking any punctuations. Therefore, they necessitate further normalization in subsequent steps.

2. **Comment Normalizer**: We employ the $\text{BARTPho}_{\text{syllable}}$ base[3] version of BARTpho (Tran et al., 2022a), a pre-trained Sequence-to-Sequence model tailored for Vietnamese to normalize user comments. We trained this model on the ViLexNorm dataset (Nguyen et al., 2024), the first corpus developed for the Vietnamese lexical normalization task. The corpus comprises over 10,000 pairs of sentences meticulously annotated by human annotators, sourced from public comments on Vietnam's most popular social media platforms. This approach ensures robust normalization of

[2] https://vlsp.org.vn/vlsp2020/eval/asr

[3] https://huggingface.co/vinai/bartpho-syllable-base

Figure 1: Our overall social listening system.



colloquial and informal Vietnamese text commonly found in social media comments. For example, the comment "son mac chinh hang nhee" is normalized to "son MAC chính hãng nhé", unifying both orthography and brand capitalization.

3. **Transcript Normalizer**: Similar to the Comment Normalizer, we used the BARTPho$_{\text{syllable}}$ base model. We created augmented data to pretrain the model then fine-tuned on human-annotated data to enhance its performance. Details about the data can be found in Appendix A. This two-stage process allows the model to handle the unique challenges of normalizing ASR output, including lack of punctuation and potential transcription errors.

4. **ABSA**: Our ABSA module is based on the PhoBERT-base-v2 version of PhoBERT (Nguyen and Tuan Nguyen, 2020). We normalized an ABSE lipstick dataset (Tran et al., 2022b), then finetune the model on it. This dataset contains 16,227 reviews about lipsticks, encompassing a total of 32,775 aspect-sentiment pairs. By focusing on specific aspects of beauty products, such as packaging, texture, and effectiveness, this fine-tuning process enables our model to deliver precise and granular insights into consumer opinions and sentiments. The detailed architecture of the ABSA module is discussed in Section 3.3.

5. **Subject Identifier**: This module enhances the granularity of our social listening system by tracking specific brands mentioned in user comments and reviews. We use a list of predefined keywords and PhoNLP (Nguyen and Nguyen, 2021), a BERT-based multi-task learning model for named entity recognition, to detect brands accurately. Despite the early stage of development and the need for further improvement in output accuracy, this feature represents a significant advancement in social listening capabilities. PhoNLP was selected over general-purpose Vietnamese NER toolkits such as Stanza due to its stronger Vietnamese-specific pretraining and higher accuracy on local benchmarks.

3.3 ABSA Architecture

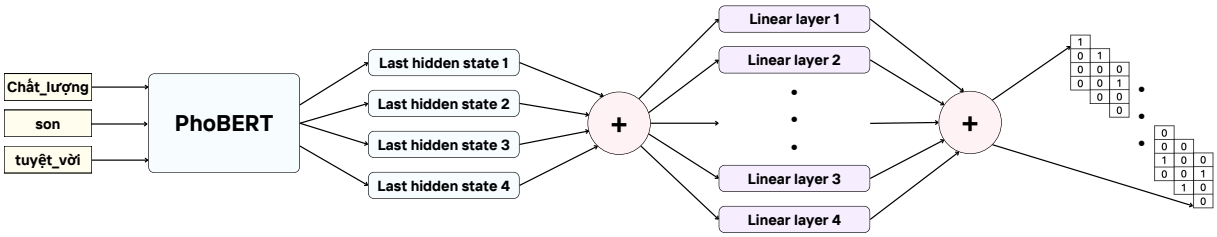
The ABSA module is a crucial component of our system, designed to identify aspects of beauty products and their associated sentiments. Our ABSA architecture is inspired by the multi-task approach for the hotel domain in the paper (Dang et al., 2022). This approach leverages PhoBERT as a pre-trained language model and processes both Aspect Category Detection and Sentiment Polarity Classification tasks simultaneously. We concatenate the last four hidden states of PhoBERT to form a comprehensive representation, which is then fed into separate layers for each task. The ABSA architecture is illustrated in Figure 2.

The architecture is as follows:

- Input Processing: The normalized text is first segmented using the VnCoreNLP toolkit (Vu et al., 2018), which ensures accurate word segmentation for Vietnamese. The segmented text is then tokenized and fed into the PhoBERT model.

- PhoBERT Encoding: The PhoBERT model processes the input and generates contextualized embeddings for each token.

Figure 2: The architecture of our ABSA module.



- Linear Layers: The last hidden states from PhoBERT are passed through a series of linear layers. Each linear layer corresponds to a specific aspect (e.g., colour, texture, smell, price). In our dataset, we define eight fixed aspects: Colour, Texture, Smell, Price, Packing, Staying Power, Shipping, and Others. Each linear head predicts the sentiment of one aspect. For unseen or ambiguous cases, predictions fall into the "OTHERS" category. This fixed-aspect design allows direct comparison with the benchmark lipstick dataset and avoids label drift when training on domain-specific data.

- Output Generation: The output of each linear layer is aggregated to produce the final predictions. For each aspect, the system determines whether it is present in the input and, if so, what the associated sentiment is (positive, negative, or neutral).

This architecture allows for simultaneous detection of multiple aspects and their sentiments within a single input, providing a comprehensive analysis of user opinions on beauty products.

4 Experiment

4.1 Experimental Setup

Each module in our pipeline was trained with specific configurations to ensure optimal performance. The training configurations for all modules are as follows:

- **ASR Module:** We utilized the pre-trained wav2vec2-base-vietnamese-250h model without further fine-tuning for our specific task.

- **Comment Normalizer:**
  - Model: $\text{BARTPho}_{\text{syllable}}$ base
  - Learning rate: 5e-05
  - Train batch size: 8
  - Evaluation batch size: 8
  - Random seed: 42
  - Optimizer: Adam
  - Number of epochs: 10

- **Transcript Normalizer:**
  - Model: $\text{BARTPho}_{\text{syllable}}$ base
  - Learning rate: 5e-05
  - Train batch size: 8
  - Evaluation batch size: 8
  - Random seed: 42
  - Optimizer: Adam
  - Training strategy: 3 epochs pretraining, followed by 5 epochs fine-tuning

- **ABSA Module:**
  - Model: PhoBERT-base-v2
  - Batch size: 32
  - Learning rate: 2e-5
  - Number of epochs: 10

- **Subject Identifier:** We utilized the PhoNLP model without further fine-tuning for our specific task.

It is important to note that our evaluation focuses specifically on the performance of the ABSA module, as it represents the final output of our pipeline. The evaluation and test results for the ABSA module are presented in Section 4.2.

To evaluate the performance of the ABSA module, we employed the F1 score metric. Weighted averages of F1-score are used to assess the overall performance of the model.

4.2 ABSA Module Results

The experimental results for the ABSA module on the development and test sets of the ABSA lipstick dataset are presented in Table 1. The results from the dataset paper (Tran et al., 2022b) are included

Table 1: ABSA module results on development and test sets (%).

Dataset	$F1_{Aspect}$	$F1_{Sentiment}$
Dev Set	98.46	97.16
Test Set	98.44	97.17
Baseline	*97.51*	*86.92*

for comparison. Note that these scores reflect the best-reported performance for the multi-task learning approach in the literature for the same dataset and serve as the baseline for our experiment.

The results demonstrate the effectiveness of the proposed system. On both the development and test sets, our ABSA module achieves high F1 scores for aspect extraction and sentiment classification, significantly surpassing the results reported in the dataset paper (Tran et al., 2022b). Specifically, our model attains an $F1_{Aspect}$ score of 98.46% on the development set and 98.44% on the test set, compared to 97.51% reported in the dataset paper. Similarly, for $F1_{Sentiment}$, our model achieves scores of 97.16% on the development set and 97.17% on the test set, outperforming the 86.92% reported in the dataset paper.

Our module's performance indicates significant improvements over the baseline, particularly in sentiment classification, where we observe a substantial increase of over 10 percentage points in F1. This improvement can be attributed to the advanced techniques employed in our system: the **Normalizer** modules. We note that both data normalization and the use of augmented samples contribute to improved consistency and reduced noise in the training data. While no separate ablation was conducted, manual inspection suggests normalization plays the dominant role in enhancing sentiment classification accuracy.

Although no formal statistical significance test was conducted, the observed improvements consistently exceeded run-to-run variance, suggesting that the gains are statistically meaningful rather than due to random fluctuations.

In addition to evaluating overall aspect and sentiment detection performance, we also analyzed the performance of our ABSA module for individual aspects. The F1-scores for aspect detection, along with the sample counts for each aspect, are presented in Table 2. These results provide a more detailed view of how well our module performs on specific aspects compared to the baseline.

The table highlights that our ABSA module performs well in most aspects, especially for "COLOUR" and "OTHERS," where it achieved the highest F1-scores of 98.95% and 97.72%, respectively. These aspects also have the highest number of samples, suggesting that the model benefits from having more data for training and evaluation. In addition, the lower performance in "SMELL" and "SHIPPING" can be attributed to the complexity and variability of these specific aspects in the dataset.

Overall, our ABSA module achieves significant advancements in aspect extraction and sentiment classification, surpassing the best-reported benchmarks on the ABSA lipstick dataset (Tran et al., 2022b). However, while excelling in most aspects with ample data, challenges persist in certain aspects, which require further refinement in handling linguistic variability and dataset balance.

5 Limitations

While our aspect-based sentiment analysis module demonstrates strong performance, the **Subject Identifier** component currently offers preliminary results and remains an area for further improvement. Without a comprehensive predefined brand list, the PhoNLP model faces challenges in recognizing newly emerging or less frequent brand names. This limitation reflects the inherent dynamics of the beauty market rather than a constraint of the model itself. To address this, future work will focus on expanding the brand lexicon and exploring adaptive named-entity recognition strategies capable of identifying unseen brands in real time.

Moreover, the current version does not yet associate sentiment polarity directly with identified brands. Integrating brand-level sentiment analysis within the **Subject Identifier** module is a promising next step toward delivering more fine-grained insights into consumer attitudes and market trends.

Finally, although the evaluation has been conducted primarily on beauty product reviews, the proposed architecture is domain-agnostic and can be readily extended to other sectors (e.g., fashion, electronics) with minimal retraining. This generalizable design highlights the scalability of our pipeline beyond the current experimental domain.

6 Conclusion

In this study, we have proposed and implemented a specialized social listening system tailored for the

Table 2: The F1-score (%) for aspect detection in each aspect with sample count

Aspect	Baseline	Our Module	Sample Count
SMELL	98.01	97.00	363
COLOUR	96.60	**98.95**	9720
STAYINGPOWER	95.67	**95.90**	464
PRICE	96.37	**96.88**	324
SHIPPING	97.15	96.10	530
PACKING	95.89	**96.72**	316
TEXTURE	94.88	**95.26**	460
OTHERS	94.68	**97.72**	754

beauty industry, leveraging advanced techniques in Aspect-Based Sentiment Analysis (ABSA), data normalization, and subject identification. Our system integrates textual and video data sources, aiming to extract detailed insights into consumer sentiments and trends surrounding beauty products. Through our experimental evaluations, we have demonstrated the effectiveness of our approach. The ABSA module achieved significant advancements in aspect extraction and sentiment classification compared to existing work. Furthermore, the incorporation of data normalization modules tailored for Vietnamese text and video transcripts has proven crucial in enhancing the accuracy and reliability of sentiment analysis outputs. We also highlights the importance of subject identification in social listening systems, particularly in tracking mentions of brands and specific entities within consumer discussions. While our current model shows promising results, further research and refinement are needed to adapt to evolving brand landscapes and improve accuracy in real-time scenarios. We hope that this research contributes a comprehensive framework for applying social listening and ABSA methodologies to gain deeper insights into consumer behaviors and preferences within the dynamic beauty market. Future work will focus on extending the system to additional domains and performing significance testing to further verify the robustness of the observed improvements.

Acknowledgments

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D4-2025-14.

References

Hoang-Quan Dang, Duc-Duy-Anh Nguyen, and Trong-Hop Do. 2022. Multi-task solution for aspect category sentiment analysis on vietnamese datasets. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 404–409.

Jay Kreps, Neha Narkhede, Jun Rao, and 1 others. 2011. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB*, volume 11, pages 1–7. Athens, Greece.

Hengyun Li, XB Bruce, Gang Li, and Huicai Gao. 2023. Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. *Tourism Management*, 96:104707.

Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2023. Improving readability for automatic speech recognition transcription. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II 14*, pages 647–658. Springer.

Wendy W Moe and David A Schweidel. 2017. Opportunities for innovation in social media analytics. *Journal of product innovation management*, 34(5):697–702.

Zarmeen Nasim and Sajjad Haider. 2017. Absa toolkit: An open source tool for aspect based sentiment analysis. *International Journal on Artificial Intelligence Tools*, 26(06):1750023.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–7.

Thai Binh Nguyen. 2021. Vietnamese end-to-end speech recognition using wav2vec 2.0.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024. ViLexNorm: A lexical normalization corpus for Vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437, St. Julian's, Malta. Association for Computational Linguistics.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022a. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Quang-Linh Tran, Phan Thanh Dat Le, and Trong-Hop Do. 2022b. Aspect-based sentiment analysis for Vietnamese reviews about beauty product on E-commerce websites. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 767–776, Manila, Philippines. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Arne Westermann and Jörg Forthmann. 2021. Social listening: a potential game changer in reputation management how big data analysis can contribute to understanding stakeholders' views on organisations. *Corporate Communications: An International Journal*, 26(1):2–22.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM*, 59(11):56–65.

A Details of Data Used for Transcript Normalization

A.1 Human-annotated Data for Finetuning

To construct our gold dataset, we leveraged a 100-hour speech public dataset from Vinbigdata[4], which serves as a clean subset of the VLSP2020 ASR competition[5]. Specifically, this dataset comprises 100 hours of speech data collected from open sources and manually transcribed with an impressive accuracy rate of 96%.

The Vinbigdata dataset comprises a total of 56,427 samples. We excluded samples with fewer than four characters to maintain quality. The transcripts of this dataset were annotated exactly as the audio, retaining all disfluencies and phonetic variations while lacking any punctuations. This raw, unprocessed data formed the basis of our gold dataset.

To generate readable target transcripts, we normalized the original transcripts from the dataset of Vinbigdata. This normalization involved removing deletable portions of disfluencies and fillers, shortening abbreviations, and normalizing phonetic variations to correspond to common Vietnamese words. Specifically, we annotated 596 pairs for training, 250 for the development set, and 250 for the test set. Importantly, all four types of post-processing (repetitions, fillers, abbreviation, and phonetic variations) were evenly distributed across these sets. The remaining samples were reserved for data augmentation purposes.

This process ensured that our gold dataset not only reflected the original transcription characteristics but also captured the nuances of ASR model errors.

A.2 Augmented Data for Pretraining

Given the time and effort-intensive nature of manual annotation, we explored data augmentation strategies for ASR post-processing task. Our approach involves two key techniques aimed at enhancing the diversity and robustness of our training data. We pretrained the Transcript Normalizer using a combined 30,000 augmented samples: 15,000 generated through predictions and 15,000 synthesized.

A.2.1 Utilizing ASR Model Predictions:

We leveraged the predictions a the ASR model as an augmentation strategy, specifically focusing on the incorrect predictions. To identify errors in the remaining data, we employed a state-of-the-art Vietnamese ASR model[6]. This model, fine-tuned

[4] https://institute.vinbigdata.org/events/vinbigdata-chia-se-100-gio-du-lieu-tieng-noi-cho-cong-dong/

[5] https://vlsp.org.vn/vlsp2020/eval/asr

[6] https://huggingface.co/khanhld/wav2vec2-base-vietnamese-160h

on approximately 160 hours of diverse Vietnamese speech data, has not yet incorporated a language model but has yielded promising results. Out of the remaining 48,927 samples, we randomly selected 15,000 samples where the model predictions were incorrect, thus capturing the errors inherent in the ASR system. We paired the incorrectly predicted samples with their corresponding transcripts from the Vinbigdata dataset. By doing so, we introduced more varied forms of mistakes that an ASR system may generate, enriching our training data and improving the model's ability to handle diverse errors.

A.2.2 Synthetic Data Generation:

In addition to using model predictions, we employed a synthetic data generation approach. We collected the most popular articles from Vietnamese Wikipedia[7] in the year 2023. After gathering the text, we pre-processed it by removing all characters except those in the alphabet or numerical characters. The entire text was then converted to lowercase, and random segments of text, each with a length of fewer than 130 tokens, were extracted to simulate the punctuation-free transcripts in the Vinbigdata dataset. To maintain quality, segments with fewer than four characters were excluded. Afterwards, we randomly selected 15,000 samples.

Subsequently, we performed synthetic data augmentation by simulating repetitions, fillers, and phonetic variations. To prevent confusion, only one of the three synthetic methods was applied to each sample, selected randomly with equal probabilities (33.33%). The synthesis methods were implemented as follows:

1. **Repetitions:**
    - Examine each token and decide whether to introduce a repetition with a 30% probability.
    - Special case: If the token being assessed is a conjunction or connective, the decision probability increases to 50%.
    - Implement repetitions: 80% of instances involve duplication, 15% triplication, and 5% quadruplication.
2. **Fillers:**
    - Examine each token and determine whether to include a filler with a 20% probability.
    - Implement: Add one token after the chosen token with the following probabilities:
        - 70% for selecting from frequently used filler words.
        - 30% for selecting from less common filler words.
3. **Phonetic Variations:**
    - Examine each token and decide whether to apply a phonetic variation with a 30% probability.
    - Implement:
        - Analyse the first and last characters of the token to identify whether they belong to a list of words prone to phonetic confusion.
        - If affirmative, replace one of the two, either the beginning or the end of the token, with a probability of 50%.

Additionally, in (1) and (2), we employed word segmentation using VnCoreNLP (Vu et al., 2018) (applied to 50% of samples) before iterating through tokens. This introduced cases where repetitions span both simple and compound words in (1) and fillers may or may not be inserted between compound words in (2).

[7] https://vi.wikipedia.org/

Balancing Heat and Clarity: Alcohol, Tea, and the Body in Premodern Chinese Texts

Clio Luo
University of Chicago, Illinois, United States
clioluvsj@uchicago.edu

This paper explores how two familiar beverages—tea (茶 cha) and alcohol (酒 jiu)—became philosophical and medical opposites in premodern China yet were always understood as partners in achieving bodily and moral balance. Drawing on eight key works ranging from the Western Han silk manuscripts to late Qing household recipe books, I show how the act of drinking was conceived as both therapy and self- cultivation. The Recipes for Fifty-Two Ailments: From the Mawangdui Silk Manuscripts (c. 200 BCE) establishes the pharmacological foundation by describing wine as a warm solvent guiding herbs through the body. By the Tang period, treatises such as Lu Yu's Cha Jing and the anonymous Jiu Jing portray tea and alcohol as moral emblems—clarity against indulgence—while Cha Jiu Lun gives them voices in debate. Later texts, from the Shi Jian Ben Cao and Qian Jin Shi Zhi to Qing works like Sui Yuan Shi Dan and Sui Xi Ju Yin Shi Pu, integrate dietary, medical, and aesthetic discourses. In addition to close reading, I employ computational text analysis to visualize how moral and physiological vocabularies cluster around these substances. Using a controlled lexicon of key terms (e.g., 清 clear, 毒 toxin, 德 virtue, 心 heart), I generate frequency charts, co-occurrence heatmaps, and network graphs to trace semantic relationships across the corpus. The results reveal two recurring clusters: a "cool" domain linking tea with clarity and integrity, and a "warm" domain linking wine with heat and passion. Together, they illustrate a persistent Chinese logic of healing in which good living depends not on abstinence but on harmonizing opposites—cool and hot, bitter and sweet, clarity and warmth.

From Tea to Symbol: A Multimodal Discourse Analysis of HeyTea's Branding Discourses

Tian Gao
The Hong Kong Polytechnic University, Guangdong University of Technology
25041138g@connect.polyu.hk

In the highly-competitive market of China's new tea beverages, the strategy of product differentiation is difficult to be put into effect. Marketing and branding has become more essential. This study investigates how leading beverage brands like HeyTea use linguistic and visual resources to build a unique brand identity that resonates with the younger generation of consumers.

Guided by a multimodal discourse analysis framework, this research tries to investigate the under-explored area of marketing discourse for Chinese tea beverages. It presumes that brand identity is discursively built through the combination of language and other visual resources. Data was collected from HeyTea's official website, WeChat public account and its online menu, comprising over 30 product names and their visual designs.

The data analyses reveal three main strategies utilized by HeyTea. First, at the linguistic level, HeyTea utilizes creative lexical innovations, such as the affectionate reduplication "芝芝" (Zhizhi, cheesy) and "啵啵" (Bobo, bubble), to create a sense of intimacy. Second, the product names incorporate traditional Chinese culture (e.g., "喜柿多多", which means "There are many happy events"), which transcend functional description, making the consumption experience as an aesthetic and emotional event. Third, the unique brand image is reinforced by a consistent visual demonstration employing muted colors and minimalist design, aiming to construct an identity of "inspiration" and premium quality.

The findings in this study indicate that the aforementioned strategies effectively reconstruct tea beverages from mere drinks into consumable symbols of modern, stylish lifestyles. The research concludes that the strategic use of discourse of Chinese new tea drinks can serve as a tool for observing the dynamics of contemporary Asian consumer culture, where consumer identity is built through the symbolic act of consumption. The current study expects to offer insights into the understanding of brand communication in the experience economy.

Foodies also Need Their Own GPTs: Evaluating Language Models in Visual Question Answering on the `WorldCuisines` Dataset

Genta Indra Winata
Capital One AI Foundations
gentaindrawinata@gmail.com

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Food is universally acknowledged as an important medium of expression of cultures, and at the same time a way for different people and traditions to connect to each other (Wahlqvist, 2007). Dishes reflect local identities and can tell stories, and those can be shared in turn across countries that share a culinary heritage (Anderson, 2014). There can be slight variations on similar recipes, often difficult to distinguish for the unexperienced eye; at the same time, with the global increase of food tourism all over the world (Ellis et al., 2018), there will be soon the need for new technologies to fill the gap between the image of a dish on a menu and the curiosity of a tourist willing to taste something new. Recognizing a dish and providing information about it are not trivial challenges even for multimodal large language models (MLLMs), given the large variety of existing dishes and recipes. In order to stimulate research on food-related visual question answering, we developed the WorldCuisines dataset (Winata et al., 2025), a multilingual and multicultural benchmark comprising text-image pairs of dishes for 30 different languages, for a total of more than 1 million data points. The dataset supports different tasks, such as identifying specific food types and their origins from their pictures. Our preliminary evaluation shows that performance varies a lot across different models. MLLMs generally perform better when they are provided with information about a typical location for a given dish in the prompt (e.g. I am in Hanoi and I am about to eat this now, together with an image showing a *bánh mì*), while they struggle when the prompt contains a location that is not typical (e.g. consider the same prompt as before, but with Hokkien fried rice instead of *bánh mì*). Such findings suggest that there MLLMs can be easily misled by unexpected textual information about the location of the user, and therefore they still have a lot of room for improvement.

References

Eugene Newton Anderson. 2014. *Everyone Eats: Understanding Food and Culture*. NYU Press.

Ashleigh Ellis, Eerang Park, Sangkyun Kim, and Ian Yeoman. 2018. What Is Food Tourism? *Tourism Management*, 68:250–263.

Mark L Wahlqvist. 2007. Regional Food Culture and Development. *Asia Pacific Journal of Clinical Nutrition*, 16:2.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025. WORLDCUISINES: A Massive-scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. In *Proceedings of NAACL*.

Conceptual Metaphors in Food Reviews: LLM-Based Implications for Korean Discourse

Charmgil Hong
Handong Global University
charmgil@handong.edu

Jong-Bok Kim
Kyung Hee University
jongbok@khu.ac.kr

Seulkee Park
Kyung Hee University
seulkeepark@khu.ac.kr

Yunseong Choe
Handong Global University
22000758@handong.ac.kr

Abstract

This study analyzes conceptual metaphors in Korean restaurant reviews, focusing on how taste, service, and price are structured through systematic TARGET–SOURCE mappings. Using corpus annotation and Conceptual Metaphor Theory, we identify metaphor types and evaluate how Korean-specific metaphors affect LLM interpretation and performance under different prompt configurations.

This study examines how conceptual metaphors structure Korean restaurant review discourse, with a particular focus on food-taste expressions and their underlying TARGET-SOURCE mappings. Building on conceptual metaphor (Lakoff and Johnson, 1980, 2008) and Conceptual Metaphor Theory (CMT) (Ahrens, 2010), where concrete source domains provide cognitive structure for more abstract target domains (e.g. IDEAS ARE FOOD: *He devoured the book*; *We don't need to spoon-feed our students*), we investigate how Korean reviewers draw on a wide range of concrete experiential domains to evaluate taste, service, and price. Conceptual mapping is modeled as a systematic correspondence between a concrete source domain and an abstract evaluative target domain (McGlone, 1996; Türker, 2013; Choi, 2017; Kim, 2024). Because GenAI directly adopts these Korean-specific abstract target-domain meanings and, through negative transfer, incorrectly judges them as concrete, its interpretations often diverge from human intuitions.

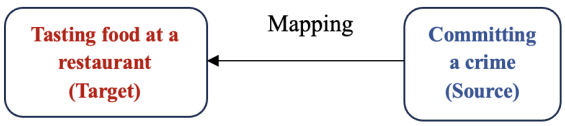


Figure 1: Conceptual mapping in a food review

A key observation in Korean discourse is that many food-related expressions do not directly talk about 'food as a physical entity', but instead denote 'the taste of food as the conceptual target'. Because Korean expressions often encode meanings implicitly, the intended abstract target remains partly covert in the surface form and is recovered through context. For instance, as in (1), what is evaluated is not the stew as an object but its taste, which is metaphorically mapped to the source domain CRIME.

(1) Ccikay-ka pap.totwuk-ida.
stew- rice.thief-
"(lit.)The stew is a rice thief; (met.) The taste of stew makes you eat up all your rice.'

Empirically, we adopt the Hwang (2024) Korean restaurant review dataset as our primary corpus. The data consist of 2,708 sentences with binary sentiment annotations (positive/negative), obtained by extracting restaurant names from Seoul's public Restaurant Business Permit Information, crawling Kakao Map reviews, segmenting them into sentences, and tagging sentiment.

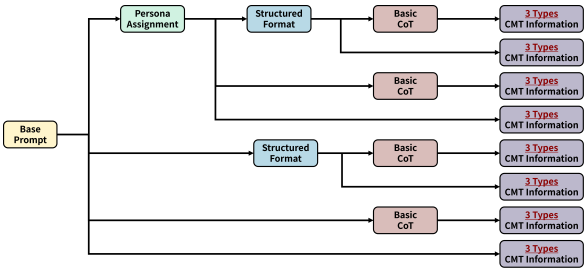


Figure 2: Cumulative prompt configuration tree

To anchor CMT analysis, we define ten evaluative target domains frequently occurring in review discourse: TASTE, KINDNESS, PRICE, FOOD, COST-EFFECTIVENESS, ATMOSPHERE, RESTAURANT, MENU, SERVICE, and ORDER. Using the Kiwi tokenizer, we identify 2,491 lexical items and select the top keywords per category, among which TASTE-, SERVICE-,

and MONEY VALUE-related items are the most prominent.

A three-stage manual annotation pipeline is then implemented. First, each sentence is assigned to one of the ten target domains (e.g., the price is reasonable → PRICE). Second, sentences are evaluated for the presence of conceptual metaphors. Third, metaphorical sentences are mapped into TARGET IS SOURCE format. This process yields 108 metaphorical expressions and 118 unique metaphorical mappings, with TASTE, SERVICE, and MONEY VALUE as central targets. Representative mappings are illustrated by naturally occurring examples:

(2) **TASTE IS A MENTAL STATE**

a. Talkthwikim-i cincca michy-ess-e-yo.
fried.chicken- really crazy—POL
"(lit.) The fried chicken really went crazy; (met.) The fried chicken is good.'

TASTE IS A CRIME

b. Mas-i cwuk-i-n-ta.
taste- ill—
'(lit.) The taste kills (someone); (met.) The taste is amazing.'

TASTE/PRICE IS OUT

c. Pwutay.cenkol-un mas-kwa
army.stew-TOP taste-
kasengpi-ka nemchye.hulu-n-ta.
money.value- overflow–
'(lit.) As for the army stew hotpot, taste and value-for-money overflow.; (met.) It's full of flavor and extremely cost-effective.'

According to the three-way distinction proposed by Lakoff and Johnson (1980), the metaphorical expressions identified in the corpus fall into structural, ontological, and orientational types. Structural metaphors conceptualize evaluative meanings through more familiar domains, and in Korean food reviews this appears primarily as personification, where non-human targets such as taste or price are described with human mental or psychological predicates, as in (2a). Ontological metaphors draw on concrete entities to express positive or negative qualities, as in mappings in (2b). Orientational metaphors, by contrast, rely on spatial schemas and movement, producing expressions in (2c) such as flavors or value "bursting out" (TASTE/PRICE IS OUT).

Finally, we outline an LLM-based methodological framework that integrates this linguistically grounded CMT annotation with Large Language Models. Three Korean-capable LLMs (EXAONE-4.0-32B, GPT-OSS-20B, Qwen3-30B-A3B) are evaluated under 32 cumulative prompt configurations combining task instructions, expert persona, structured XML-like tags, basic chain-of-thought reasoning, and different types of CMT information injection. This design allows us to assess how well LLMs identify metaphors and produce appropriate TARGET IS SOURCE mappings, and how factors such as structured prompts and domain-specific metaphor knowledge influence their performance. The study thus combines theoretical conceptual metaphors, corpus-based analysis, and LLM-based modeling to provide a detailed account of conceptual metaphors in Korean food review discourse and to lay the groundwork for extending this approach to other evaluative domains, including hospitals, institutions, and schools.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2025S1A5C3A02006302).

References

Kathleen Ahrens. 2010. Mapping principles for conceptual metaphors. In Graham Low, Zazie Todd, Alice Deignan, and Lynne Cameron, editors, *Researching and Applying Metaphor in the Real World*, pages 185–208. John Benjamins Publishing Company.

Young-ju Choi. 2017. Comparison of the concept eat as a metaphorical source in korean and english. *The Mirae Journal of English Language and Literature*, 22(3):285–306.

Chaemin Hwang. 2024. 24-1_dsl_modeling_nlp2_restaurant_review_sentiment_analysis. https://github.com/Chaemin-Hwang/24-1_DSL_Modeling_NLP2_Restaurant_Review_Sentiment_Analysis. GitHub Repository.

Jong-Bok Kim. 2024. *English and Korean in Contrast: A Linguistic Introduction*. John Wiley & Sons.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Matthew S McGlone. 1996. Conceptual metaphors and figurative language interpretation: Food for thought? *Journal of memory and language*, 35(4):544–565.

Ebru Türker. 2013. A corpus-based approach to emotion metaphors in korean: A case study of anger, happiness, and sadness. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, 11(1):73–144.

Landscapes of Matcha: Cultural Semiotics in Japan and Taiwan

Miwa Morishita
Kobe Gakuin University
miwa@gc.kobegakuin.ac.jp

Yasunari Harada
Waseda University
harada@waseda.jp

Abstract

This presentation explores the linguistic and visual landscapes of matcha in cafés and shops in Japan and Taiwan. The analysis is based on about forty photographs of signage, menus, and product labels collected during research visits between 2022 and 2024. The presenters, who conducted extended fieldwork in Taiwan, examine how matcha is linguistically and culturally represented across both countries.

Originally a ceremonial and prestigious beverage in Japan, matcha has been transformed into a popular flavor for sweets and drinks. Once linked mainly to the tea ceremony, it is now found in convenience stores and drugstores in products such as KitKat, cream puffs, and parfaits. This shift reflects both the popularization of a high-culture taste and the commodification of cultural symbols in everyday consumption.

Taiwanese tea culture is highly differentiated, with green tea (unfermented), white tea (lightly fermented), yellow tea (post-fermented), oolong (semi-fermented), black tea (fully fermented), and dark tea (after-fermented). Even in ordinary cafés, customers are asked to choose sugar levels from 0% to 100%, a question that surprises Japanese visitors used to unsweetened tea. In this context, matcha is understood as a type of green tea but is marketed under names such as "Japanese Kyoto Matcha" or "Japanese-style Matcha Milk," reflecting creative localization.

In Japan, certain expressions rarely seen abroad - usucha (thin tea) and koicha (thick tea) - remain integral to cultural discourse. The matcha most people imagine corresponds to usucha, while koicha uses two to three times more powdered tea and is reserved for formal occasions. Phrases such as "koicha no" ("rich matcha-flavored") convey luxury and authenticity. The spelling matcha itself marks linguistic innovation: unlike the traditional maccha (Hepburn) or mattya (Kunrei), matcha has become a globally recognizable hybrid form signifying new international identity. Photographs show that Taiwanese cafés and dessert shops often draw on Japanese regional imagery. Names such as Kyoto, Uji, and Shizuoka - major tea regions - appear in shop titles and labels, revealing both admiration for and familiarity with Japan. Global brands like Starbucks reinforce this framing by listing origins such as Kagoshima, Shizuoka, Mie, Nara, Kyoto, and Fukuoka. Meanwhile, hojicha (roasted green tea), now popular internationally and known in Taiwan as Fukichi-cha, is promoted as a specialty from Kyoto and Nara. Alongside these authenticity-oriented examples, creative fusions such as matcha pistachio desserts are common in Taiwan, though many Japanese find the combination overpowering. In Japan, matcha ramen and matcha curry udon from Uji exemplify playful experimentation within traditional cuisine. Together, these examples show how matcha functions as a semiotic bridge between Japan and Taiwan—symbolizing both tradition and innovation. Through these images, the presentation illustrates how the linguistic landscapes of matcha express national identity and intercultural adaptation, revealing the global recirculation and recontextualization of Japanese taste and language.

Toward the Development of a Japanese Food Culture QA

Waka Ito
Japan Women's University
m2016013iw@ug.jwu.ac.jp

Manaka Odagaki
Japan Women's University

Haruka Tsuchida
Japan Women's University

Yuha Nishigata
Japan Women's University

Yui Obara
Japan Women's University

Kimio Kuramitsu
Japan Women's University
kuramitsuk@fc.jwu.ac.jp

Abstract

"Food" is one of the most complex domains of cultural knowledge, as it reflects historical, regional, and social contexts. It has also been identified as one of the most challenging domains for large language models (LLMs) to reproduce accurately. This study aims to evaluate the food culture understanding of LLMs by developing a question-answering (QA) dataset focused on Japanese food culture.

We have been developing SakuraQA, a QA dataset designed to assess knowledge of contemporary trends in Japan. In this study, we extend the food culture category of SakuraQA and construct a new Japanese Food Culture QA dataset.

The dataset consists of multiple-choice questions designed to evaluate both traditional and contemporary aspects of Japanese food culture, organized into two main categories: (1) Traditional and Regional: Questions assessing knowledge of Japan's traditional and regionally grounded food culture, such as local specialties and seasonal customs (e.g., "What dish do Japanese people traditionally eat during the New Year holidays?"). (2) Contemporary and Popular: Questions reflecting modern trends in Japanese food culture, including neologisms, abbreviations, restaurant chains, new products, and social media–driven food trends (e.g., "What dish is commonly referred to as 'TKG' in Japan?").

Through these QA tasks, we aim to reveal how well LLMs comprehend Japanese cultural understanding and adapt to both traditional and contemporary cultural contexts. We compare Japanese, multilingual, and closed-source models to analyze differences in cultural understanding across categories. Based on this comparison, we provide insights into how Japanese LLMs can further improve their handling of cultural knowledge.

This study addresses the question of how cultural understanding can be effectively integrated into Japanese LLMs. By situating the discussion in an international context, we aim to contribute to a broader perspective on cultural understanding in LLMs and to the development of evaluation frameworks that capture the depth of food culture understanding in LLMs.