LREC 2026

Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026

Workshop Proceedings

Editors
Arturo Montejo-Ráez, Cristina Grisot, Joanna Blochowiak

11 May 2026

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-85-2

Preface

Welcome to the proceedings of **LLMs4SSH: Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities**, held in conjunction with **LREC 2026** in **Mallorca, Spain**, on **11 May 2026**.

Large Language Models (LLMs) are rapidly evolving in terms of multilinguality, multimodality, reasoning capabilities, and agentic behaviour. These developments have significantly advanced the state of the art in language and data processing. However, despite their impressive performance, current LLM-based technologies remain only partially aligned with the research needs, methodological traditions, and epistemic values of the Social Sciences and Humanities (SSH). Addressing this gap requires closer dialogue between the Language Technologies (LT) community and SSH scholars, as well as careful consideration of the broader ethical, societal, and interpretability challenges posed by these models.

The LLMs4SSH workshop aims to provide a dedicated forum for researchers and practitioners working at the intersection of AI, Natural Language Processing, Digital Humanities, linguistics, social sciences, cultural heritage, and related disciplines. The workshop explores how multilingual, multimodal, and reasoning-oriented LLMs can support SSH research tasks, including the processing of historically and culturally diverse materials, the interpretation of complex narratives, and the analysis of social and cultural phenomena. At the same time, it encourages critical reflection on the methodological implications of using LLMs in SSH contexts, with particular attention to transparency, bias, representativeness, and responsible use.

This full-day workshop brings together experts from both LT and SSH communities to exchange perspectives, share recent research, and identify key challenges and opportunities for future collaboration. Through the contributions presented in these proceedings, LLMs4SSH seeks to define research priorities and foster interdisciplinary approaches that shape the development of LLMs in ways that better support SSH scholarship.

We hope that the workshop and these proceedings will stimulate further discussion and inspire new collaborations toward developing language technologies that are not only powerful, but also meaningful and responsible in the context of SSH research.

The Organizing Committee.

Organizing Committee

Arturo Montejo-Ráez (Universidad de Jaén)

Jeremy Barnes (HiTZ Center, UPV/EHU)

Elena Battaner (Universidad Rey Juan Carlos)

Joanna Blochowiak (University of Zurich)

Cristina Grisot (University of Zurich)

Nikola Ljubešić (Jožef Stefan Institute)

Maciej Piasecki (Wrocław University of Science and Technology)

German Rigau (HiTZ Center, UPV/EHU)

Marko Tadić (University of Zagreb)

Friedel Wolff (SADiLaR, North-West University)

Table of Contents

Workshop Program

Monday, May 11, 2026

9:00–9:30	**Welcome** Chair: Cristina Grisot
9:30–10:30	**Invited talk** Chair: Arturo Montejo-Ráez
9:30–10:30	*Advanced LLMs for Humanities Research* Sarah Oberbichler, University of Luxembourg
10:30–11:00	**Coffe break and poster session**
10:30–11:00	*State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?* Taja Kuzman Pungeršek, Peter Rupnik, Ivan Porupski, Vuk Dinić and Nikola Ljubešić
10:30–11:00	*Exploring the Use of Large Language Models in Critical Discourse Analysis: A Consensus-Based Pilot Study* Emiliano Giovannetti and Francesca Cristiano
10:30–11:00	*LLM Evaluation in Practice: A Review of Metrics, Practitioner Insights, and Lessons Learned* Roos M. Bakker, Marianne Witte-Schaaphok, Julia García-Fernández, Tom Brand, Jens van der Weide and Stephan Raaijmakers
10:30–11:00	*Is Human–LLM Interaction Culture-Dependent? A Cross-Linguistic NLP Analysis of Student Interviews on AI-Assisted Thesis Writing* Madalina Chitez, Karla Csuros, Dejana Jelena Milićević, Petya Osenova, Stefan Marinov, Teodor Valchev, Nikolay Paev, Otto Kruse, Christian Rapp, Andreea Dinca, Roxana Rogobete, Claudia Doroholschi, Loredana Punga, Anabella Costache, Dumitru Tucan and Cristina Banciceru
10:30–11:00	*Next Reply Prediction X (NRP-X) Dataset: Linguistic Discrepancies in Naively Generated Content* Simon Münker, Nils Schwager, Kai Kugler, Michael Heseltine and Achim Rettinger
10:30–11:00	*Quid est VERITAS? A Modular Framework for Archival Document Analysis* Leonardo Bassanini, Ludovico Biancardi, Alfio Ferrara, Andrea Gamberini, Sergio Picascia and Folco Vaglienti

Monday, May 11, 2026 (continued)

10:30–11:00 *Do We Still Need Corpora and Corpus Analysis Platforms? Discourse Analysis in Times of LLMs*
Julia Krasselt, Dolores Lemmenmeier-Batinić and Philipp Dreesen

10:30–11:00 *GaelEval: Benchmarking LLM Performance for Scottish Gaelic*
Peter Devine, William Lamb, Beatrice Alex, Ignatius Ezeani, Dawn Knight, Mícheál J. Ó Meachair, Paul Rayson and Martin Wynne

10:30–11:00 *Argumentation through Discourse Relations and Subjectivity: Introducing FreCaDiS, a French Multi-Genre Corpus*
Joanna Blochowiak and Cristina Grisot

11:00–12:30 **Oral presentations**
Chair: Elena Battaner

11:00–11:30 *Text-only Domain Adaptation for Low-Resource ASR Using Large Language Models*
William Lamb, Dongge Han, Ondrej Klejch and Peter Bell

11:30–12:00 *Benchmarking LLMs for Aspect-Based Sentiment Classification in Slovene Historical Periodicals*
Tina Munda, Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Ciril Bohak, Vojko Gorjanc and Darja Fišer

12:00–12:30 *Automatic Metrical Scansion of Poetry in a Low-Resource Setting*
Pablo Ruiz Fabo, Anxo Alonso Pérez, Pablo Rodríguez Fernández and Pablo Gamallo

12:30–14:00 **Lunch break**

14:00–15:30 **Oral presentations**
Chair: Joanna Blochowiak

14:00–14:30 *SACRED: A Faithful Annotated Multimedia Multimodal Multilingual Dataset for Classifying Connectedness Types in Online Spirituality*
Qinghao Guan, Yuchen Pan, Donghao Li, Zishi Zhang, Yiyang Chen, Lu Li, Flaminia Canu, Emilia Volkart and Gerold Schneider

14:30–15:00 *Integrating Knowledge Graphs and Multilingual Scholarly Corpora for Domain-Adaptive LLMs in SSH*
Adam Faci, Alessio Miaschi, Anne Combe, Pascal Cuxac, Francesca Frontini, Nicolas Larrousse and Stéphane Pouyllau

15:00–15:30 *From One-Hot to Semantic Encoding: Entity Embedding for Small and Heterogeneous Digital Humanities Datasets*
Isabelle Gribomont

15:30–16:30	**Coffe break and poster session**
15:30–16:30	*Design and Methodological Architecture of a Multilingual Corpus of Interpreter-mediated Public Service Telephone Interactions* Raquel Lazaro Gutierrez, Daniel López Padilla, Jorge Rico, María José Vilella Sánchez and Fernando Manuel Espinoza-Cuadros
15:30–16:30	*Toward Responsible and Epistemically Grounded Multilingual LLMs for Computational Social Science and Humanities* Wajdi Zaghouani
15:30–16:30	*Automatic Evaluation of Multiple-Choice Items for Reading Comprehension: Effects of Question and Distractor Categories* John S. Y. Lee, Yin Poon, Shunjie Wang and Kai Wah Chu
15:30–16:30	*Reflexive Research with LLMs: Considering the Positionality of Users and Systems* Eleanor L.T. Smith, Luis Morgado da Costa and Antske Fokkens
15:30–16:30	*Small Can Be Beautiful in LLMs for SSH: a Case for Bulgarian* Kiril Simov, Nikolay Paev, Petya Osenova, Teodor Valchev and Stefan Marinov
15:30–16:30	*A Multimodal LLM-Based Nutrition Label for Analyzing Social Media Feed Exposure* Tim Gollub, Armin Heidari, Cem Ertürkan and Benno Stein
15:30–16:30	*Charting the European LLM Benchmarking Landscape: A New Taxonomy and Registry* Spela Vintar, Mojca Brglez, Taja Kuzman Pungeršek and Nikola Ljubešić
15:30–16:30	*Cross-Lingual Abstractive Keyphrase Generation for Historical Newspapers* Simon Clematide, Jenifer L. Meyer, Juri Opitz, Maud Ehrmann and Kaspar Beelen
15:30–16:30	*LLM Probe: Evaluating LLMs for Low-Resource Languages* Hailay Kidu Teklehaymanot, Wolfgang Nejdl and Gebrearegawi Gebremariam
16:30–17:30	**General roundtable** Chair: Nikola Ljubešić
17:30–18:00	**Concluding remarks and next steps** Chair: German Rigau Claramunt

State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?

Taja Kuzman Pungeršek[*], **Peter Rupnik**[*], **Ivan Porupski**[*], **Vuk Dinić**[*], **Nikola Ljubešić**[*†‡]
[*]Jožef Stefan Institute;
[†]Faculty of Computer and Information Science, University of Ljubljana;
[‡]Institute of Contemporary History;
Ljubljana, Slovenia
{taja.kuzman, peter.rupnik, ivan.porupski, vuk.dinic, nikola.ljubesic}@ijs.si

Abstract

Until recently, fine-tuned BERT-like models provided state-of-the-art performance on text classification tasks. With the rise of instruction-tuned decoder-only models, commonly known as large language models (LLMs), the field has increasingly moved toward zero-shot and few-shot prompting. However, the performance of LLMs on text classification, particularly on less-resourced languages, remains under-explored. In this paper, we evaluate the performance of current language models on text classification tasks across several South Slavic languages. We compare openly available fine-tuned BERT-like models with a selection of open-weight and closed-source LLMs across three tasks in three domains: sentiment classification in parliamentary speeches, topic classification in news articles and parliamentary speeches, and genre identification in web texts. Our results show that LLMs demonstrate strong zero-shot performance, often matching or surpassing fine-tuned BERT-like models. Moreover, when used in a zero-shot setup, LLMs perform comparably in South Slavic languages and English. However, we also point out key drawbacks of LLMs, including less predictable outputs, significantly slower inference, and higher computational costs. Due to these limitations, fine-tuned BERT-like models remain a more practical choice for large-scale automatic text annotation.

Keywords: LLM evaluation, text classification, large language models, South Slavic languages, sentiment identification, topic classification, genre identification

1. Introduction

Until recently, the dominant approach for text classification tasks relied on fine-tuning BERT-like transformer models on thousands of manually-annotated training examples. Recently, however, the field has shifted with the development of instruction-tuned decoder-only transformer models. These models, also commonly referred to as large language models (LLMs), which were originally developed primarily for text generation tasks, have demonstrated remarkable capabilities across a broad range of natural language processing (NLP) tasks, including text classification (Kuzman et al., 2023; Huang et al., 2023).

In this paper, we focus on South Slavic languages, where research on text classification tasks included in our study has, until recently, been limited or even non-existent (Kuzman and Ljubešić, 2023; Mochtak et al., 2024; Kuzman and Ljubešić, 2025). We take a first step toward systematically evaluating the current state of the art for text classification in these languages. Our evaluation is based on three text classification tasks in three different domains for which manually-annotated test datasets in South Slavic languages and fine-tuned BERT-like classifiers are freely available: sentiment classification of parliamentary speeches, topic classification in news articles, topic classification in parliamentary speeches, and automatic genre identification in web texts. These tasks span different domains and language styles, allowing for a comprehensive analysis of the performance of transformer-based models on text classification tasks. Specifically, we compare the performance of openly available fine-tuned BERT-like models with the zero-shot capabilities of both open-weight and closed-source LLMs used via prompting.

An important aspect of our study is to examine whether the performance of multilingual models on South Slavic languages is on par with their performance on English. This question is particularly relevant given that the evaluated large language models have been predominantly pretrained and instruction-tuned on English data.

By evaluating various models on a selection of text classification tasks in English and various South Slavic languages, we set out to test the following two hypotheses that are based on previous experiments with fine-tuned BERT-like models and LLMs on automatic genre identification (Kuzman et al., 2023), news topic classification (Kuzman and Ljubešić, 2025) and sentiment analysis in parliamentary texts (Mochtak et al., 2025):

H1 Zero-shot prompting with instruction-tuned large language models (LLMs) can achieve

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 1–17
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Dataset	Lang	# Instances	# Labels	% Most and Least Frequent Label
Sentiment classification in parliamentary speeches				
ParlaSent-EN-test	EN	2600	3	40.8% (Neutral), 26.8% (Positive)
ParlaSent-HR-test	HR	1336	3	41.9% (Negative), 17.2% (Positive)
ParlaSent-SR-test	SR	1074	3	46.2% (Negative), 17.6% (Positive)
ParlaSent-BS-test	BS	190	3	47.9% (Negative), 14.7% (Positive)
Genre classification in web texts				
EN-GINCO	EN	272	8	23.5% (Information/Explanation), 0.4% (Legal)
X-GINCO-SL	SL	80	8	15% (Prose/Lyrical), 8.8% (Opinion/Argumentation)
X-GINCO-HR	HR	80	8	16.3% (Promotion), 7.5% (Instruction)
X-GINCO-MK	MK	80	8	15% (News), 1% (Opinion/Argumentation)
Topic classification in news articles				
IPTC-test-HR	HR	291	17	11.0% (Economy), 3.8% (Conflict, War and Peace)
IPTC-test-SL	SL	282	17	10.6% (Society), 3.2% (Conflict, War and Peace)
Topic classification in parliamentary speeches				
ParlaCAP-test-EN	EN	876	22	6.4% (Law and Crime), 2.1% (Culture)
ParlaCAP-test-HR	HR	869	22	8.5% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-SR	SR	874	22	7.1% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-BS	BS	824	22	10.4% (Other), 0.5% (Culture)

Table 1: Information on test datasets in English (EN), Croatian (HR), Serbian (SR), Bosnian (BS), Slovenian (SL), and Macedonian (MK).

results comparable to the use of BERT-like models fine-tuned on training data that are similar to the test data.

H2 The performance of LLMs used in a zero-shot setup on text classification tasks on South Slavic test datasets is comparable to the performance on English test datasets.

2. Related Work

After the introduction of transformer architectures, BERT (bidirectional encoder representations from transformers) models have achieved state-of-the-art results in text classification tasks, outperforming earlier non-neural approaches, such as support vector machines (SVMs). They have also demonstrated strong cross-lingual zero-shot capabilities in various classification tasks, including automatic genre identification (Kuzman and Ljubešić, 2023), news topic classification (Petukhova and Fachada, 2023; De Clercq et al., 2020), and sentiment classification (Mochtak et al., 2024). However, these models still require fine-tuning on a training dataset, developed during manual annotation campaigns that are time-consuming and costly.

Instruction-tuned decoder-only transformer models, commonly referred to as large language models (LLMs), have recently shown strong performance in a range of classification tasks, even in zero-shot prompting setups that require no training data (Kuzman et al., 2023; Ljubešić et al., 2024a; Huang et al., 2023; Kuzman Pungeršek et al., 2026). They have achieved promising results on various natural language processing tasks, including stance detection (Zhang et al., 2022), implicit hate speech categorization (Huang et al., 2023), news topic classification (Kuzman and Ljubešić, 2025), automatic genre identification (Kuzman et al., 2023), causal commonsense reasoning (Ljubešić et al., 2024b), and machine translation (Hendy et al., 2023). Due to their promising performance, researchers have even started using them as data annotators, either by generating text and labels (Meng et al., 2022) or by annotating pre-existing texts (Kuzman and Ljubešić, 2025; Kuzman Pungeršek et al., 2026). Despite the growing interest in this topic, the majority of evaluations of LLMs used in text classification tasks are limited only to English (Sun et al., 2023; Zhang et al., 2025; Kostina et al., 2025; Zhao et al., 2024). Systematic multilingual evaluations, especially which would include less-resourced languages such as those in the South Slavic group remain limited. Our work addresses this gap by providing a comparative evaluation of open-weight and closed-source LLMs with openly-available fine-tuned BERT-like models across four benchmark families comprising three diverse classification tasks and three different domains in South Slavic languages and English.

3. Benchmarks

The benchmarks (evaluation datasets) used in this study cover three text classification tasks, namely, sentiment identification, topic classification, and automatic genre identification, and three domains: parliamentary speeches, news articles and web texts. An overview of the datasets is provided in Table 1. The four benchmark families differ significantly in terms of language coverage, number of test instances, and label granularity.

The topic classification task is evaluated on two domains: 1) news articles, namely, the Croatian and Slovenian IPTC test datasets (Kuzman and Ljubešić, 2025), which comprise around 300 text instances per language, and 2) parliamentary speeches, namely, the Bosnian, Croatian, English and Serbian ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) that consist of approximately 820 to 880 instances per language. In the ParlaCAP benchmarks, an instance is a transcription of an utterance given by a parliamentary member in a parliamentary session.

The topic classification task involves the highest number of labels, that is, 17 news topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema[1] (IPTC, 2022), and 22 policy topic labels (21 major topics and a label *Other*) from the Comparative Agendas Project (CAP; Baumgartner et al., 2019) Master Codebook (Bevan, 2019).[2]

In contrast, the Bosnian, Croatian, English, and Serbian ParlaSent sentiment identification datasets (Mochtak et al., 2024; Mochtak et al., 2023) have a significantly lower granularity of labels, with only 3 categories. They are represented by the largest number of instances, ranging from 190 (Bosnian part) to 2600 (English part) sentence-level instances.

With 8 labels, the Croatian, English, Macedonian, and Slovenian GINCO genre datasets (Kuzman et al., 2023) represent a midpoint in label granularity among the four benchmark families. However, the genre identification task might be the most difficult one, as genre identification depends on the interpretation of full texts with the focus on author's purpose, the common function of the text, and the text's conventional form (Orlikowski and Yates, 1994). This complexity has also contributed to smaller test datasets in terms of the number of text instances, as manual annotation is more time-consuming. It is also important to note that, unlike the parliamentary datasets, the English portion of the genre datasets is not fully comparable to the South Slavic portions, which are label-balanced and contain fewer ambiguous instances. Nevertheless, the genre datasets remain valuable for evaluating model performance within each language.

All test datasets were manually annotated by annotators that are deemed reliable based on their satisfactory inter-annotator agreement, namely, Krippendorff's alpha (Krippendorff, 2018) values close to or above the 0.667 threshold for reliable annotation. To prevent large language models from incorporating the test datasets during their training phase, the test datasets are not publicly available, except for the ParlaSent benchmark family. Access to other datasets is granted on request from the corresponding authors. Further details on the test datasets are provided in Section A.1 of the Appendix.

4. Methodology

In this paper, we evaluate the main machine learning approaches that have recently been used for our selection of text classification tasks, with a focus on the comparison between the freely available fine-tuned BERT models and the open-weight and closed-source LLMs.[3] The models are evaluated on four families of test datasets that comprise South Slavic languages. The performance of the models is evaluated based on the micro-F1 and macro-F1 metrics, which enable assessment of the model performance at both the instance and label levels, respectively.

The following machine learning models are included in the evaluation:

- **dummy classifier**: a dummy classifier that predicts the most frequent class in the training data. To allow comparison, the dummy classifiers were trained on the same datasets that were used for fine-tuning the BERT-like models, mentioned below.

[1] https://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

[2] https://www.comparativeagendas.net/pages/master-codebook

[3] The code for the model evaluation and analysis of results is available at https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic.

- **fine-tuned BERT-like classifiers**: in our study, we evaluate previously developed openly accessible multilingual fine-tuned BERT-like models that have been fine-tuned for the respective task, namely, the XLM-R-ParlaSent (Rupnik et al., 2023; Mochtak et al., 2024) model for sentiment identification in parliamentary texts, the X-GENRE classifier (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d, 2023) for automatic genre identification, the IPTC News Topic classifier (Kuzman and Ljubešić, 2025; Kuzman and Ljubešić, 2025) for news topic classification, and the ParlaCAP classifier (Kuzman Pungeršek et al., 2026; Kuzman Pungeršek and Ljubešić, 2025) for topic classification in parliamentary speeches. The XLM-R-ParlaSent and the ParlaCAP models are based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on parliamentary proceedings in 30 European languages (Mochtak et al., 2024). The XLM-R-ParlaSent model was fine-tuned on 13 thousand instances from the ParlaSent sentiment training dataset (Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian; Mochtak et al., 2024), while the ParlaCAP model was fine-tuned on the ParlaCAP-train dataset (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026) that comprises around 30 thousand speeches from parliamentary debates annotated with CAP topic labels, originating from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The X-GENRE classifier is based on the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a) in English and Slovenian; while the IPTC News Topic classifier is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) that was fine-tuned on the EMMediaTopic dataset (Kuzman and Ljubešić, 2024c) in Catalan, Croatian, Greek, and Slovenian. All fine-tuned models use the same classes as the test datasets used in our study.

- **open-weight and closed-source large language models**: we use closed-source OpenAI models, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`; OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`; OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`; OpenAI, 2025); a closed-source Gemini 2.5 Flash model (Comanici et al., 2025) by Google DeepMind; a closed-source Mistral Medium 3.1 model (`mistral-medium-2508`; Mistral AI, 2025) by Mistral AI; and four open-weight models, namely, the Meta LLaMA 3.3 model (Meta, 2024), the Gemma 3 model (Gemma Team et al., 2025), the Qwen 3 model (Yang et al., 2025), and the DeepSeek-R1-Distill model (`DeepSeek-R1-Distill-Qwen-14B`; Guo et al., 2025). It is important to note that while the LLaMA model was pretrained on a web text collection in various languages, it is said to support only 8 languages, namely English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai (Meta, 2024). The DeepSeek-R1-Distill model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that provides support for more than 29 languages – not including South Slavic languages though. In contrast, the Gemma 3 model is reported to support over 140 languages (Gemma Team et al., 2025), and the Qwen 3 model was pretrained on 119 languages (Yang et al., 2025). While closed-source models are said to be massively multilingual, with Gemini 2.5 models being pretrained on over 400 languages (Comanici et al., 2025), details on their language coverage are very limited.

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). OpenAI models were used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[4] that provides a unified API access to various closed-source models. To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. More details on the models and their implementation, including information on the availability of openly available models and fine-tuning datasets, are provided in Section A.2 of the Appendix.

All instruction-tuned LLMs are used in a zero-shot prompting setup, meaning that they receive only a task description and label definitions. All prompts and label definitions are written in English, while the instances are provided in the original languages (English or South Slavic). Changing the language of the prompt could introduce an additional factor that affects performance. In the presented experiments, our goal is to assess model performance on a specific task and language, rather than to evaluate their instruction-following abilities across different languages. The models are instructed to output a label, represented by a digit. The same prompt per benchmark family is used for all LLMs.

[4] `https://openrouter.ai/`

Prompts are provided in Figure 4 in Section A.2 of the Appendix.

5. Results

In this section, we evaluate the performance of the fine-tuned BERT-like models and the instruction-tuned LLMs on a selection of text classification tasks that include test datasets in South Slavic languages. First, in Section 5.1, we provide results on the four benchmark families with a focus on hypothesis H1, which expects that zero-shot prompting with LLMs can provide performance that is comparable to that of fine-tuned BERT-like models. In Section 5.2, we compare in more detail the performance of the closed-source and open-weight LLMs on the three text classification tasks, which is followed by a discussion on the advantages and limitations of LLMs for data annotation based on text classification tasks (Section 5.3). Lastly, in Section 5.4, we compare the performance of LLMs on English test datasets with their performance on South Slavic datasets, addressing hypothesis H2, which presumes that the available multilingual LLMs perform similarly on South Slavic languages as on English.

Model	Rank	Rank (EN)	Rank (South Slavic)
GPT-5	2.29	1.33	2.55
GPT-4o	2.36	2.00	2.45
Fine-Tuned BERT-Like Model	3.21	4.67	2.82
Gemini 2.5 Flash	3.50	3.33	3.55
Mistral Medium 3.1	5.36	5.00	5.45
Gemma 3	5.71	5.67	5.73
LLaMA 3.3	6.00	6.67	5.82
Qwen 3	7.43	7.00	7.55
GPT-3.5-Turbo	8.79	9.00	8.73
DeepSeek-R1-Distill	10.00	10.00	10.00

Table 2: Comparison of models based on their average rank (1 = best-performing, 10 = worst-performing) across all test datasets (first column), and averaged across English (second column) or South Slavic (third column) test datasets.

5.1. State of the Art in Text Classification Tasks

Figure 1 provides the results of model evaluation on our selection of text classification tasks. A consistent pattern emerges across all four benchmark families: LLMs, when used in a zero-shot prompting setup, achieve some of the highest scores. As shown in Table 2, which compares model rankings across tasks, LLMs achieve first place more often on average than the fine-tuned BERT-like model.

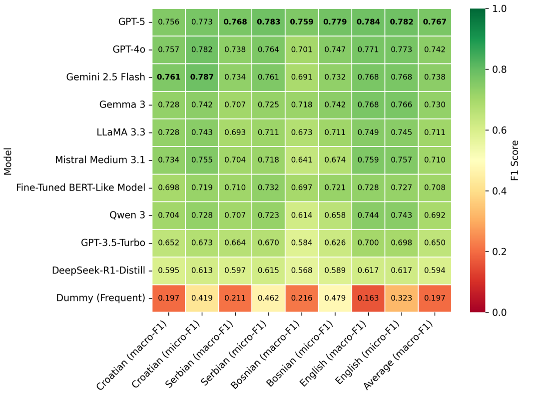
Figure 1a shows that both open-weight and closed-source LLMs, used in a zero-shot prompting setup on the sentiment identification task, achieve performance that is comparable or even significantly higher than that of a fine-tuned BERT-like model trained on a large manually-annotated sentiment dataset. The only models that consistently perform worse than the fine-tuned BERT-like model are GPT-3.5-Turbo and DeepSeek-R1-Distill. Sentiment classification appears broad enough that more potent LLMs can interpret label definitions effectively without task-specific fine-tuning, reducing the benefit of additional training.

In contrast, fine-tuned BERT-like models outperform most LLMs on automatic genre identification and topic classification tasks. These tasks depend on predefined label sets based on specific guidelines, and the strong performance of fine-tuned BERT-like models indicates that domain-specific fine-tuning on labelled data still offers an advantage over the general knowledge leveraged by LLMs in zero-shot setups. This advantage is particularly clear in genre identification for South Slavic texts, where the fine-tuned BERT-like model significantly outperforms LLMs. The likely reason for the fine-tuned model's very strong performance on South Slavic genre datasets is the curated nature of the test data – more challenging examples were removed before and during manual annotation, unlike in the English genre test dataset where the instances were randomly sampled from an English web corpus. Nevertheless, despite this limitation, the South Slavic test dataset remains valuable for comparing the performance of LLMs.

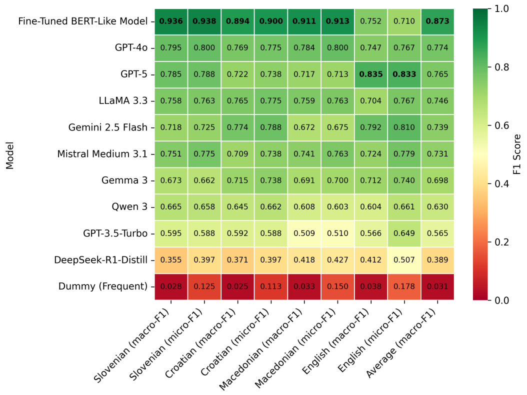
To conclude, since some LLMs used in a zero-shot prompting setup achieve higher or comparable results to fine-tuned BERT-like models across all classification tasks and languages, as shown in Table 2, we can confirm hypothesis H1, which proposed that zero-shot prompting with LLMs can perform comparably to fine-tuned BERT-like models.

5.2. Comparison of Large Language Models

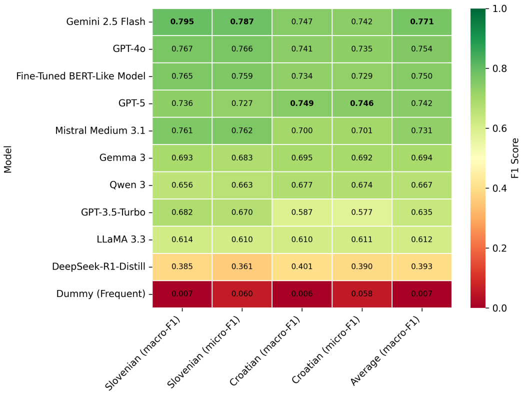
Figure 2 shows the performance of open-weight and closed-source LLMs, used via prompting, on the tasks of sentiment identification, automatic genre identification, news topic classification, and parliamentary topic classification. The DeepSeek-R1-Distill model is not included in the comparison, as it performs significantly worse than the other



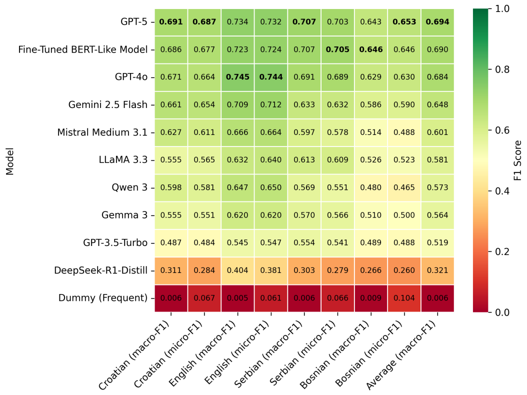
(a) Sentiment classification.



(b) Automatic genre identification.



(c) News topic classification.



(d) Parliamentary topic classification.

Figure 1: Micro-F1 and macro-F1 scores across models and languages on the test datasets for sentiment classification (Figure 1a), automatic genre identification (Figure 1b), and topic classification on news (Figure 1c) and parliamentary speeches (Figure 1d).

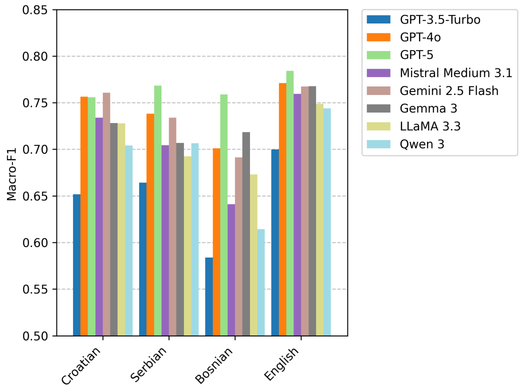
models, as shown in Figure 1.

While different models perform best across different languages and test datasets, a clear trend emerges: the top-performing models across all four benchmark families are the closed-source GPT-4o and GPT-5 from OpenAI, along with Gemini 2.5 Flash. Although GPT-5 is newer and reportedly more powerful, it does not outperform GPT-4o on all benchmarks. Among open-weight models, Gemma 3 generally achieves the best results in sentiment identification (Figure 2a) and news topic classification (Figure 2c). For automatic genre identification (Figure 2b) and parliamentary topic classification (Figure 2d), rankings of open-weight models vary by language. Overall, the weakest performance is observed with the older closed-source GPT-3.5-Turbo model, highlighting the rapid progress in both open-weight and closed-source model development.

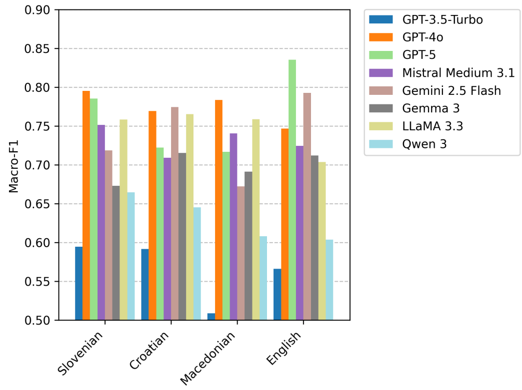
5.3. Advantages and Disadvantages of LLMs

A clear advantage of LLMs is that they do not require manually-annotated training data for specific tasks, yet still achieve strong performance when provided only with task instructions and brief label descriptions. However, these models are significantly more computationally expensive than fine-tuned BERT-like models. While closed-source models deliver the best performance, as shown in previous sections, they come with several limitations: they are costly to use, their architectures and pretraining data are not publicly disclosed, and access through APIs hinders reproducibility, in contrast to open-weight LLMs and fine-tuned BERT-like models.

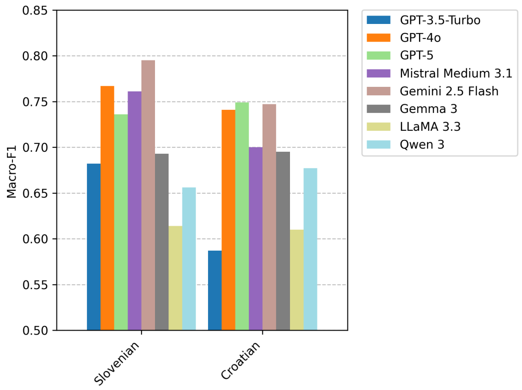
What is more, the inference speed of all LLMs is significantly slower than that of a fine-tuned BERT-like model. As shown in Figure 3, the fine-tuned BERT-like model achieves one of the highest macro-F1 scores on the topic classification task for parliamentary speeches, while maintaining a very low inference time of just 0.02 seconds per



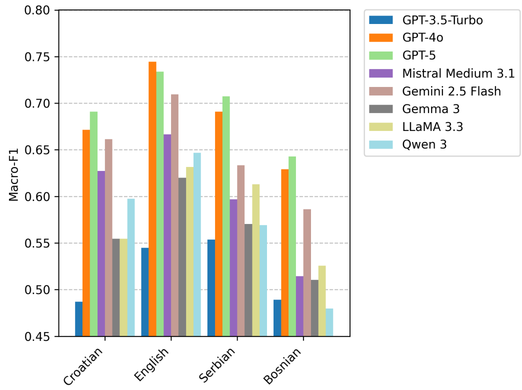
(a) Sentiment classification.



(b) Automatic genre identification.



(c) News topic classification.



(d) Parliamentary topic classification.

Figure 2: Comparison of LLMs used in a zero-shot prompting setup on sentiment identification (Figure 2a), automatic genre identification (Figure 2b), and topic classification on news (Figure 2c) and parliamentary speeches (Figure 2d).

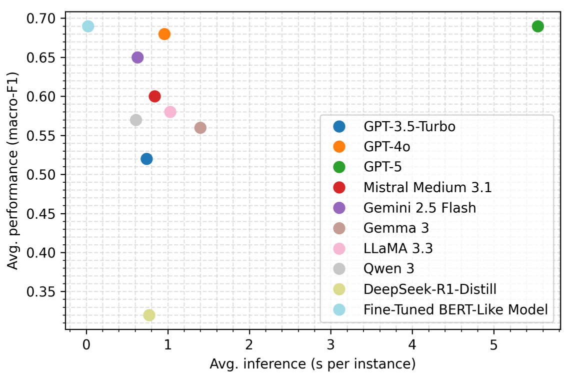


Figure 3: Comparison of models on the parliamentary topic classification based on their inference speed (seconds per instance) and performance (macro-F1 scores), both averaged across all four languages.

instance. In contrast, most LLMs have inference times between 0.6 and 1.4 seconds per instance, making them three to seven times slower for annotating the same dataset. The slowest model, GPT-5, takes 5.5 seconds per instance, which renders it impractical for large-scale automatic annotation of text collections. In this regard, fine-tuned BERT-like models offer a key advantage due to their lower computational cost and higher inference speed. Moreover, they can be trained on training data that is annotated by LLMs using the recently introduced LLM teacher-student paradigm (Kuzman and Ljubešić, 2025), which considerably reduces the effort needed to develop task-specific models.

Another limitation of LLMs, as revealed by the experiments, is their occasional deviation from the defined label set. This issue was especially noticeable in topic classification and, to a lesser extent, in genre identification. The highest rate of label hallucination was found in the DeepSeek-R1-Distill model, which produced non-existing labels for 8% of instances in the news topic test datasets and 4% in the genre test dataset. Similar issues were also observed, though much less frequently (less than 1%), with the LLaMA 3.3, Gemma 3, Qwen 3 and Mistral Medium 3.1 models. In contrast, fine-tuned

BERT-like models do not suffer from this issue, as they output probabilities for the predefined classes.

Model	Difference (sentiment)	Difference (topic)
GPT-5	0.02	0.05
GPT-4o	0.04	0.08
Gemini 2.5 Flash	0.04	0.08
Gemma 3	0.05	0.07
LLaMA 3.3	0.05	0.07
Mistral Medium 3.1	0.07	0.09
Qwen 3	0.07	0.10
GPT-3.5-Turbo	0.07	0.03

Table 3: Difference between model performance in macro-F1 scores obtained on sentiment and topic classification in parliamentary texts on English versus the average macro-F1 scores on South Slavic languages.

5.4. Performance on English versus on South Slavic languages

The sentiment identification ParlaSent and the topic classification ParlaCAP benchmark families comprise test datasets in South Slavic languages and English that were constructed with the same methodology. Thus, they also allow for a comparison of the performance of the LLMs on English, a highly resourced language, with South Slavic languages, which are significantly less represented in the pretraining and instruction-tuning datasets used to develop large language models.

As shown in Table 3, the differences in macro-F1 scores between English and the average of macro-F1 scores for South Slavic languages are relatively small for sentiment identification, ranging from 2 to 7 points. For topic classification, the performance gap is slightly larger, ranging from 3 to 10 points. This is likely due to the increased difficulty of the task, which involves greater label granularity: 22 labels compared to just 3 in sentiment classification. These findings partially confirm hypothesis H2, which stated that LLMs, when used in a zero-shot setup, perform comparably on text classification tasks in South Slavic languages as they do on English.

Interestingly, even the open-weight LLaMA 3.3 model – reported to support only eight languages, excluding the South Slavic group – does not show a substantial performance drop when applied to South Slavic languages compared to English.

6. Conclusion

In this paper, we evaluated how well current machine learning technologies handle text classification tasks in South Slavic languages. We compared fine-tuned BERT-like models with decoder-only generative large language models (LLMs) that are used in a zero-shot prompting setup across three tasks and three text domains: sentiment classification in parliamentary texts, news topic classification, topic classification in parliamentary texts, and automatic genre identification on web texts.

Our results show that LLMs used with prompting, where only a brief task description and labels were provided, achieved strong results across all tasks and languages, particularly the closed-source GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025) and Gemini 2.5 Flash (Comanici et al., 2025) models. The performance of LLMs is comparable or higher than that of fine-tuned BERT-like models specialized for the tasks. On the sentiment identification task, most open-weight and closed-source LLMs outperformed the fine-tuned model, demonstrating strong general knowledge of the notion of sentiment. For genre and topic classification, however, fine-tuning BERT-like models remains beneficial, as these tasks rely on predefined label sets and fine-tuning aligns the models more closely with the task requirements.

Interestingly, LLMs perform similarly in English and South Slavic languages, with rather minor drops in micro- and macro-F1 scores, namely a drop of 2 to 7 points in terms of macro-F1 scores on sentiment classification, and a slightly higher drop from 3 to 10 points on topic classification in parliamentary texts. This suggests that the gap in multilingual performance is smaller than expected, even for open-weight models not explicitly dedicated to these languages.

Although large language models offer impressive zero-shot performance and reduce the need for annotated data, they come with higher computational costs and are more prone to producing invalid labels. Moreover, their inference speed is at least three times slower than that of the fine-tuned BERT-like models. Thus, their use in use cases with extensive data to be processed, such as automatic enrichment of large corpora with text categories, remains impractical due to their high computational demands. In contrast, fine-tuned BERT-like models are more computationally efficient and can be better tailored to the specific characteristics of a task and its domain. They remain a practical and reliable choice for text classification tasks, especially when computational resources are limited, high inference speed is desired or output reliability is critical. Moreover, it is possible to combine the strengths of both approaches, as proposed by the LLM teacher-student paradigm (Kuzman and Ljubešić, 2025): LLMs can be used to automatically annotate training data, reducing the need for costly and time-consuming manual annotation, while fine-tuned BERT-like models can then be trained on these datasets.

This study represents only an initial step to systematically benchmark text classification performance in South Slavic languages. Although our evaluation includes four diverse benchmark families, some of the test datasets remain relatively small. Future work will aim to increase dataset sizes, include more South Slavic languages and dialects, and introduce additional classification tasks. As new large language models continue to emerge rapidly, it will also be important to establish ongoing evaluations to track whether their performance continues to improve, particularly on South Slavic languages. Importantly, this study only evaluated the performance of LLMs in a zero-shot prompting setup. In future work, we plan to extend the evaluation to include few-shot prompting and fine-tuning on training data. To support further research and facilitate reproducibility, we have made all code, evaluation scripts, and results publicly available.[5] Additionally, we have developed an interactive dashboard[6] that enables users to explore the results of our evaluation of large language models on the tasks presented in this paper, as well as on additional commonsense reasoning tasks. The dashboard is an ongoing project that monitors the performance of newly released large language models on South Slavic languages and dialects. In future work, we plan to expand both the range of tasks and the coverage of South Slavic languages and dialects included in the dashboard.

7. Ethical Considerations and Limitations

Our study has several limitations that should be acknowledged. First, while we aimed to include a broad set of South Slavic languages, some – most notably Bulgarian – were not covered in our experiments. We assume that the performance on Bulgarian would be similar to that observed for Macedonian, given their close linguistic proximity, or the results for Bulgarian could be slightly better, as Macedonian is comparatively more low-resourced. Moreover, due to the high computational cost of evaluating the LLMs on all the test datasets and the financial cost associated with the use of closed-source models, each model was evaluated on each dataset only once. This setup prevents us from fully estimating the variance of the results, however, based on our preliminary experiments, we expect this variance to be relatively small. Finally, the scope of our evaluation remains limited in terms of test datasets, language coverage and tasks. Expanding the range of benchmarks would allow for a more comprehensive validation of our findings, particularly regarding the hypothesis that LLMs can perform on par with fine-tuned BERT-like models across diverse natural language understanding tasks, languages and language varieties.

8. Acknowledgements

We would like to thank the developers of the llm.ijs.si service (Marić et al., 2025) for establishing the LLM inference platform deployed at the Jožef Stefan Institute, which provided convenient access to the open-weight large language models used in this study. We also thank the annotators of the test datasets for their diligence and the time devoted to manual annotation, which resulted in the high-quality evaluation datasets used in this work. Lastly, we would like to thank the CLASSLA knowledge centre for South Slavic languages and the Slovenian CLARIN.SI infrastructure for their valuable support.

This work was supported in part by the projects "Spoken Language Resources and Speech Technologies for the Slovenian Language" (Grant J7-4642), "Large Language Models for Digital Humanities" (Grant GC-0002), the research programme "Language Resources and Technologies for Slovene" (Grant P6-0411), all funded by the ARIS Slovenian Research and Innovation Agency, and the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070), co-funded by the Kliping d.o.o. agency. The authors acknowledge the OSCARS project – and its ParlaCAP cascading grant project –, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This research is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

[5] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

[6] See the CLASSLA LLM Evaluation Dashboard for South Slavic Languages at https://www.clarin.si/classla-llm-dashboard/.

Marta Bañón, Miquel Esplà-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation*, pages 301–302.

Frank R Baumgartner, Christian Breunig, and Emiliano Grossman. 2019. *Comparative Policy Agendas: Theory, Tools, Data*. Oxford University Press.

Shaun Bevan. 2019. Gone Fishing. *Comparative Policy Agendas: Theory, Tools, Data*, pages 17–34.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Orphée De Clercq, Luna De Bruyne, and Véronique Hoste. 2020. News topic classification as a first step towards diverse news recommendation. *Computational Linguistics in the Netherlands Journal*, 10:37–55.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2025. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

IPTC. 2022. Groups of NewsCodes. https://iptc.org/standards/newscodes/groups/#descrncd. Accessed October 29, 2024.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Lancaster University.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage Publications.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey. *Language Resources and Evaluation*, pages 1–34.

Taja Kuzman and Nikola Ljubešić. 2025. LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, 13:35621–35633.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In *Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Taja Kuzman Pungeršek, Peter Rupnik, Daniela Širinić, and Nikola Ljubešić. 2026. Supercharging Agenda Setting Research: The ParlaCAP Dataset of 28 European Parliaments and a Scalable Multilingual LLM-Based Classification. *arXiv preprint arXiv:2602.16516*.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024a. DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Ivan Vulić, Fabian Schmidt, and Goran Glavaš. 2024b. JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 209–219.

Nikola Marić, Boshko Koloski, Damjan Demšar, Jan Jona Javoršek, and Sašo Džeroski. 2025. Running large language models locally: design and operational insights with llm.ijs.si. *International conference AI for science 2025: Ljubljana, Slovenia, 22.09.2025-26.09.2025*, page 77.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Meta. 2024. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: June 26, 2025.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv preprint arXiv:2004.03705*.

Mistral AI. 2025. Medium is the new large. https://mistral.ai/news/mistral-medium-3. Accessed: October 10, 2025.

Michal Mochtak, Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2025. Parlasent: mapping sentiment in political discourse with large language models. *Political Research Exchange*, 7(1):2508377.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036.

OpenAI. 2023. ChatGPT General FAQ. https://help.openai.com/en/articles/6783457-chatgpt-general-faq. Accessed: June 26, 2025.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed September 11, 2024.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: October 10, 2025.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Alina Petukhova and Nuno Fachada. 2023. MN-DS: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5):74.

Qwen Team. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen Team. 2024b. Qwen2.5: A party of foundation models. Accessed: June 26, 2025.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. 2025. Pushing the limit of LLM capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528.

Hang Zhao, Qile P Chen, Yijing Barry Zhang, and Gang Yang. 2024. Advancing Single and Multi-task Text Classification through Large Language Model Fine-tuning. *arXiv preprint arXiv:2412.08587*.

10. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and others. 2024. *Multilingual comparable corpora of parliamentary debates ParlaMint 4.1*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1912.

Kuzman, Taja and Ljubešić, Nikola. 2023. *Multilingual text genre classification model X-GENRE*. Hugging Face. PID https://doi.org/10.57967/hf/0927.

Kuzman, Taja and Ljubešić, Nikola. 2024a. *English-Slovenian text genre dataset X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1960.

Kuzman, Taja and Ljubešić, Nikola. 2024b. *Genre-enriched web corpora MaCoCu-Genre*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1969.

Kuzman, Taja and Ljubešić, Nikola. 2024c. *Multilingual IPTC Media Topic dataset EMMediaTopic 1.0*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1991.

Kuzman, Taja and Ljubešić, Nikola. 2024d. *Multilingual text genre classification model X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1961.

Kuzman, Taja and Ljubešić, Nikola. 2025. *Multilingual IPTC News Topic Classifier*. Hugging Face. PID https://doi.org/10.57967/hf/4709.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2025. *Multilingual ParlaCAP model for CAP Topic Classification in Parliamentary Speeches*. Hugging Face. PID https://doi.org/10.57967/hf/6684.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2026. *Multilingual training dataset for CAP policy topic classification ParlaCAP-train*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2093.

Ljubešić, Nikola and Rupnik, Peter and van Noord, Rik. 2023. *Multilingual parliamentary model XLM-R-parla*. Hugging Face. PID https://doi.org/10.57967/hf/6717.

Mochtak, Michal and Rupnik, Peter and Meden, Katja and Ljubešić, Nikola. 2023. *The multilingual sentiment dataset of parliamentary debates ParlaSent 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1868.

Rupnik, Peter and Ljubešić, Nikola and Mochtak, Michal. 2023. *Multilingual parliament sentiment regression model XLM-R-ParlaSent*. Hugging Face. PID https://doi.org/10.57967/hf/6718.

A. Appendix

A.1. Benchmarking Datasets

In this section, we provide additional information on the datasets used for benchmarking the models on sentiment identification, topic classification, and genre identification tasks in this study.

ParlaSent test datasets for sentiment classification in parliamentary speeches include Croatian, Serbian, Bosnian, and English data from the multilingual sentiment dataset of parliamentary debates ParlaSent 1.0 (Mochtak et al., 2024, Mochtak et al., 2023).[7] The dataset comprises sentences that were randomly sampled from Croatian, Serbian, Bosnian and British parliamentary corpora and manually annotated with reported inter-annotator agreement ranging from 0.53 to 0.66 in Krippendorff's alpha (Krippendorff, 2018). The annotation involved a more granular six-level sentiment polarity scale that has been mapped to a three-level sentiment polarity scale which we use in our experiments: negative (0), neutral (1), and positive (2).

GINCO datasets for automatic genre identification comprise the English EN-GINCO dataset (Kuzman et al., 2023) and a multilingual X-GINCO dataset from the AGILE benchmark for Automatic Genre Identification.[8] The test instances were sampled from the enTenTen20 English web corpus (Jakubíček et al., 2013) and the MaCoCu multilingual web corpus collection (Bañón et al., 2022). They were manually annotated by experts with a background in linguistics and computational linguistics who had experience with previous genre annotation campaigns (Kuzman et al., 2022, 2023) where they reached an acceptable inter-annotator agreement of 0.71 in nominal Krippendorff's alpha (Krippendorff, 2018). While the X-GINCO dataset comprises numerous European languages, for the purposes of this study, we focus on three South Slavic languages: Croatian, Macedonian, and Slovenian. The test datasets use the X-GENRE annotation schema (Kuzman et al., 2023) that includes the following genre labels: *Information/Explanation*, *News*, *Instruction*, *Opinion/Argumentation*, *Forum*, *Prose/Lyrical*, *Legal* and *Promotion*. While EN-GINCO and X-GINCO datasets have been annotated by the same annotator with the same schema, one should note that there are important differences between them in terms of their construction – the English test dataset was sampled randomly from the web corpus, resulting in an unbalanced label distribution, while the X-GINCO datasets were curated with more deliberate interventions to ensure a balanced label distribution and a more controlled sampling process. Consequently, the X-GINCO datasets comprise fewer ambiguous instances and could be regarded as an easier test dataset.

IPTC News Topic test datasets (Kuzman and Ljubešić, 2025) comprise Croatian and Slovenian news articles extracted from the MaCoCu-Genre web corpus collection (Kuzman and Ljubešić, 2024b) and manually annotated by one annotator. The reliability of the annotator was confirmed on a sample of data that was annotated by an additional annotator. The two annotators reached an acceptable inter-annotator agreement of 0.73 in nominal Krippendorff's alpha (Krippendorff, 2018). Text instances are annotated with 17 topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema, developed by the International Press Telecommunications Council (IPTC) (IPTC, 2022). The datasets are more or less balanced by labels.

ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) comprise parliamentary speeches in Bosnian, Croatian, English, and Serbian, sourced from the ParlaMint 4.1 dataset (Erjavec et al., 2024; Erjavec et al., 2025). These speeches were annotated by a single expert annotator using the 21 CAP categories from the official CAP schema (Baumgartner et al., 2019), along with an additional *Other* label. The datasets are approximately balanced across labels. To assess the annotation quality, the Croatian dataset was independently annotated by two additional annotators. Inter-annotator agreement between the expert annotator and the others ranged from 0.62 to 0.68 in Krippendorff's alpha, which is around the threshold of 0.67 typically considered acceptable for annotation reliability (Krippendorff, 2018).

A.2. Models

In the following subsections, we outline the models included in the evaluation – the fine-tuned BERT-like classifiers (Section A.2.1) and the open-weight and closed-source LLMs (Section A.2.2).

A.2.1. Fine-Tuned BERT-like Models

BERT (bidirectional encoder representations from transformers) deep neural models (Kenton and Toutanova, 2019) have revolutionized the field of

[7] Available in the CLARIN.SI repository at http://hdl.handle.net/11356/1585 and in the Hugging Face repository at https://huggingface.co/datasets/classla/ParlaSent.

[8] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

natural language processing (NLP), outperforming non-neural methods across various NLP tasks. They have a more complex and computationally expensive architecture featuring transformers – neural networks with self-attention mechanisms (Vaswani et al., 2017) – that significantly improves the efficiency of training the models on massive text data. Similarly to decoder-only transformer models, BERT models are pretrained on massive amounts of texts, possibly in multiple languages, which establishes their ability to encode the words and texts in high-dimensional vector spaces (Minaee et al., 2020) and enables their application even across languages in a zero-shot classification scenario. To develop BERT-based classifiers, the pretrained models are trained, that is, fine-tuned, on a training dataset comprising text instances annotated with labels. In our study, we evaluate openly-accessible multilingual fine-tuned BERT-like models that have already been developed in recent related research. Namely, we evaluate the following models:

- **IPTC News Topic classifier**[9] (Kuzman and Ljubešić, 2025) is a multilingual fine-tuned BERT-like model for news topic classification according to the top-level IPTC NewsCodes schema (IPTC, 2022). The model is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on 15,000 training text instances from the EMMediaTopic[10] dataset (Kuzman and Ljubešić, 2024c). The training dataset contains news article instances in four languages: Catalan, Croatian, Greek, and Slovenian. The training dataset was annotated using an LLM that was shown to achieve annotation reliability comparable to that of human annotators (Kuzman and Ljubešić, 2025). This approach is based on the novel methodology that uses the LLM teacher-student pipeline to develop BERT-like classifiers in the absence of manually-annotated training data.

- **XLM-R-ParlaSent** (Rupnik et al., 2023; Mochtak et al., 2024) is a domain-specific multilingual transformer model for sentiment identification in parliamentary texts. It is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on 1.72 billion words from parliamentary proceedings in 30 European languages. To develop the XLM-R-ParlaSent model,[11] the pretrained XLM-R-Parla model was fine-tuned on the ParlaSent sentiment training dataset (Mochtak et al., 2024; Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian). The training dataset[12] comprises 13,000 instances sampled from parliamentary proceedings and manually annotated with sentiment labels.

- **ParlaCAP classifier**[13] (Kuzman Pungeršek and Ljubešić, 2025; Kuzman Pungeršek et al., 2026) is a domain-specific multilingual transformer model for topic classification in parliamentary texts based on the CAP schema (Baumgartner et al., 2019). As the XLM-R-ParlaSent model, this model is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023; Mochtak et al., 2024). The XLM-R-parla model was then fine-tuned on the ParlaCAP-train dataset[14] (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026). The training dataset comprises around 30 thousand speeches from parliamentary debates from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The training dataset was annotated with the CAP categories by a GPT-4o (OpenAI, 2024) model used in a zero-shot prompting setup, following the LLM teacher-student framework (Kuzman and Ljubešić, 2025). Based on the inter-annotator agreement, calculated on a sample that was annotated by three human annotators and the LLM annotator, the agreement between the LLM and the human annotators was comparable to the agreement between the human annotators themselves. This indicates that the LLM performs as reliably as human annotators on this task, supporting its use for annotating the training data.

- **X-GENRE classifier** (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d) is a multilingual fine-tuned BERT-like model for automatic genre identification.[15] The model is based on

[9]The IPTC News Topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier.

[10]The EMMediaTopic training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/1991.

[11]The XLM-R-ParlaSent model is accessible in the Hugging Face repository at https://huggingface.co/classla/xlm-r-parlasent.

[12]The ParlaSent training and test datasets are freely available in the CLARIN.SI repository at http://hdl.handle.net/11356/1868.

[13]The ParlaCAP topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/ParlaCAP-Topic-Classifier.

[14]The ParlaCAP-train training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/2093.

[15]The X-GENRE classifier is freely available in the

the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a), which contains 1,772 text instances in Slovenian and English, manually-annotated with genre labels from the X-GENRE schema (Kuzman et al., 2023).

A.2.2. Instruction-Tuned Large Language Models

As the BERT models, decoder-only large language models are based on a transformer deep neural architecture and are pretrained on massive text collections. However, while the development of fine-tuned BERT-like classifiers necessitates large amounts of annotated training data, recent advances in the field have shown that the instruction-tuned LLMs are capable of text classification in a zero-shot or few-shot prompting setups which do not require any training data. We assess the performance of the following large language models:

- **OpenAI models**, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`) (OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`) (OpenAI, 2025). These closed-source instruction-tuned LLMs were developed by OpenAI. OpenAI states that the models are trained on large multilingual web corpora, however, specific details about the training data, procedures, and architecture are not publicly known.

- **Gemini 2.5 Flash model** (Comanici et al., 2025) is a closed-source multilingual and multimodal instruction-tuned LLM by Google DeepMind. The model is reported to be pretrained on over 400 languages (Comanici et al., 2025), however, details on the language coverage are not available.

- **Mistral Medium 3.1 model** (`mistral-medium-2508`) (Mistral AI, 2025) is a closed-source multimodal instruction-tuned model by Mistral AI. Available details on the model architecture and language coverage are very limited.

- **LLaMA 3.3 model**[16] (Meta, 2024) is an open-weight instruction-tuned multilingual LLM, developed by Meta, with 70 billion parameters. The model was pretrained on a web text collection in various languages, however, it is reported to support only 8 languages, namely, English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.

- **Gemma 3 model**[17] (Gemma Team et al., 2025) is an open-weight multilingual instruction-tuned LLM, developed by Google DeepMind. The model was pretrained on multimodal data with large quantities of multilingual texts and is reported to support over 140 languages. We use the model in 27 billion parameter size.

- **DeepSeek-R1-Distill**[18] (Guo et al., 2025) is an open-weight reasoning LLM, developed by DeepSeek AI. We use the distilled model in 14 billion parameter size, namely the `DeepSeek-R1-Distill-Qwen-14B` model. The model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that was fine-tuned using a dataset curated with the DeepSeek-R1 reasoning model. The Qwen 2.5 model provides multilingual support for over 29 languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, and Arabic.

- **Qwen 3**[19] (`Qwen3-2504`) (Yang et al., 2025) is an open-weight LLM, developed by Alibaba Cloud. We use the model with the 32 billion parameter size, namely, the `qwen3:32b` model. The model is said to support over 100 languages and dialects (Yang et al., 2025).

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). We use the quantized versions of the models as they are available through the Ollama library.[20] OpenAI models are used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[21] that provides a unified API access to various closed-source models.

To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. The same prompts were used for all open-weight and closed-source models. In Figure 4, we provide prompts that were provided to the LLMs for zero-shot text classification, namely for sentiment classification (Figure 4a), automatic genre identification (Figure 4b), news topic classification (Figure 4c) and topic classification in parliamentary

Hugging Face repository at https://doi.org/10.57967/hf/0927 and the CLARIN.SI repository at http://hdl.handle.net/11356/1961.

[16] https://ollama.com/library/llama3.3

[17] https://ollama.com/library/gemma3

[18] https://ollama.com/library/deepseek-r1:14b

[19] https://ollama.com/library/qwen3

[20] https://ollama.com/library

[21] https://openrouter.ai/

speeches (Figure 4d). For more details on the setups used to apply fine-tuned BERT-like models and instruction-tuned LLMs to the test datasets, refer to the code published on GitHub.[22]

[22] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

```
### Task
    Your task is to classify the provided parliamentary text into a sentiment
label, meaning that you need to recognize whether the speaker's sentiment towards the topic
is negative, neutral, positive or somewhere in between. You will be provided with an excerpt
from a parliamentary speech in {lang} language, delimited by single quotation marks. Always
provide a label, even if you are not sure.


### Output format
    Return a valid JSON dictionary with the following key: 'sentiment' and a
value should be an integer which represents one of the labels according to the following
dictionary: {sentiment_description}.

    Text: '{text}'
```

(a) Sentiment classification.

```
### Task
    Your task is to classify the following text according to genre. Genres are
text types, defined by the function of the text, author's purpose and form of the text.
Always provide a label, even if you are not sure.

### Output format
    Return a valid JSON dictionary with the following key: 'genre' and a
value should be an integer which represents one of the labels according to the following
dictionary: {labels_dict}.


    Text: '{text}'
```

(b) Automatic genre identification.

```
### Task
Your task is to classify the provided text into a topic label, meaning that you
need to recognize what is the topic of the text. You will be provided with a news text,
delimited by single quotation marks. Always provide a label, even if you are not sure.

### Output format
Return a valid JSON dictionary with the following key: 'topic' and a value
should be an integer which represents one of the labels according to the following
dictionary: {label_dict_with_description}.


Text: '{text}'
```

(c) News topic classification.

```
### Task
    Your task is to classify the provided text into a policy agenda topic label, meaning that you need to recognize what
is the predominant topic of the text. You will be provided with an excerpt from a parliamentary speech from the {parl} parliament in
{language} language, delimited by single quotation marks. Always provide a label, even if you are not sure.


    Follow the following rule: if the speech mentions a policy area and a policy instrument (e.g., taxes, laws), pick
the label based on the area, not the instrument (e.g., annotate mortgage tax changes with 14 (Housing), law on education with 6
(Education)).


### Output format
    Return a valid JSON dictionary with the following key: 'topic' and a value should be an integer which represents one
of the labels according to the following dictionary: {majortopics_description}.


    Text: '{text}
```

(d) Parliamentary topic classification.

Figure 4: The prompts that are provided to the LLMs for the sentiment identification task (Figure 4a), automatic genre identification (Figure 4b), and topic classification on news (Figure 4c) and parliamentary speeches (Figure 4d). The prompts comprise the description of the task and labels with short descriptions.

Exploring the Use of Large Language Models in Critical Discourse Analysis: A Consensus-Based Pilot Study

Emiliano Giovannetti, Francesca Cristiano

Cnr-Istituto di Linguistica Computazionale “A. Zampolli”
Via Moruzzi 1, 56124 Pisa, Italy
{emiliano.giovannetti, francesca.cristiano}@ilc.cnr.it

Abstract

Large Language Models (LLMs) are increasingly used in the social sciences and humanities (SSH) to support the analysis of complex textual data, raising methodological questions about evaluation and interpretive reliability. This paper explores the use of LLMs in Critical Discourse Analysis (CDA), considered here as a paradigmatic case of interpretive research in SSH, through a preliminary consensus-based evaluation framework. The study reports on a pilot experiment conducted on a small, theory-driven corpus of opinion articles addressing the October 7, 2023 attack and its aftermath. An LLM is asked to answer analytically motivated questions targeting different levels of discourse structure. Its responses are compared with annotations produced by multiple human analysts and aggregated through a consensus-based procedure. The results reveal an asymmetry in model performance: while LLMs align well with human consensus on macro- and superstructural features, they struggle with microstructural phenomena involving implicit meaning. These findings support the view of LLMs as epistemic support tools rather than replacements for human interpretation.

Keywords: critical discourse analysis, large language models, interpretive evaluation, consensus-based analysis, social sciences and humanities

1. Introduction

The growing availability of Large Language Models (LLMs) has generated considerable interest across the social sciences and humanities (SSH), where they are increasingly explored as tools for supporting the analysis of complex textual data (Underwood, 2025). LLMs produce semantically rich outputs and natural-language explanations, making them attractive for interpretive research domains. At the same time, their adoption raises methodological questions concerning evaluation, reliability, and the role of human judgment (Abdurahman et al., 2025).

To examine these questions in a concrete interpretive setting, this study focuses on Critical Discourse Analysis (CDA). CDA investigates how discourse contributes to the construction and reproduction of power relations and ideologies (Fairclough, 1995; van Dijk, 1998). Because it targets implicit meaning, evaluative framing, and ideologically loaded lexical choices, CDA represents a particularly demanding test case for assessing LLMs in SSH research. In recent years, the advent of LLMs has opened new possibilities for assisted CDA, while raising some methodological concerns. Although LLMs are highly efficient in implementing more mechanical tasks, they face greater challenges in performing complex analyses that require deep reasoning and significant critical distance (Gillings et al., 2025). LLMs appear flexible in conducting a range of analytical functions across different types of texts, and in their application to Critical Discourse Analysis. However, the specific role that they play in this field, and whether they should be applied autonomously or in combination with human work, remains still to be defined (DeJeu, 2025).

In the light of this, and within a field yet in an exploratory phase, the present study explores the use of LLMs in CDA through a consensus-based evaluation framework designed to assess their analytical behaviour in interpretive contexts. We report on an experiment conducted on a small, theory-driven corpus of opinion articles addressing a highly polarized and widely mediatized event, namely the October 7, 2023 attack and the subsequent escalation of the conflict. The choice of this case study reflects the high density of ideological positioning and discursive polarization characteristic of such contexts. Rather than asking whether LLMs can “perform” CDA autonomously, this study explores the conditions under which they may contribute to interpretive research practices. By comparing model outputs with consensus-based human annotations across different levels of discourse analysis, we aim to shed light on both the potential and the limitations of LLMs as epistemic support tools in the social sciences and humanities.

2. Theoretical and Methodological Framework

This study is grounded in Critical Discourse Analysis, with a particular reference to Teun A. van Dijk’s theoretical framework since he devotes significant attention to CDA applied to the study of the news. This analytical approach articulates discourse analysis across multiple, interconnected levels, namely macrostructure, superstructure, and microstructure, each associated with different degrees of explicitness and interpretive complexity (van Dijk, 1988).

The macrostructural level concerns the global meaning of a text. The main themes of a text

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 18–22
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(topics) are produced through well-defined rules and organized into a set of propositions. In journalistic texts, topics are not presented in a linear or sequential manner, but rather in an order according to which the most specific information precedes the less detailed one. Topics are an important aspect of news texts and they represent what the authors consider to be the most important information. The superstructural level refers to the conventional organization of texts according to genre-specific categories, such as headlines, summaries, and commentary, which guide readers' expectations and interpretive trajectories. The microstructure concerns local linguistic elements, namely words and sentences that make up a text, as well as the strategies of local meaning and their underlying ideology. This level of analysis focuses on semantics, syntax, style, and rhetoric.

This distinction is methodologically relevant for LLM evaluation: macro- and superstructural features rely more on surface regularities, whereas microstructural phenomena require pragmatic inference and contextual sensitivity. The contrast provides a principled lens for identifying structural limits in model behaviour. Within interpretive SSH traditions, CDA exemplifies inquiry where multiple theoretically grounded readings may coexist (Wodak & Meyer, 2009). Here, disagreement is an epistemic resource rather than noise to be minimized.

These considerations have direct implications for evaluating LLMs in SSH contexts. If interpretation is inherently plural and negotiated, rigid gold standards and accuracy-based metrics are insufficient to capture model behaviour meaningfully. Instead, evaluation must account for plausible human interpretations and the theoretical assumptions underlying them. For this reason, the present study adopts a consensus-based framework, aggregating multiple human annotations to define a shared interpretive reference. Consensus is not treated as absolute agreement, but as bounded convergence reflecting common ground and residual uncertainty. CDA thus functions as a methodological testbed for exploring how LLMs can be used in interpretive contexts, with evaluation serving as a structured means of assessing their analytical alignment.

3. The Experiment

The experiment reported in this paper is designed to explore how LLMs can be evaluated when applied to CDA. Rather than aiming at large-scale validation or performance benchmarking, the study adopts a small-scale, theory-driven design, intended to foreground methodological issues related to interpretation, consensus, and evaluation.

The corpus adopted in the experiment consists of thirty English-language opinion articles drawn from three newspapers with diverse political and ideological orientations: *The Jerusalem Post*, *The Electronic Intifada*, and *The Washington Post*. Opinion pieces were selected because they make ideological positioning particularly explicit and, therefore, constitute a suitable object for CDA. The articles cover the period from October 7, 2023 to January 7, 2024 and focus on the attack carried out by Hamas against Israel on October 7, 2023, together with its immediate political and military aftermath. This event was chosen as a case study due to its high degree of media visibility and discursive polarization, which results in sharply contrasting representations across different outlets.[1]

As previously stated, the analytical framework is grounded in van Dijk's model of discourse analysis and operationalized through a set of analytically motivated questions targeting different levels of textual organization.

We defined eight questions, distributed across the three main dimensions of the model. Q1–Q4 concern the macrostructure: Q1 asks raters to estimate the proportion of the article devoted to the events of 7 October 2023; Q2 focuses on the connotative framing of the attack, asking whether it is described in positive, negative, neutral terms, or not mentioned; Q3 and Q4 identify, respectively, the actors represented as the main agents of the action and those represented as its targets. Q5 addresses the superstructure and focuses on the function of the headline, distinguishing between informative, persuasive, and emotionally oriented titles. Q6–Q8 concern the microstructure: Q6 examines the presence of negatively connoted lexical items, Q7 the use of euphemisms as mitigating strategies in the representation of violent events, and Q8 investigates forms of linguistic dehumanization directed at specific groups. Taken together, these questions translate key CDA categories into a controlled annotation scheme that can be applied comparatively to both human raters and LLMs.

The resulting scheme includes both ordinal categories (Q1, Q6, Q7) and nominal categories (Q2–Q5, Q8), reflecting different degrees of interpretive gradience. The aim is not to exhaustively annotate all the possible discursive features, but to test how an LLM responds to questions that vary in terms of interpretive explicitness and contextual dependence. The full list of analytical questions, together with the prompts and the corpus of articles used in the experiment and all resulting data are publicly available in an online repository, in line with

[1] Given the political sensitivity of the topic, it is important to acknowledge that LLM outputs may also reflect alignment constraints and training-data biases embedded in the model. While a systematic bias analysis is beyond the scope of this pilot study, this factor should be considered when interpreting the results.

principles of transparency and reproducibility in SSH research.[2]

Three human analysts with expertise in CDA independently annotated the entire corpus by answering the same set of questions for each article. To reduce potential bias, the articles were anonymized and presented in random order, without reference to their source. In parallel, the same questions were submitted to GPT-4o via the ChatGPT web interface, conducting five independent runs for each article. Each run was performed in a separate temporary chat session to avoid cross-instance memory effects. This procedure was adopted to account for the non-deterministic nature of the model outputs and to assess the internal stability of the model's responses.

Given the interpretive nature of the task, the evaluation did not rely on a single authoritative annotation. Instead, both human and model-generated responses were aggregated using a consensus-based procedure, primarily based on majority agreement. In cases where no simple majority emerged, a predefined deterministic rule was applied to ensure consistency: for the ordinal categories (Q1, Q6, Q7), consensus was defined by selecting the median category, whereas for the nominal categories (Q2–Q5, Q8), a fixed tie-breaking criterion was used. To ensure the reliability of this reference standard, an inter-annotator agreement was first assessed among the human raters and, separately, across the multiple runs of the LLM.

Agreement was assessed using complementary metrics capturing both chance-corrected agreement (Fleiss' κ for human annotators) and distributional convergence; inter-run consistency of the LLM was quantified using the Mean Modal Agreement Ratio[3] (Artstein & Poesio, 2008), reported in Figure 1. Fleiss' κ indicated overall moderate agreement among human annotators, with lower values observed for microstructural categories, confirming their higher interpretive variability.

All the inter-agreement measures and the underlying data are made openly available in the repository cited above, in the interest of transparency and reproducibility. This approach reflects the assumption that, in interpretive SSH research, analytical validity emerges from negotiated convergence rather than from absolute correctness. The model's outputs were therefore evaluated by comparing their consensus labels with the consensus derived from human annotations, allowing us to assess whether the LLM's analyses fall within the range of plausible human interpretations.

Overall, this experimental design is intended to shift the focus from performance measurement to methodological reflection. By combining human disagreement, model stability, and consensus-based comparison, the experiment provides a framework for examining the conditions under which LLMs may act as epistemic support tools for interpretive analysis, rather than as autonomous analytical agents.

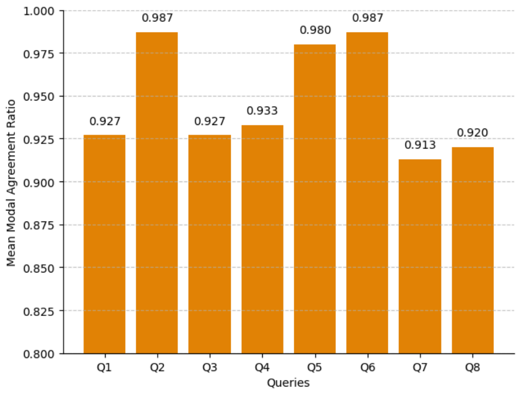


Figure 1: Inter-run agreement of the LLM across the eight queries, measured using the Mean Modal Agreement Ratio

4. Results and analysis

The results of the experiment reveal a differentiated pattern in the alignment between the LLM's outputs and the consensus-based human annotations, highlighting both the potential and the limitations of LLMs when applied to CDA (Fig. 2). Quantitatively, LLM consensus accuracy ranged from 0.67 to 0.97 across the eight analytical questions. Performance was the highest for superstructural and macrostructural categories (Q1–Q5), while lower scores were observed for microstructural phenomena (Q6–Q8), particularly euphemism detection. In particular, the headline function (Q5) reached 0.97 accuracy, while euphemism detection (Q7) dropped to 0.67, marking the largest divergence between model and human consensus. When a distribution-sensitive agreement measure was considered, alignment scores decreased consistently, indicating that strict majority comparison slightly overestimates convergence in cases of human disagreement.

---

[2] https://github.com/klab-ilc-cnr/critical-discourse-analysis

[3] Mean Modal Agreement Ratio measures how often annotators or model runs converge on the most frequent category. It provides an intuitive estimate of agreement and it is particularly suitable for interpretive tasks where some degree of disagreement is expected; in this study, it is used to assess the internal consistency of the LLM across multiple runs.

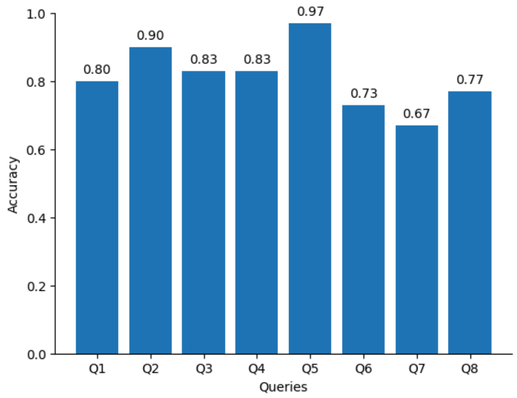


Figure 2: Accuracy of the LLM with respect to the human consensus across the eight queries

At the macrostructural level (Q1–Q4), the model reliably identifies the global themes, the relative emphasis on events, and the representation of social actors. An even higher convergence emerges for the superstructural features, particularly the headline function (Q5), which reaches 0.97 accuracy. This pattern aligns with van Dijk's (1985) notion of news schemata, where fixed organizational categories such as headlines follow predictable conventions that LLMs can effectively capture.

A different pattern emerges in microstructural analysis (Q6–Q8), where ideological meaning is often implicit. In euphemism detection (Q7), the model systematically misclassified highly explicit expressions such as "*mass slaughter*," "*carnage*," and "*terrible blow*" as euphemistic. From a CDA perspective, these terms are not mitigating but overtly explicit and emotionally charged. This suggests that the model equates euphemisms with lexical salience rather than pragmatic attenuation. Notably, these misclassifications were stable across the runs, indicating a consistent internal heuristic rather than random variation.

Similar issues arise in the analysis of dehumanization (Q8). In several cases, the model classified politically charged expressions such as "*Zionist enemy*" or references to "*Zionist aggression*" as instances of dehumanizing discourse. Human analysts, by contrast, did not reach the same conclusion, as these expressions articulate ideological positioning and moral condemnation without necessarily denying the humanity of the targeted group. In another instance, the model identified the Israeli military as the dehumanized subject on the basis of formulations describing it as inherently criminal or genocidal, even in the absence of animalistic metaphors or explicit denial of moral status. These examples point to a systematic tendency to conflate strong evaluative stance with dehumanization, overlooking the more specific linguistic mechanisms through which dehumanization operates in discourse.

Taken together, these cases illustrate a broader pattern in the model's behaviour: when confronted with microstructural phenomena that require pragmatic inference and contextual interpretation, the model tends to overextend analytical categories on the basis of lexical cues alone. Rather than assessing whether a given expression functions rhetorically as mitigation or dehumanization within its discursive context, the model appears to rely on surface features such as emotional intensity or ideological polarity. This behaviour does not occur sporadically, but recurs across texts and questions, indicating a structural limitation rather than isolated error.

From a methodological perspective, these findings gain further significance when considered alongside the contrast between the human interpretive variability and the model stability. The human annotations display non-negligible disagreement for precisely these microstructural categories, confirming that such phenomena are intrinsically open to interpretation. The model, by contrast, produces highly stable responses across multiple runs, suggesting internal consistency. In an interpretive SSH context, however, such stability may signal epistemic rigidity rather than analytical robustness, as consistent outputs can still reflect systematically simplified or biased readings.

Overall, the results highlight a fundamental asymmetry in the applicability of LLMs to interpretive analysis. While LLMs appear well suited to supporting exploratory work at the level of global structure and discursive organization, they struggle with forms of meaning that rely on implication, mitigation, and contextual negotiation—precisely the dimensions that CDA identifies as central to ideological critique (Fairclough, 1995; van Dijk, 1993). These limitations should not be interpreted merely as performance deficits, but as indicators of the epistemic boundary between computational pattern recognition and human critical interpretation.

5. Conclusions

This pilot study explored how LLMs can be integrated into CDA practice, employing a consensus-based evaluation framework to examine their interpretive behaviour. Instead of evaluating LLMs' ability to "perform" Critical Discourse Analysis independently, the study has focused on how their outputs relate to shared human interpretations and on what this relationship reveals about the epistemic role of LLMs in SSH research.

The experimental findings highlight a systematic asymmetry in the model behaviour. LLMs align closely with human consensus when addressing macro- and superstructural aspects of discourse, such as thematic framing, the actor representation, and the headline function. These dimensions rely on relatively explicit cues and genre conventions, which appear well suited to the model's strengths in pattern recognition and contextual association. By contrast, the analysis of microstructural phenomena—where ideological meaning is often implicit, mitigated, or rhetorically mediated—reveals recurring limitations.

In conclusion, the findings brace a view of LLMs as epistemic support tools rather than replacements for human analysts. Their value lies in their ability to provide stable, reproducible perspectives that can assist exploratory analysis and highlight salient discursive patterns. At the same time, their limitations underscore the continued centrality of human judgment in the analysis of ideologically charged and rhetorically complex texts.

6. Bibliographical References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., and Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8(2):1–25.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

DeJeu, E. B. (2025). Can (and should) LLMs perform critical discourse analysis? *Journal of Multicultural Discourses*, 19(3):188–195.

Fairclough, N. (1995). *Critical discourse analysis: the critical study of language*. Longman, London and New York.

Gillings, M., Kohn, T., and Mautner, G. (2025). The rise of large language models: challenges for Critical Discourse Studies. *Critical Discourse Studies*, 22(6):625–641.

Underwood, T. (2025). The impact of language models on the humanities and vice versa. *Nature Computational Science*, 5(9):695–697.

Van Dijk, T. A. (1985). Structures of news in the press. In T. A. Van Dijk, (Ed.), *Discourse and Communication*. Berlin, New York: DE GRUYTER, pp. 69–93.

Van Dijk, T. A. (1988). *News as discourse*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2):249–283.

Van Dijk, T. A. (1998). Opinions and ideologies in the press. In A. Bell, and P. Garrett (Eds.) *Approaches to Media Discourse*. Oxford: Blackwell. pp. 21–63.

Wodak, R., and Meyer, M. (2009). *Methods of critical discourse analysis*. SAGE Publications, London, 2nd edition.

LLM Evaluation in Practice: A Review of Metrics, Practitioner Insights, and Lessons Learned

R.M. Bakker[1,2], M.W. Witte-Schaaphok[1], J. García-Fernández[1], T. Brand[1], J. van der Weide[1], S.A. Raaijmakers[1,2]
[1] The Netherlands Organization for Applied Scientific Research (TNO),
[2] Leiden University Centre for Linguistics
{roos.bakker, marianne.schaaphok, julia.garciafernandez, tom.brand, jens.vanderweide,stephan.raaijmakers}@tno.nl

Abstract

The rapid, widespread adoption of Large Language Models (LLMs) highlights the need to understand their performance, strengths, and limitations. However, evaluating LLMs presents significant challenges due to the broad range of tasks and model capabilities, especially in practice or low-resource settings where benchmark datasets are not available. In text generation tasks, answer diversity has always complicated automatic evaluation, and the enhanced fluency and creativity of LLMs lead to further challenges. Existing metrics and frameworks often fail to account for these complexities. Furthermore, recent research into the replicability of benchmarks has demonstrated serious issues when reproducing historical benchmark results. This paper makes two key contributions: (1) a categorisation of challenges and metrics in LLM evaluation, and (2) lessons learned from practice through a survey and a use case. To this end, a literature study was conducted to identify challenges and metrics in scientific work. A survey among developers working with LLMs provided insights into practical challenges. Furthermore, selected metrics were implemented in a practical use case to gain insights into their strengths and limitations. By combining theoretical analysis with real-world experiences and lessons learned from practice, this work provides an overview and best practices for users evaluating LLM performance.

Keywords: LLM Evaluation, NLP Evaluation, Applied Evaluation

1. Introduction

The evaluation of generative Large Language Models (LLMs) has become increasingly critical as these models have gained widespread use in recent years. In the past, language models were evaluated on specific tasks using manual annotations against benchmarks. While this approach remains effective for classification tasks, it is not always suitable for text generation due to the variability of natural language. Correct answers can be formulated in many acceptable ways, which makes it difficult to compare LLM answers to a fixed ground truth. Another challenge is that commonly used benchmarks are not always representative of real-world applications or practical usage scenarios (Kiela et al., 2021). Furthermore, many benchmarks are publicly available and may be included in LLM training (Ravaut et al., 2024), raising concerns about overfitting (Zhang et al., 2024). In this work, we contribute to a better understanding of these challenges in a practical setting through two contributions: (1) a categorisation of the key challenges and metrics in evaluating LLMs, (2) lessons learned from practice through a survey and a use case.

This paper is structured as follows: in Section 2, we will discuss related work on LLMs, LLM evaluation, and we will present a categorisation of evaluation challenges. In Section 3, we will compare a set of evaluation metrics. In Section 4, we will discuss evaluation challenges in practical contexts, gathered through a survey and a practical use case. In Section 5, we will discuss the insights from the survey and the use case. Finally, we will conclude our work and discuss ideas for future work.

2. Related Work

LLMs are a progression of standard statistical language models that condition word generation on word context by assigning probabilities to word sequences. LLMs are produced by deep neural networks based on Transformer architectures (Vaswani et al., 2017), and condition the generation of words on vectorised representations of left context. In the original Transformer architecture, a separate bi-directional encoder optimises this vectorisation in conjunction with a left-to-right operating, autoregressive decoder that provides feedback to the encoder. Most contemporary LLMs, such as GPT-3, are decoder-only architectures that integrate the encoding process within the model itself (Brown et al., 2020). These models are large according to three dimensions: the amount of words they are initially trained on, the amount of parameters in the underlying neural network models, and the training budget (Minaee et al., 2025; Naveed et al., 2024).

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 23–38
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2.1. LLM Evaluation

LLMs rapidly gained popularity across diverse applications, increasing the need for robust evaluation. Traditional evaluation methods, such as perplexity, BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), have been widely used. However, due to the fast advancements of LLMs, those metrics no longer capture the full scope of model performance (Meister and Cotterell, 2021; Wu et al., 2023a). This sparked a search for new metrics and a deeper understanding of existing ones.

Several surveys review evaluation methods for text generation and large language models, and try to categorise methods along different dimensions. For instance, the surveys of Celikyilmaz et al. (2021); Belz and Reiter (2006) make a distinction between intrinsic evaluation, which focuses on model behaviour, and extrinsic evaluation, which considers downstream task performance. The recent survey of Chang et al. (2024) takes a different approach with grouping tasks ("what"), datasets ("where"), and protocols ("how") (Chang et al., 2024) or automatic versus human (Belz and Reiter, 2006). Some surveys take a more narrow scope, such as benchmarks only (Ni et al., 2025) or evaluation challenges and limitations (Laskar et al., 2024) (which we further discuss in Section 2.1. Further categories can be added, such as alignment and safety (Guo et al., 2023; Liu et al., 2023b), but it is not always clear how those align with existing work, reflecting the lack of consensus on evaluation standards (Chang et al., 2024).

We introduce a categorisation of aspects of LLM evaluation in Table 1, which includes categories such as the ones discussed above. At the top level, we have four categories. The scope indicates the focus of the evaluation, which can be on intrinsic qualities, such as correctness of generated texts (Celikyilmaz et al., 2021), or extrinsic qualities such as a user's improved comprehension after interacting with a downstream task (Belz and Reiter, 2006). The second category of evaluation is by the approach or method of evaluation: by humans, metrics, benchmarks, or a hybrid form. Metrics and benchmarks can both be considered automatic methods (Celikyilmaz et al., 2021), align with the "where" category of Chang et al. (2024), and fall under the 'organisation' category of Guo et al. (2023). Third, evaluation can be distinguished by task type, similar to Chang et al. (2024). Within the task types, a broad distinction can be made between Natural Language Understanding (NLU) and Natural Language Generation (NLG) (Jurafsky and Martin, 2008; Khurana et al., 2023). NLU tasks, like sentiment analysis or semantic role labelling, are generally easier to evaluate using accuracy or F1. NLG tasks, such as summarisation or question answering, are more challenging due to the diversity of valid outputs and the need to assess correctness, fluency, and relevance (Liu et al., 2023a). Finally, we identify four levels of evaluation: input or prompt level, output, model, and system level.

LLM Evaluation			
Scope	**Approach**	**Task**	**Level**
Intrinsic	Human	NLG	Input
Extrinsic	Metrics	NLU	Output
	Benchmarks	Combined	Model
	Hybrid		System

Table 1: Categorisation of LLM evaluation dimensions.

2.2. Evaluation Challenges

Several challenges recur in the literature, summarised in Table 2 across scope, approach, task, and level. Many challenges cut across multiple evaluation aspects; for example, sustainability concerns can arise at multiple tasks and approaches. We focus on representative challenges at the scope and approach levels to illustrate key issues, rather than attempting an exhaustive list.

Intrinsic Challenges Intrinsic challenges can be summarised by a lack of robustness, reproducibility, variation in possible answers and a lack of model transparency. The lack of robustness can be explained by the fact that LLMs have an inherent instability, which can result in different answers for small variations in prompts. Therefore, different punctuation, wording, or order of prompts can affect the outcome of the system (Sclar et al., 2024) (Loya et al., 2023). Mizrahi et al. (2024) and Hida et al. (2024) show that varying the prompts and the prompt formatting can result in different rankings of LLMs in terms of performance. Additionally, for few-shot learning, the order of samples affects performance (Lu et al., 2022), and similarly, the order of answers for multiple choice tasks also matters (Pezeshkpour and Hruschka, 2024).

Related to this is the lack of reproducibility: generally, LLMs lack the ability to consistently obtain the same results under the same conditions due to their non-deterministic nature and confounding variables such as errors in benchmarks and different LLM model versions. This affects evaluation trustworthiness on local (where the same task can yield inconsistent results) and global scale (where reproducing research results is hindered by insufficient documentation and version control) (Laskar et al., 2024; Biderman et al., 2024). We note that non-determinism can be limited through greedy decoding settings, though this is not common practice. Recent research by Vaugrante et al. (2024) demonstrates that many benchmark results that led

Challenge	Scope	Approach	Task	Level
Robustness	Intrinsic	Any	NLG	Model
Reproducibility	Intrinsic	Any	NLG	Model
Model transparency	Intrinsic	Any	NLG	Model
Lack of standardisation	Extrinsic	Human	Any	Any
Subjectivity	Extrinsic	Human	Any	Any
Data leakage	Extrinsic	Metrics	NLU	System
Data quality	Extrinsic	Benchmarks/Metrics	Any	Output
Sustainability	Extrinsic	Any	Any	System
Bias and fairness	Intrinsic/Extrinsic	Any	Any	Any
Answer diversity	Intrinsic/Extrinsic	Metrics	NLG	Output

Table 2: LLM evaluation challenges mapped to scope, approach, task, and level.

to published landmark results (such as the effectiveness of zero-shot chain-of-thought prompting, expert prompting and sandbagging) are not reproducible, even when the same benchmarks are run again on the same model versions. This is indicative of a fundamental replicability "crisis" in LLM research. Further, and importantly, the statistical underpinning of evaluation results is only a relatively new topic in the LLM field (see e.g. (Miller, 2024)).

Sallou et al. (2024) identify three key reproducibility challenges: output variability, time-based output drift (due to retraining or user feedback), and traceability (linking outputs to specific prompts and configurations). Additionally, Atil et al. (2024) note that LLM stability varies across tasks and is rarely deterministic.

Another challenge that can be both intrinsic and extrinsic is the diversity of correct answers in NLG tasks (Wang et al., 2023). Lexical and semantic matching techniques can evaluate answers to a certain degree, but it is not useful if the given answer is not included in the ground truth (Kamalloo et al., 2023). Finally, lack of model transparency complicates LLM evaluation. Many models do not disclose their training data, architecture, or weights, making them harder to assess (Liu et al., 2023c; Liesenfeld and Dingemanse, 2024). Even with disclosed data, the low explainability of LLMs limits understanding of how they produce answers (Wu et al., 2023b)

Human-related Evaluation Challenges We can identify two evaluation challenges that mainly relate to human aspects: 1) lack of universally accepted standards, and 2) human subjectivity in evaluation. The rapid growth of LLMs has led to numerous benchmarks and evaluation methods, but no universally accepted standard exists. Researchers create their own benchmarks, raising challenges in benchmark selection, implementation, prompt variations, and fair model comparison (McIntosh et al., 2024; Post, 2018; Biderman et al., 2024). While human evaluation is essential, it remains challenging due to inter-annotator disagreement, biases, and sensitivity to question framing (Abeysinghe and Circi, 2024). Besides, human evaluation is often time-consuming and costly. Subjectivity also affects automatic evaluation, as benchmarks, examples, and annotations are influenced by human judgment.

Metrics & Benchmark Challenges There are several challenges due to the limitations of current evaluation metrics and benchmarks. First of all, data leakage or data contamination gives rise to the question whether results on benchmarks and test sets can still be trusted (Sainz et al., 2023; Zhou et al., 2023). Balloccu et al. (2024) show that the GPT-3.5 and GPT-4 models have been exposed to 4.7M samples from 263 different benchmarks during retraining. Second, whether models treat individuals or social groups fairly has been a complex, yet pressing evaluation challenge. Even current state-of-the-art LLMs have been proven to exhibit biased behaviour (Plaza-del Arco et al., 2024). There is no consensus on the conceptualisation of ‘social bias’ in its many forms (Blodgett et al., 2020). For LLMs specifically, existing benchmarks have been shown to be inconsistent in measuring different forms of bias (Blodgett et al., 2021). While there exist many different metrics and datasets, a ’gold standard’ is lacking (Gallegos et al., 2024).

Furthermore, there is the issue of data quality. A ground truth dataset is often required, but they are costly to construct and thus limited in size and variety (Nasution and Onan, 2024). Annotation by humans is not always consistent (Hashemi et al., 2024), and annotator quality also influences the quality of the dataset, and therefore the evaluation (Grosman et al., 2020; de León Languré and Zareei, 2024). Finally, LLMs require a large amount of computing resources. A challenge is how to measure the sustainability of both training and using these models (Khowaja et al., 2024).

3. Metrics

When evaluating LLMs in practice, relying on human evaluation is often costly and time-intensive, while standard benchmarks may be unsuitable because they are neither domain- nor data-independent. As a result, automatic metrics are frequently used as a practical alternative. This section presents an overview of existing metrics, their strengths, and weaknesses.

In general, the assessed metrics show moderate correlation with human judgment and are inadequate for evaluations requiring 100% correlation with human evaluators due to the subjective nature of text evaluation (Celikyilmaz et al., 2021; Zhu et al., 2024; Liu et al., 2023a). Most metrics require references (a ground truth), which are often low-quality or unavailable (Ke et al., 2022). Reference-free metrics need unambiguous criteria definitions. Metrics like BLEU (Papineni et al., 2002) can vary in implementation, affecting comparability across studies (Post, 2018). There is no ultimate metric for LLM evaluation; the choice depends on the task, scope, and level of evaluation, the level of correlation with human judgment needed, and other factors such as computational resources and reference availability. In Table 3, we have compiled a set of metrics. These are categorised into classification, ranking, statistical, model-based, and LLM-based, each with its strengths and weaknesses, discussed below.

Traditional classification metrics They categorise the outputs as correct or incorrect and provide simple performance metrics. The metrics included in this category in Table 3 are accuracy, precision, recall, and F-score. **Strengths**: Traditional classification metrics tend to be simple and intuitive, making them easy to implement and easy to interpret. They are a straightforward tool for binary classification tasks, such as comparing generated text to a gold standard reference. **Weaknesses**: These metrics require a narrow definition of correctness, which is not available in all text generation tasks. They often do not work well with unbalanced datasets, and high-quality labelled data is required to compute these metrics, which is not always available. Furthermore, the binary categorisation approach fails to account for partially correct outputs, which is crucial in open-ended text generation tasks.

Ranking metrics These are used to evaluate the performance of models that produce a ranked set of possible solutions to the given task. These metrics compare these ranked outputs against a correct solution or reference. The metrics included in this category in Table 3 are SaCC (strict accuracy), LaCC (lenient accuracy), and MRR (mean reciprocal rank) (Voorhees and Tice, 2000). **Strengths**: Ranking metrics are easy to understand and allow for more nuanced evaluation of non-deterministic models by considering a ranking of multiple different responses to the same query. They are particularly valuable for recommendation systems (e.g. search engines), where the position of the correct answer in a list is significant (Jadon and Patil, 2025). These metrics offer a more detailed insight into text generation performance than metrics based on only a single output per query.

Weaknesses: The quality of these metrics relies heavily on the references used and can be sensitive to class imbalance. They require multiple outputs per query and ranking, making them more computationally expensive than single-output metrics. Ranking metrics often only consider the highest-ranked correct answer, ignoring other correct answers or when the text generator indicates that the answer is unknown.

Statistical metrics They measure the level of correspondence or matching between n-grams, which can be characters, word pairs or word sequences. The metrics included in this category in Table 3 are chrF (Popović, 2015), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and Perplexity (Jelinek et al., 1977). **Strengths**: They are simple and easy to implement, and especially powerful for NLG tasks that require exact matching, such as speech recognition or machine translation (MT). **Weaknesses**: They do not account for the fact that the same idea can be correctly expressed in various ways, i.e., they do not account for meaning, only for lexical overlap. They are only representative if high-quality references are available.

Model-based metrics These metrics use the tokenizer functions and/or embeddings of language models to encode the text, and evaluate text by computing the similarity between the generated text and the expected answer. The metrics included in this category in Table 3 are BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019), Mauve (Pillutla et al., 2024), MoverScore (Zhao et al., 2019), COMET (Rei et al., 2020), FrugalScore (Kamal Eddine et al., 2022), and CTRLEval (Ke et al., 2022). **Strengths**: Model-based metrics take semantics into consideration, which traditional metrics cannot do. Whereas traditional metrics tend to only work with proxies for the semantics of a text, the LLMs that these metrics are based on are designed to capture meaning. In some cases, references are not needed. **Weaknesses**: The effectiveness of these metrics depends on the performance of the model. They require significant computational re-

	Metric name	Task	Measures	GT	Resources
Classification	**Accuracy**	-	Fraction of correct answers	✓	- -
	Precision	-	Fraction of true positives over all positive responses	✓	- -
	Recall	-	Fraction of true positives over all actual positive responses	✓	- -
	F-score	-	Harmonic mean of precision and recall	✓	- -
Ranking	**SaCC**	-	How often highest ranked answer is correct	✓	- -
	LaCC	-	How often correct answer in top 5	✓	- -
	MRR	-	How high correct answer ranks on average	✓	- -
Statistical	**chrF**	MT	Character N-gram F-score	✓	-
	BLEU	MT	Similarity of translation to a reference text	✓	-
	ROUGE	Summ.	Lexical overlap between summary and reference text	✓	-
	METEOR	MT	Similarity of translation to a reference text	✓	-
	Perplexity	-	Model's probability to predict a given text	✗	+
Model-based	**BLEURT**	-	Non-trivial semantic similarities between sentences	✓	+
	BERTScore	-	Contextualised embedding-based semantic similarity	✓	+
	Mauve	-	Statistical gap between two text distributions	✓	+
	MoverScore	-	Semantic similarity of generated text to reference text	✓	+
	COMET	MT	Quality of generated translation	✗	++
	FrugalScore	-	Semantic similarity of generated text to reference text	✓	+
	CTRLEval	Gen.	Coherence, consistency and attribute relevance of controlled text generation	✗	+
LLM-based	**G-Eval**	-	Quality of NLG output based on user-defined criteria	✗	+++
	Prometheus 1 and 2	-	Any long-form text based on customised score rubric provided by the user	✓	++
	GEMBA	MT	Quality of generated translation	✗	+++

Table 3: Overview of metrics to evaluate text generation. *GT* = Ground Truth required. For the column *Task* (task specific metrics): *MT* = Machine Translation, *Summ.* = summarization, *Gen.* = text generation.

sources and inherit some of the problems of language models, such as biases. These metrics become a 'moving target' when new and improved versions of the model are released, where the question of whether the metrics should be updated is hard to answer, because although the model might have improved, scores are only comparable between identical models.

Generative LLM as evaluator (LLM-based) These methods leverage the reasoning and instruction-following capabilities of generative LLMs by proposing a tool, framework, or set of steps to use an LLM to evaluate generated text based on some criteria defined by the user. This approach is also often called LLM-as-a-Judge (Gu et al., 2025). The metrics included in this category in Table 3 are G-Eval (Liu et al., 2023a), Prometheus 1 and 2 (Kim et al., 2024a,b), and GEMBA (Kocmi and Federmann, 2023). **Strengths**: These evaluation methods are flexible and can be used with any text generation task. The scores given can be explained by the LLM in natural language. They can be improved by evaluating the same task several times and computing the average or distribution. **Weaknesses**: These metrics are highly dependent on the model, instructions, and language used. They inherit issues like inconsistency, lack of transparency, and bias from language models. Using LLMs as evaluators demands significant computational resources and raises ethical and environmental concerns, especially with proprietary models. Generative LLMs have larger computational requirements than those used in simpler approaches like BERTScore.

4. Evaluation in Practice

To collect practical evaluation challenges, we conducted a survey with 19 LLM developers at a Dutch applied research institute. The survey aimed to gain insights into the evaluation methods currently used and the challenges encountered. Additionally, we implemented metrics in a practical use case to gain insight into their strengths and weaknesses.

4.1. Survey Results Overview

The 19 respondents varied in experience, sector (e.g., government, health), and type of developed applications. We asked the respondents questions on applied evaluation methods, arguments for their choice of evaluation methods, challenges they face in their evaluation, which challenges they find most urgent and what they would need for better evaluations[1]. From the survey results, we focused on respondents' evaluation methods, the reasons for using those, and their questions and needs regarding LLM evaluation in practice. For each response, we manually extracted a list of needs, then grouped similar or identical needs. Finally, we categorised the extracted needs into five themes: Resources, Datasets, Metrics, System, and Governance, which are discussed below[2].

Resources 14 respondents expressed a strong need for comprehensive and practical guidance to support evaluation workflows. This includes general and task-specific evaluation frameworks, prompting strategies, and human-in-the-loop methodologies. There is a clear demand for structured approaches that help practitioners tailor metrics to use cases, integrate manual and automated methods, and identify suitable domain experts. The community also seeks clarity on when and how to apply different evaluation techniques, supported by up-to-date lists of tools and relevant literature. Handling LLM-specific challenges, such as hallucinations, multiple correct answers, and instruction-following failures, was also highlighted as a critical area requiring dedicated guidelines.

Datasets 5 evaluation practitioners emphasised the importance of transparency in dataset construction and usage. Key concerns include contamination of evaluation datasets, the quality and reliability of reference data, and the availability of benchmarks that are both accessible and widely accepted. There is also a need for support in creating custom benchmarks tailored to specific tasks or domains. Notably, respondents highlighted the challenge of evaluating without ground truth, pointing to a gap in methods that can assess model outputs in contexts without a ground truth available.

Metrics 19 practitioners expressed a need for metrics that are interpretable, robust, and aligned with human judgment, especially in terms of correctness over fluency. There is a desire for flexibility in selecting metrics across multiple dimensions, such as factuality, creativity, and bias, and for tools that support custom metric design. Respondents also indicated a desire for visibility into the computational and environmental costs of metrics, as well as guidance on their applicability across different tasks, budgets, and domains. Specific needs include metrics for multilingual evaluation, fuzzy matching, and RAG systems.

System 3 practitioners indicated the importance of contextualising evaluation within the broader system in which the LLM operates. This includes understanding the required quality level for a given application and clarifying who the intended end user of each evaluation method is. There is also a need for clearer guidance on when LLM-based evaluators are appropriate, especially in relation to human judgment and task complexity. These insights point to a growing recognition that evaluation cannot be isolated from the product or pipeline context and must be adapted to real-world deployment scenarios.

Governance The governance theme reflects concerns about the broader implications of evaluation practices. 2 respondents called for consideration of the environmental impact of evaluation methods, as they noted that robust evaluation requires repeated trials to achieve statistically significant results, which can conflict with sustainability goals due to the increased computational and energy demands. There is also a need for scoring or assessing evaluation approaches based on their alignment with ethical and legal standards, particularly EU and Dutch values. These concerns highlight the importance of responsible evaluation practices that go beyond technical performance to include sustainability and compliance dimensions.

4.2. Metric tests in a practical use case

To understand the challenges of evaluating LLM-generated text in practice, we implemented three metrics in a practical use case involving a chatbot prototype for a government dashboard. The chatbot helps users interpret complex health statistics from a psychological/behavioural model. The chatbot is designed to handle simple questions only, allowing psychologists more time to address complex queries from the users of the dashboard. Examples of questions that the chatbot can answer are: *what does 'technostress' mean?*; or *what is the score (or status) of 'social support' in my team*; or *how does 'work-life balance' relate to 'work engagement'?*. The chatbot can also answer more complex filter questions such as: *which indicators in the positive functioning category have a 'green' status?*.

[1] All survey questions can be found in Appendix A.

[2] The full list of needs for LLM evaluation in practice extracted from the survey and categorised by theme can be found in Appendix B.

To answer the questions, the chatbot retrieves data using predefined functions (function calling), which ensures responses are based on actual data. If users ask unrelated questions, the chatbot politely declines to answer. Metrics were chosen based on task type, availability of reference text, and chatbot response type (textual/numerical output). For all tasks, we check that the correct function is being called. Then, by using Table 3 and based on their popularity and availability off-the-shelf, we chose a set of metrics to apply to the content of the responses by the chatbot.

Exact string matching Although not an LLM evaluation metric, in this case, it is suitable for questions in which the chatbot is expected to answer with a numerical score, a predefined status, or a partially fixed message.

For a numerical score or a predefined status (e.g., 'green' or 'good', 'orange' or 'attention', or 'red' or 'urgent'), we built a helper function that extracts the number or status word from the answer and matches it with the expected number or status word from the database.

For queries to which the chatbot refuses to answer, we compiled a set of semantically equivalent formulations in Dutch. In our automatic evaluation, we check that at least one of these pieces of text is present in the chatbot's answer.

This method is simple, reliable and deterministic, and it successfully detects when a chatbot's answer is wrong, according to human judgment.

BLEU (Papineni et al., 2002) For questions about the definition of a term, which involves retrieving definitions from the database, BLEU can measure lexical overlap between the references and the chatbot's output. We used SacreBLEU (Post, 2018), which standardises tokenisation for consistent scores. As the desired output is composed of two different text components (definition and interpretation), we also applied fuzzy string matching with these two texts to complement the BLEU scores. BLEU measures lexical overlap well when small deviations from a reference answer are acceptable. This metric is relatively simple, consistent and deterministic. The score successfully detects when the chatbot response is very different to what is expected. However, because BLEU is a continuous score, it was necessary to define a threshold to classify responses as correct or incorrect. To determine this threshold, we explored three approaches.

The first approach was to use the arithmetic mean of all BLEU scores as the threshold. This method is simple and intuitive, as it considers the overall distribution of scores. However, it is sensitive to outliers, potentially leading to a threshold that does not reflect the majority of cases.

To mitigate the effect of outliers, we also considered the median, which represents the middle value when all scores are sorted. The median is more robust to skewed distributions and extreme values, providing a more stable threshold in cases where the BLEU scores are not symmetrically distributed.

Finally, we applied the Jenks natural breaks optimisation (Fisher-Jenks algorithm) (Jenks, 1967), which partitions the data into classes by minimising variance within each class and maximising variance between classes. In our case, we specified two classes (correct vs. incorrect), and the algorithm identified the break point that best separates the two clusters of BLEU scores. Unlike the mean or median, this method adapts to the actual distribution of the data, making it particularly suitable when scores form distinct groups.

To validate these thresholds, we conducted a manual evaluation comparing expected answers to model outputs. The threshold identified by the Jenks-based approach corresponded best with human judgment. Consequently, we incorporated fixed lower and upper bounds (10 and 30) alongside the Jenks-based threshold for classifying ambiguous responses, ensuring that only responses with a reasonable degree of overlap are considered correct, while also automating the classification of clearly inadequate or highly satisfactory responses.[3]

BERTScore (Zhang et al., 2019) When a user asks for concepts that are not in the database, or about any information unrelated to the dashboard, the chatbot should politely decline to answer. We evaluated this using exact string matching. In addition, we wanted to assess whether a model-based metric would be more suitable than string matching in this case, so we implemented BERTScore. Model-based metrics that calculate the similarities using the contextualised embeddings are more suitable than metrics based on lexical overlap, such as BLEU, for textual outputs in which the response can convey the correct message but using different words. We compared BLEU, fuzzy string matching, and BERTScore on a sample reference text. BERTScore yielded results comparable to BLEU and fuzzy string matching. BERTScore is more lenient towards small syntactic variations, such as synonyms, but the broader context of the sentence is overlooked. The variability in scores is even lower than with BLEU, often lacking a clear cut-off point.

G-Eval (Liu et al., 2023a) In the absence of reference texts, an LLM as evaluator offers flexibility similar to human evaluation, as the model can be provided with specific criteria for assessment. For this reason, we applied G-Eval to complex, open-ended questions in which the user asks the chatbot

[3] A BLEU score of 30 was selected as the upper threshold, as the chatbot responses include additional text beyond the retrieved definition. In our tests, a score of 30 consistently indicated high-quality, complete definitions.

for advice, given the data shown in the dashboard. We used GPT-4o as the backbone LLM for this.

Within this task, the chatbot performs function calling to aggregate data from different parts of the model and provides a recommendation. One of the developers of the chatbot evaluated the chatbot's output according to a list of criteria. We then used the same criteria within G-Eval (Liu et al., 2023a). We focused on the advice part of the response, where the chatbot recommends which scores for terms associated with a broader category should be improved. A requirement was that the chatbot must not generate any information that cannot be derived from the dashboard data. We executed this evaluation repeatedly over the same set of four different responses from the chatbot to the same question. We observed that the scores given by G-Eval are not consistent across the different executions. We also saw that G-Eval did not always comply with the predefined criteria.

5. Lessons Learned

This section summarises the key lessons learned from the literature review, survey, and use case. We share a set of best practices to help developers design effective evaluation strategies for LLM evaluation.

1. The first step for LLM evaluation is to clearly define the evaluation task. This starts with establishing the goal: different objectives require distinct evaluation methods, as we observed in Section 4.2.

2. Identify and describe the end-user of the generated text that is being evaluated.

3. Based on the evaluation description and end-user, choose a suitable evaluation method and metric. When possible, it is advisable to include a form of human evaluation, especially for complex generated answers, or to ensure that the evaluation method performs as expected.

4. Finally, document the choice for the evaluation method and the reasons behind the choice, so that it can be reviewed and adjusted if the evaluation scenario changes in the future. Note that it is also wise to consider intermediate outputs and the number of prompts or the prompting technique needed to reach the final correct answer.

There are several key considerations throughout this process. When using references or a ground truth, these should be of the highest quality possible (i.e., reputable source, verified by domain experts, in line with the defined task). When using non-deterministic metrics (i.e. LLM as evaluator), scores should be computed several times due to potential inconsistencies. Averages and standard deviations should be computed to make the outcomes more robust.

Finally, based on our analysis, survey, and use case, we conclude that human evaluation remains essential for assessing LLM performance. The use case demonstrated that evaluation metrics can be combined and tailored to the specific task or domain. However, even well-designed metrics cannot capture all relevant aspects of model performance. Automated benchmarks have limitations that cannot easily be overcome, making it unwise to rely solely on their outcomes. Instead, meaningful evaluation should combine task-specific metrics with human judgment in real-world contexts, ensuring a more comprehensive and reliable assessment.

6. Conclusion

The evaluation of LLMs is a crucial topic given their rapid development and widespread adoption across domains. As the number of LLM-based applications grows, the need for robust evaluation frameworks becomes evident, not only to guide development and to enable meaningful comparisons between models, but also to ensure their reliability and correctness in critical applications.

In this work, we gathered insights and lessons learned both from scientific literature and from practice. We outlined literature on the evaluation of LLMs and gave a categorisation of the main challenges in LLM evaluation, with a focus on LLM intrinsic, human-related, and automatic evaluation challenges. Additionally, we presented a categorisation of popular and available text generation metrics.

To gain insights into challenges from practice, we conducted a survey targeted to LLM developers. The results show that the main problems mentioned are the absence of references or a ground truth and the difficulty of finding trustworthy benchmarks. According to the respondents, integrating automatic methods with human expertise in evaluation is a relevant direction of future research.

In applying LLM evaluation metrics to a real use case, we found that starting with the simplest metrics that best fit the task and available resources was effective. More sophisticated metrics like BERTScore (Zhang et al., 2019) do not necessarily provide better evaluation quality than simpler ones like BLEU (Papineni et al., 2002) or even string matching, depending on the evaluation's goal. While more flexible metrics like G-Eval (Liu et al., 2023a) can be powerful when no reference text is available, they are difficult to control in terms of consistency and require clearly defined criteria. Ultimately, we found that combining simpler and sophisticated metrics strikes a good balance between evaluation performance, explainability, and resource efficiency. When paired with human evaluation, such a combination can reduce the amount of manual work while still ensuring a proper and reliable assessment of model performance.

For future work, it would be valuable to test more metrics and combinations of them in a wider array of practical use cases to further highlight evaluation challenges beyond benchmarks and scientific tasks. Expanding metrics to address bias, fairness, and sustainability is critical, given that current frameworks fail to adequately capture these dimensions. Finally, greater transparency from language models about their evaluation processes is needed to facilitate more robust and interpretable assessments.

7. Limitations

This study has several limitations. While we reviewed relevant recent literature, we did not conduct a systematic literature review, so some relevant works may have been missed. The goal of this work was to compare scientific challenges to practical insights, rather than to provide a comprehensive overview. Similarly, while we described and categorised commonly used metrics, the list is not exhaustive.

Another limitation is that our findings and lessons learned are based on a small survey and a single use case, which may not fully capture the broader landscape of LLM evaluation challenges. The survey participants were LLM engineers working in applied research, so our best practices may not fully reflect end-user satisfaction or concerns arising in industry deployment contexts. Involving a broader and more diverse group, such as end users or researchers from different domains, could provide additional perspectives.

Regarding the use case, we tested a limited set of metrics on a single application, which may limit the generalisability of our findings. Due to time constraints, we did not perform a comprehensive statistical analysis of the metrics applied to the different tasks performed by the chatbot, which may affect the reliability and robustness of our observations. Furthermore, due to confidentiality constraints, we are unable to share all details of the use case, which impacts transparency and reproducibility.

Finally, although we touch on broader LLM evaluation challenges, this study primarily focuses on text generation. Evaluating tasks like reasoning or retrieval may require different approaches and further investigation.

8. Ethical Considerations

The survey conducted as part of this study was anonymous, and participants provided informed consent before beginning the survey. No personally identifiable information was collected, and responses were used solely for research purposes. As the survey focused on professional experiences with LLMs and did not involve sensitive personal data, no formal ethics board approval was required.

Acknowledgements

The authors extend their gratitude to the GRAIL project for funding this research. In addition, they would like to thank the participants who filled in the survey, and the use cases for gathering lessons learned in practice.

9. Bibliographical References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. Llm stability: A detailed analysis with some surprises.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. 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.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024.

Lessons from the trenches on reproducible evaluation of language models.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. 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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. *arXiv preprint arXiv:2006.14799*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).

Alejandro de León Languré and Mahdi Zareei. 2024. Improving Text Emotion Detection Through Comprehensive Dataset Quality Analysis. *IEEE Access*, 12:166512–166536. Conference Name: IEEE Access.

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.

Jonatas S. Grosman, Pedro H. T. Furtado, Ariane M. B. Rodrigues, Guilherme G. Schardong, Simone D. J. Barbosa, and Hélio C. V. Lopes. 2020. Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*, 93:101553.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations.

Aryan Jadon and Avinash Patil. 2025. A comprehensive survey of evaluation techniques for recommendation systems. In *Computation of Artificial Intelligence and Machine Learning*, pages 281–304, Cham. Springer Nature Switzerland.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

George F. Jenks. 1967. The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography*, 7:186–190.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Prentice Hall.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning Cheaper, Lighter and Faster Evaluation Metrics for Automatic Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRL-Eval: An Unsupervised Reference-Free Metric for Evaluating Controlled Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.

Sunder Ali Khowaja, Parus Khuwaja, Kapal Dev, Weizheng Wang, and Lewis Nkenyereye. 2024. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*, pages 1–23.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1774–1787, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. 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, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao

Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023c. LLM360: Towards Fully Transparent Open-Source LLMs. ArXiv:2312.06550 [cs].

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. 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 5328–5339, Online. Association for Computational Linguistics.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Arbi Haza Nasution and Aytuğ Onan. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, 12:71876–71900. Conference Name: IEEE Access.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A survey on large language model benchmarks.

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.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2024. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024.

How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

June Sallou, Thomas Durieux, and Annibale Panichella. 2024. Breaking the silence: the threats of using llms in software engineering. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, ICSE-NIER'24, page 102–106, New York, NY, USA. Association for Computing Machinery.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Laurène Vaugrante, Mathias Niepert, and Thilo Hagendorff. 2024. A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions. ArXiv preprint arXiv:2409.20303.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36:77013–77042.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023a. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 695–707. Springer.

Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, and Noah A. Smith. 2023b. Transparency Helps Reveal When Language Models Learn Meaning. *Transactions of the Association for Computational Linguistics*, 11:617–634.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. ArXiv:2311.01964 [cs].

Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2024. Human Bias in the Face of AI: The Role of Human Judgement in AI Generated Text Evaluation. *arXiv preprint arXiv:2410.03723*.

A. Appendix A: Survey Questions

Best practices & practical challenges in the evaluation of LLMs As part of our research on the evaluation of Large Language Models (LLMs) we are identifying the challenges with the evaluation of LLMs and the state-of-the-art methods both from scientific and practical perspective. As part of the practical perspective we want to collect both best practices, research and practical challenges developers run into at company XXX. This survey has two goals: 1) Gain insight in how the evaluation is approached in current research on LLMs within company XXX 2) Collect practical challenges that developers/projects run into when evaluating their LLM-based applications The results of this survey will be used anonymised and aggregated (such that they are not traceable to you or your project) in our report and possibly in a scientific paper. By filling this survey you agree with this use of your response. In case of any questions about the survey, please contact XXXX or XXXX Thank you!

General information First we would like to know some general information about your background, experience and sector you are working in.

Q1: What is your background?

- Artificial Intelligence
- Linguistics
- Computer Science
- Mathematics
- Social Sciences
- Other

Q2: What is your experience in working with (Large) Language Models?

- < 1 year
- 1-2 years
- 3-5 years
- 5-10 years
- > 10 years

Q3 For which sector is (most of) your work aimed?

- Health
- Mobility
- Sustainability
- Safety & security
- Government
- Not sector specific
- Other

Your research and applications First we would like to know about your research, in which way do you use LLMs?

Q4: How many projects related to LLMs do/did you work on this year?

- 0
- 1
- 2
- 3
- 4
- 5+

Q5: In what ways do you use LLMs in your research?

- Research on evaluation of LLMs
- Research on increasing performance of LLMs
- Research on integrating LLMs, knowledge and/or tools
- Chatbot
- RAG-application
- Text analysis applications
- Multi-modal applications
- Other

Q6: In which phase is your research? (e.g. problem analysis, design, implementation, evaluation, pilot, deployed) - if you have multiple projects, please answer the question for all of your projects.

Evaluation in practice We would like to hear more about your choices for the evaluation and which challenges you run/ran into?

Q7: Which methods of evaluation do you (aim to) use?

- automatic evaluation (metrics)
- human evaluation
- LLM-based evaluation
- Other

Q8: Can you give a description of the evaluation method you (aim to) use? For example what do you aim to evaluation, which metrics do you use, what do you evaluate with humans?

Q9: Why did you choose this way of evaluation?

Q10: What are challenges that you run/ran into for the evaluation of your LLM?

Q11: What are questions that you have regarding the evaluation of LLMs?

Research on LLM evaluation If you research LLM evaluation specifically, we are curious to hear more about your work.

Q12: Do you research LLM evaluation specifically?

- Yes
- No

Q13: What does your research focus on?

Q14: Do you have results that you can share?

Your opinion on LLM evaluation research? We are curious to hear more about what you think are the most relevant/urgent research topics and challenges.

Q15 What do you think is the most relevant method of evaluation of LLMs? (rank from most to least relevant)

- Human evaluation
- Automatic evaluation
- LLM-based evaluation
- Combination of human and automatic evaluation
- Combination of all three forms of evaluation

Q16: What do you think is the biggest challenge in LLM evaluation (rank form biggest to smalles challenge)

- Robustness, i.e. the ability to produce the similar results for (small) variations in prompts and orders
- Reproducibility, i.e. the ability to produce the same result multiple times
- Data quality, i.e. achieving a qualitative ground truth data set
- Data Leakage, i.e. the fact that test data/ test scenario's might be included in the training data
- Lack of universally accepted benchmarks, i.e. large variety of benchmarks that all have limited variability in prompts/scenario's, are biased to English data, do not consider alternative answers.
- Subjectivity of humans, i.e. different humans give different answers in both annotation and evaluation
- Fairness evaluation, i.e. challenges in gaining insight in biases of the models
- Sustainability, i.e. the energy usage of these models during evaluation and deployment
- Lack of model transparency, i.e. no/limited access or insight in training data and weights
- Other challenge, i.e. a challenge you encounter that is not in this list

Q17: If you ranked 'Other challenge' in the previous question, please let us know which challenge you mean?

Q18: What solution or research would you be helped with?

Next steps This survey is meant as a way to collect more general insights, but does not allow for in-depth discussions. We might be interested to discuss your research and challenges in more detail. These are the last two questions of the survey, thank you very much for helping us and don't forget to submit! :)

Q19: If you are open to a discussion, please leave your email here or send a message to XXX or XXX in case you don't want your answers linked to you.

Q20: If you have any other comments or questions you can leave them here.

B. Appendix B: Survey Results

Resources - Total: 14
15. General guidelines for conducting evaluations. 16. Identification of state-of-the-art evaluation techniques. 20. Methods, frameworks, or guidelines for human/manual evaluation. 23. Guidelines for tailoring metrics to specific use cases. 24. Guidance on integrating task-specific and general evaluation methods. 25. A structured approach to evaluation. 26. Guidance on identifying domain experts for evaluation. 30. Recommendations for effective prompting strategies. 32. Human-in-the-loop evaluation methodologies. 34. An always up-to-date list of available evaluation tools and methods. 35. Guidance on when to use each evaluation method. 36. Clear definitions of text generation tasks to support evaluation. 38. Relevant and rigorous scientific literature on LLMs as evaluators. 43. Guidelines for handling LLM inconsistencies, multiple correct answers, hallucinations, and instruction-following failures.
Datasets - Total: 5
2. Transparency about data contamination in evaluation datasets. 4. An accessible overview of available benchmarks and their quality or acceptance level. 5. Support for creating custom benchmarks. 27. Transparency about the quality of reference data (ground truth). 41. Methods for evaluating without references or ground truth.
Metrics - Total: 19
3. Consistency metrics and repetition guidance tailored to each task. 7. Support for designing custom metrics with heuristic guidance. 8. Specification of the domain or range of applicability for each metric. 9. Visibility into the computational cost of each metric. 10. Clear interpretation of evaluation metrics. 11. Information on the robustness of metrics. 12. Metrics that prioritize correctness over fluency. 13. Frameworks that include multiple evaluation dimensions (e.g., factuality, correctness, conciseness, creativity, bias, interpretability, tone). 14. Flexibility in selecting metrics across different dimensions. 17. Specific metrics for evaluating Retrieval-Augmented Generation (RAG) systems. 18. Indication of how well metrics correlate with human judgment. 19. Guidance on which metrics are suitable under different budget constraints. 22. Mapping between evaluation accuracy (e.g., repetitions needed for significant scores) and environmental cost. 28. Metrics for evaluating the efficiency of LLM responses. 29. Guidance on handling fuzzy matching in evaluation. 31. Indication of speed, accuracy, and completeness of metrics. 39. Capabilities for multilingual evaluation. 40. Indication of which evaluation aspects are prerequisites for others. 42. Insights into the generalizability of metrics.
System - Total: 3
1. A clear understanding of the required quality level for evaluation. 6. Clarity on the intended end user of each metric or evaluation method. 33. Clarity on when LLM-based evaluators are appropriate or not.
Governance - Total: 2
21. Consideration of the environmental impact (e.g., carbon footprint) of metrics. 37. Scoring of evaluation methods based on alignment with EU and Dutch values.

Table 4: Needs for LLM evaluation in practice reported by survey respondents, and grouped under the themes Resources, Datasets, Metrics, System, and Governance.

Is Human–LLM Interaction Culture-Dependent? A Cross-Linguistic NLP Analysis of Student Interviews on AI-Assisted Thesis Writing

Madalina Chitez[1], Karla Csuros[1], Dejana Jelena Milićević[2], Petya Osenova[3], Stefan Marinov[3], Teodor Valchev[3], Nikolay Paev[3], Otto Kruse[2], Christian Rapp[2], Andreea Dinca[1], Roxana Rogobete[1], Claudia Doroholschi[1], Loredana Punga[1], Anabella Costache[1], Dumitru Tucan[1], Cristina Baniceru[1]
[1]West University of Timisoara
[2]Zurich University of Applied Sciences
[3]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
madalina.chitez@e-uvt.ro

Abstract

This study investigates whether human–LLM interaction in academic writing exhibits cross-cultural variation. Using NLP-informed corpus methods, we analyze nine semi-structured student interviews from three national contexts (Romania, Bulgaria, Switzerland) to examine how AI use is linguistically constructed across three dimensions of epistemic positioning: agency strength, authority dynamics, and discourse-level stance. Results show a strong predominance of distancing and hedging strategies, with AI consistently framed as a functional writing support tool rather than an epistemic authority. At the same time, modest but systematic cross-country differences indicate culturally embedded variation in how students discursively negotiate epistemic responsibility and evaluation in AI-assisted writing practices.

Keywords: Human–LLM interaction, Epistemic positioning, Cross-linguistic discourse analysis, AI-assisted writing, Student interviews, Academic thesis writing, Corpus-informed NLP

1. Introduction

Academic thesis writing is a central higher-education practice through which students construct and justify knowledge rather than reproduce information. Writing functions not only as communication but also as a cognitive process that supports claim formulation, evidence evaluation, and epistemic responsibility (Bean, 2011; Hofer and Pintrich, 1997). At the same time, academic writing conventions are shaped by cultural and educational traditions described in contrastive rhetoric (Kaplan, 1966) and subsequent discourse research (Connor, 2002). The rapid integration of large language models (LLMs) introduces a new dimension to these culturally embedded practices. While AI tools increasingly support idea generation and text formulation, they lack epistemic awareness and cannot assume responsibility for knowledge claims. Human–AI collaboration may therefore reflect culturally situated norms of authorship, intellectual autonomy, and knowledge evaluation. Despite growing research on AI-assisted writing, little is known about whether human–LLM interaction itself exhibits cross-cultural variation in academic discourse, as most existing studies rely primarily on surveys or experimental tasks rather than natural language data. To address this gap, the study proposes semi-structured student interviews as a cross-linguistic discourse resource suitable for corpus-informed NLP analysis. We examine how human–LLM interaction is linguistically constructed across three dimensions of epistemic positioning: epistemic agency, authority dynamics, and discourse-level stance toward AI outputs. By focusing on observable linguistic patterns, the study provides preliminary corpus-based insights into how students discursively position AI within academic knowledge construction across national contexts.

2. Related Work

2.1. Cross-Cultural Academic Writing and Digital Writing Contexts

Research in contrastive rhetoric shows that academic writing conventions vary across linguistic and educational traditions (Kaplan, 1966). Subsequent discourse studies document differences in textual organization, authorial positioning, and argumentative structure across academic cultures (Clyne, 1987; Connor, 2002). Corpus-informed work further highlights distinctive writing practices in Central and Eastern Europe shaped by historical educational traditions (Chitez et al., 2018). In the Romanian context, analyses of BA theses reveal systematic differences in rhetorical structuring and epistemic positioning (Băniceru et al., 2012), which continue to influence students' English academic writing, particularly in stance expression (Bercuci and Chitez, 2023). At the same time, research on digital writing environments shows that writing technologies reshape knowledge construction processes by supporting iterative drafting, feedback in-

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 39–44
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

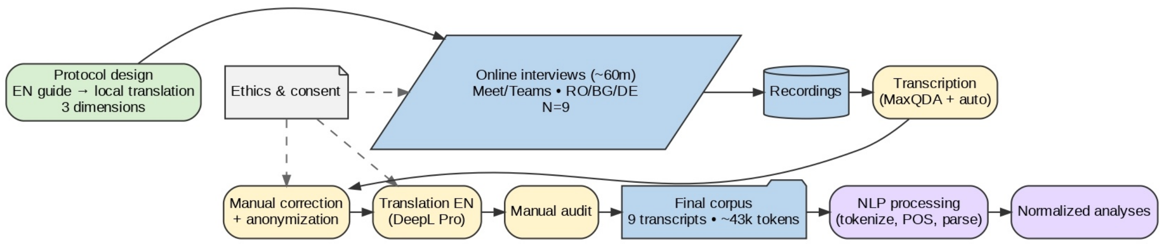


Figure 1: Pipeline

tegration, and interaction with information sources (Kruse and Rapp, 2019). Broader syntheses indicate that digital tools influence epistemic engagement by mediating how students evaluate and assume responsibility for knowledge claims (Kruse et al., 2023).

2.2. Epistemic Positioning and AI-Assisted Writing

Corpus-based research demonstrates that epistemic positioning is systematically encoded through recurrent lexico-grammatical patterns such as stance verbs, modality, and complement structures (Biber et al., 1999). These patterns have been operationalized in corpus and NLP studies to model certainty, evaluation, and authority in academic discourse (Biber, 2006; Biber et al., 2004). Such approaches provide a methodological basis for computational analyses of epistemic agency and stance. Recent work on AI-assisted writing shows that generative systems increasingly support planning, drafting, and revision processes in higher education (Strobl et al., 2019; Imran and Almusharraf, 2023). Empirical syntheses further indicate that generative AI reshapes authorship dynamics and knowledge-construction practices (Sanz-Tejeda et al., 2026). However, NLP research highlights persistent epistemic challenges, as language models may express certainty without reliable evidential grounding (Ghafouri et al., 2024).

3. Data and Methods

3.1. Data collection pipeline

We conducted nine semi-structured interviews across three national contexts (Romania, Bulgaria, Switzerland; three participants per site). Interviews were carried out in participants' native languages (Romanian, Bulgarian, German) to support accurate reflection on writing practices. To ensure cross-context comparability, the interview guide was collaboratively designed in English and then translated locally; it covered (i) intellectual development, (ii) thesis writing practices, and (iii) the role of generative AI. Interviews were held online (Google Meet / Microsoft Teams), lasting approximately one hour each. Recordings were transcribed using MaxQDA (GmbH, 2026), for the Romanian and Swiss datasets, and an in-house model based on the Open AI Whisper speech-to-text model[1] for the Bulgarian dataset, with subsequent intensive manual correction (approx. 4 hours per interview), during which all identifying information was removed to ensure full anonymization. The verified transcripts were translated into English using DeepL Pro (DeepL, 2026), and translations were manually audited to preserve meaning and reduce interpretive drift. Ethical safeguards (see Figure 1) included written informed consent, secure storage and restricted access to recordings/transcripts, and the use of privacy-compliant tools for transcription and translation.

3.2. Dataset

The dataset consists of nine anonymized semi-structured interview transcripts, totaling approximately 43,000 tokens in English translation. The corpus is evenly distributed across three national contexts, Bulgaria, Switzerland, and Romania, with three interviews included in each sub-corpus. However, substantial variation exists in document length. Bulgarian interviews range from approximately 4,100 to 4,600 tokens, Swiss interviews from about 3,200 to 5,200 tokens, and Romanian interviews from roughly 3,400 to 9,600 tokens, the latter showing the greatest internal variability and containing the longest individual transcript in the dataset. Given this heterogeneity in transcript size, all quantitative analyses were conducted using normalized frequency measures to ensure comparability across interviews and national sub-corpora.

3.3. Computational Pipeline and Operationalization

All computational linguistic processing was conducted using the Stanza NLP library (version 1.11.0) configured for English. We utilized the `combined` models for tokenization and

[1] https://github.com/openai/whisper

multi-word token (MWT) expansion, the `combined_nocharlm` model for lemmatization, and the `combined_charlm` models for part-of-speech (POS) tagging and dependency parsing.

To mitigate parsing and tagging errors inherent in translated conversational discourse, we implemented a multi-stage validation protocol. First, spoken-language artifacts (e.g., timestamps, speaker labels, platform artifacts) were programmatically removed using regular expressions to prevent dependency parsing disruption. The transcripts, which were translated into English using DeepL Pro, were manually audited prior to NLP processing to preserve semantic meaning and reduce interpretive drift. Following the automated extraction of first-person agency constructions, a random sample ($N = 15$) of the dependency-parsed outputs was extracted. This subset underwent manual verification by the research team to confirm the accuracy of the first-person subject tagging, verb lemmatization, and the subsequent semantic class mapping, ensuring algorithmic reliability before conducting the full corpus analysis.

Within this pipeline, AI-related contexts were operationalized using a dictionary-based sentence-window approach. We defined an explicit lexicon of target terms: *chatgpt, gpt, gpt-4, openai, copilot, claude, gemini, bard, llm,* and *generative ai*. The transcript texts were lowercased, and sentences were flagged if they contained any of these strings. To accurately link these mentions to student agency, we employed a co-occurrence window of ± 1 sentence. Specifically, a first-person agency verb was classified as "AI-related" if a target term appeared in the exact same sentence, the immediately preceding sentence, or the immediately following sentence. While this approach prioritizes precision over recall, it introduces specific edge cases. Primarily, it misses broader co-references; if a participant names an AI tool and subsequently refers to it as "the tool" or "it" outside the ± 1 sentence window, the associated agency verbs are not captured. Furthermore, because the algorithm relies on direct string matching, there is a minor risk of substring collisions, though the specificity of the chosen lexicon largely mitigates this risk.

4. Results

4.1. Agency Strength in Human–LLM Interaction

Weak or distancing agency markers overwhelmingly dominate the corpus, accounting for between 86.9% and 90.3% of all constructions across national contexts. These include hedging verbs such as *think* and procedural verbs such as *use*, which primarily describe AI interaction in terms of assistance rather than knowledge evaluation. Students frequently frame their interaction through statements such as "I use AI mainly for translation," emphasizing functional support rather than epistemic authority. Moderate agency markers, representing between 4.8% and 7.6% of instances, include inquiry and evaluative verbs such as *ask* and *consider*, which signal active engagement without strong epistemic commitment. These constructions typically describe exploratory interaction, as illustrated by expressions such as "I ask whether the text makes sense." Strong agency markers remain relatively rare, accounting for only 4.9% to 6.8% of constructions. These include verification and control verbs such as *check*, *decide*, and *adjust*, which explicitly signal responsibility for evaluating AI outputs. When such verbs occur, they consistently emphasize human oversight, for example in statements such as "I check whether the information is accurate."

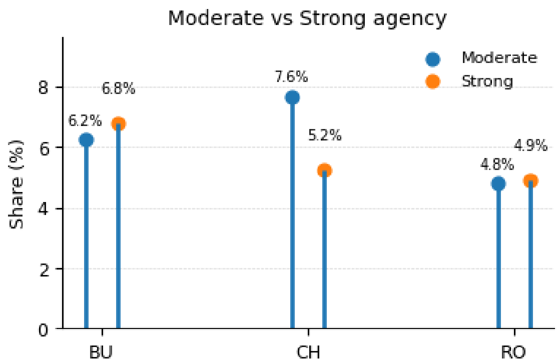


Figure 2: Epistemic positioning

Cross-country differences are present but modest. Swiss interviews show the highest proportion of moderate agency markers (7.6%), suggesting a slightly stronger orientation toward inquiry and evaluative engagement. Bulgarian interviews exhibit the highest proportion of strong agency markers (6.8%), indicating a somewhat greater tendency to linguistically express verification or decision-making processes. Romanian interviews display the lowest levels of both moderate (4.8%) and strong (4.9%) agency markers, reflecting a stronger reliance on distancing or procedural descriptions.

4.2. Authority Dynamics in AI-Related Discourse

To examine how epistemic authority is linguistically constructed when students explicitly refer to AI, we analyzed agency markers occurring in AI-related contexts and grouped them into delegation, hedging, and supervision functions. Across all national sub-corpora (Figure 3), delegation markers overwhelmingly dominate AI-related discourse. They account for approximately 65% of epistemic expressions in Romania, 70% in Bulgaria, and 74% in Switzerland. These patterns indicate that students primarily describe AI interaction through task-oriented verbs such as *use*, *write*, and *find*, which frame AI as a functional writing assistant rather than a knowledge source requiring evaluation.

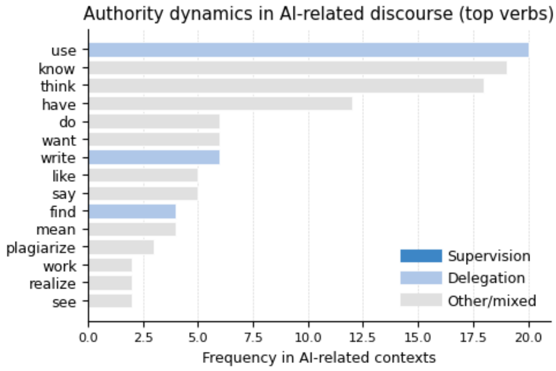


Figure 3: Authority verbs (all data)

Hedging markers constitute a secondary layer of epistemic positioning, representing roughly 27–30% of AI-related expressions across countries. This suggests that when discussing AI, students frequently maintain linguistic distance from the reliability of its outputs, signaling uncertainty or caution rather than authority. In contrast, supervision markers remain consistently low across contexts. They account for only 7.7% of AI-related agency constructions in Romania, 9.8% in Bulgaria, and 7.5% in Switzerland. These markers include verbs associated with verification, correction, and decision-making, and their limited presence indicates that explicit linguistic expressions of epistemic control over AI outputs are relatively rare. Cross-country variation (Figure 4) is modest but systematic. Bulgarian interviews show the highest relative proportion of supervision markers, suggesting a slightly stronger orientation toward linguistic expressions of oversight. Swiss interviews display the strongest dominance of delegation markers, indicating a more pronounced framing of AI as a task-execution tool. Romanian interviews fall between these patterns but show the highest relative presence of hedging, reflecting a somewhat stronger tendency toward epistemic distancing.

4.3. Discourse-Level Epistemic Stance Toward AI

The epistemic stance categories in Figure 5 capture three complementary dimensions of how students linguistically position knowledge in AI-related discourse. *Hedging markers* include modal verbs and stance expressions (e.g., *may*, *might*, *perhaps*, *I think*) that signal epistemic uncertainty and reduced

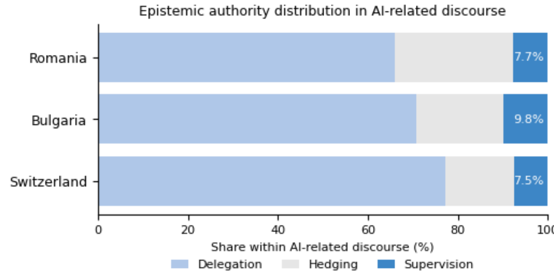


Figure 4: Authority verbs (per country)

commitment, consistent with corpus-linguistic accounts of modality (Biber et al., 1999; Biber, 2006). *Attribution markers* assign information to external agents (e.g., “AI says,” “it suggests”), thereby shifting epistemic authority away from the speaker. *Evaluation markers* consist of adjectives expressing judgments of informational quality (e.g., *accurate*, *helpful*, *problematic*). Importantly, hedging is operationalized here at the discourse level rather than as first-person epistemic verbs used in earlier agency analyses. Thus, the two measures capture complementary aspects of epistemic positioning: agency-related hedging reflects responsibility for knowledge claims, whereas modality-based hedging reflects degrees of certainty toward information.

Discourse-level stance markers show a clear predominance of hedging across all national subcorpora (normalized per 1,000 words). Hedging occurs at approximately 6.9 in Bulgarian interviews and 9.3–9.4 in Swiss and Romanian datasets, making it the most frequent epistemic feature. Attribution markers remain rare (0.23–1.29), while evaluation markers occur at intermediate levels (4.33–6.31). These patterns indicate that AI-related discourse is characterized primarily by epistemic uncertainty rather than explicit attribution or evaluation. Cross-country differences are modest but consistent: Swiss interviews show the highest levels of hedging and evaluation, whereas Bulgarian interviews display the lowest attribution rates, suggesting that students generally frame AI-supported writing through caution rather than trust or authority.

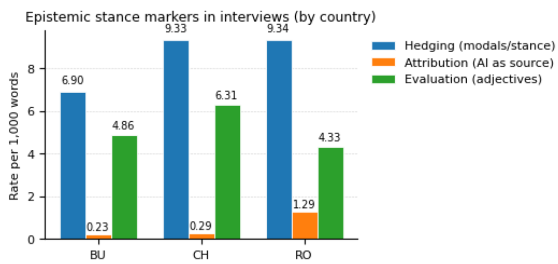


Figure 5: Epistemic stance markers

5. Discussion

This exploratory study provides corpus-based evidence that human–LLM interaction in academic writing is linguistically constructed through systematic patterns of epistemic positioning. Across contexts, students predominantly use distancing and hedging strategies, framing AI as a functional support tool rather than an epistemic authority. At the same time, modest but consistent cross-country differences suggest culturally shaped variation in how epistemic responsibility is discursively negotiated. Higher levels of verification language in the Bulgarian data and stronger evaluative engagement in the Swiss interviews may reflect differing academic socialization practices regarding knowledge control, autonomy, and critical assessment. In contrast, the Romanian interviews show a stronger reliance on procedural and distancing formulations, which may indicate a greater tendency to frame AI as a technical writing aid rather than an epistemic partner, consistent with previous findings on locally shaped academic discourse conventions (Kaplan, 1966; Connor, 2002).

Methodologically, the study demonstrates the value of combining qualitative interview data with corpus-informed, computationally tractable feature extraction to analyze epistemic positioning in natural discourse rather than self-reported attitudes. However, the findings remain exploratory due to the small sample size, reliance on translated transcripts, and the focus on reported rather than observed writing practices. Future research should extend this approach to larger multilingual datasets, integrate writing-process data, and develop automated stance-detection methods to further investigate culturally embedded patterns of human–AI collaboration.

5.1. Limitations and Robustness

The findings regarding agency and epistemic authority are highly sensitive to the strict syntactic operationalization employed in this study. The extraction algorithm is constrained to explicit first-person pronouns (e.g., *I, we, my*) that hold an exact nominal subject (`nsubj`) dependency relation to a head token tagged as a verb (`VERB`) or auxiliary (`AUX`). Consequently, this operationalization is conservative. It successfully isolates explicit claims of agency but does not parse complex evaluative multi-word expressions (e.g., "I am of the opinion that") or passive constructions ("It was checked by me"), which may result in an underestimation of total epistemic positioning. Furthermore, shifting to a modality-only operationalization, which captures adverbs or modal verbs (e.g., *may, might, perhaps*) regardless of the grammatical subject, yields distinct frequency distributions. This complementary dynamic is evident in our data, where discourse-level stance markers highlight uncertainty, while first-person agency markers emphasize functional delegation.

Additionally, because the current computational pipeline processes the English translations exclusively, systemic shifts in modality or hedging introduced by the DeepL translation algorithm constitute a limitation of the current findings. To rigorously estimate these translation effects, future iterations of this methodology should leverage multilingual NLP capabilities. By instantiating native language pipelines for Romanian, Bulgarian, and German, researchers could execute the identical dependency extraction on the original source-language transcripts. Mapping these native verbs to our epistemic classes and comparing their normalized frequencies against the English-translated results would allow for a quantitative assessment of whether machine translation processes artificially inflate or deflate specific markers of stance and agency across national sub-corpora.

6. Conclusions

This study shows that human–LLM interaction in academic writing is systematically shaped by patterns of epistemic positioning, with students across contexts predominantly framing AI as a functional support tool rather than an epistemic authority. At the same time, cross-country differences indicate that AI-supported writing practices remain embedded in culturally specific norms of knowledge construction, responsibility, and evaluation. These findings highlight the importance of analyzing human–AI interaction through multilingual discourse data and interpretable linguistic features, demonstrating how language technologies can support empirical investigations of socially situated communication practices. By operationalizing epistemic positioning in naturalistic interview discourse, the study illustrates how computational methods can provide theoretically grounded tools for examining emerging forms of AI-mediated knowledge work in academic and research contexts. Future research can extend this approach to larger multilingual datasets, incorporate writing-process and multimodal evidence, and develop automated stance-detection methods to further support interdisciplinary investigations of human–AI collaboration.

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF) under the Multilateral Academic Projects (MAPS) grant (No. IZ11Z0_230816), awarded for the project "Exploring the Triad of Academic Writing, Critical Thinking and AI Literacy: Interdisciplinary Perspectives on

Undergraduate Thesis Writing" (TRAI), co-financed by the Romanian Ministry of Education and Research through UEFISCDI (Grant No. 15ROCH / 01/07/2025) and by the Bulgarian Ministry of Education and Science (Reg N D03-34/30.07.2025).

7. Bibliographical References

John C. Bean. 2011. *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom*, 2 edition. Jossey-Bass.

L. Bercuci and M. Chitez. 2023. The influence of the romanian academic style on student writing in english. In M. Vránová, editor, *New to the LSP classroom? A selection of monographs on successful practices*, pages 169–185. Vernon Press.

D. Biber. 2006. *University language: A corpus-based study of spoken and written registers*. John Benjamins.

D. Biber, S. Conrad, and V. Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written English*. Longman.

C. Băniceru, M. I. Borchin, C. Doroholschi, and D. Tucan. 2012. Academic writing in romania: A contrastive analysis of ba thesis introductions in romanian and english. *Quaestiones Romanicae*, 1:331–344.

M. Chitez, C. I. Doroholschi, O. Kruse, Ł. Salski, and D. Tucan, editors. 2018. *University writing in Central and Eastern Europe: Tradition, transition, and innovation*. Springer.

Michael Clyne. 1987. Cultural differences in the organization of academic texts: English and German. *Journal of Pragmatics*, 11(2):211–241.

Ulla Connor. 2002. New directions in contrastive rhetoric. *TESOL Quarterly*, 36(4):493–510.

DeepL. 2026. Deepl translate.

B. Ghafouri, S. Mohammadzadeh, J. Zhou, P. Nair, J. J. Tian, H. Tsujimura, and K. Pelrine. 2024. Epistemic integrity in large language models.

VERBI GmbH. 2026. Maxqda.

Barbara K. Hofer and Paul R. Pintrich. 1997. The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67(1):88–140.

Muhammad Imran and Norah Almusharraf. 2023. Analyzing the role of chatgpt as a writing assistant at higher education level: A systematic review. *Contemporary Educational Technology*, 15(4):ep464.

Robert B. Kaplan. 1966. Cultural thought patterns in inter-cultural education. *Language Learning*, 16(1–2):1–20.

O. Kruse and C. Rapp. 2019. Seamless writing: How the digitisation of writing transforms thinking, communication, and student learning. In *Seamless learning: Perspectives, challenges and opportunities*, pages 191–208. Springer.

O. Kruse, C. Rapp, C. M. Anson, K. Benetos, E. Cotos, A. Devitt, and A. Shibani. 2023. *Digital writing technologies in higher education: Theory, research, and practice*. Springer.

Aránzazu Sanz-Tejeda, Juana Celia Domínguez-Oller, Josep Maria Baldaquí-Escandell, Raquel Gómez-Díaz, and Araceli García-Rodríguez. 2026. The impact of generative AI on academic reading and writing: a synthesis of recent evidence (2023–2025). *Frontiers in Education*, 10:1511718.

Carolin Strobl, Emily Ailhaud, Katerina Benetos, Amy Devitt, Otto Kruse, Antje Proske, and Christian Rapp. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, 131:33–48.

Next Reply Prediction X (NRP-X) Dataset: Linguistic Discrepancies in Naively Generated Content

Simon Münker[1], **Nils Schwager**[1], **Kai Kugler**[1], **Michael Heseltine**[2], **Achim Rettinger**[1]

[1] Trier University, Computational Linguistics
Universitätsring 15, Trier, Germany
{muenker, schwager, kuglerk, rettinger}@uni-trier.de

[2] University of Oxford, Sociology
42-43 Park End Street, Oxford, England
michael.heseltine@sociology.ox.ac.uk

Abstract

The increasing use of Large Language Models (LLMs) as proxies for human participants in social science research presents a promising, yet methodologically risky, paradigm shift. While LLMs offer scalability and cost-efficiency, their "naive" application, where they are prompted to generate content without explicit behavioral constraints, introduces significant linguistic discrepancies that challenge the validity of research findings. This paper addresses these limitations by introducing a novel, history-conditioned reply prediction task on authentic X (formerly Twitter) data, to create a dataset designed to evaluate the linguistic output of LLMs against human-generated content. We analyze these discrepancies using stylistic and content-based metrics, providing a quantitative framework for researchers to assess the quality and authenticity of synthetic data. Our findings highlight the need for more sophisticated prompting techniques and specialized datasets to ensure that LLM-generated content accurately reflects the complex linguistic patterns of human communication, thereby improving the validity of computational social science studies.

Keywords: human simulacra, synthetic content, linguistic authenticity

1. Introduction

The widespread adoption of Large Language Models (LLMs) began with the release of ChatGPT and similar conversational AI systems, fundamentally transforming how humans interact with artificial intelligence (Aïmeur et al., 2023). This technological advancement initiated a paradigm shift in computational social science research, where LLMs are increasingly deployed to simulate or substitute for human participants in behavioral studies (Park et al., 2023; Pérez et al., 2023). The appeal is clear: researchers can use LLMs to efficiently generate responses that would traditionally come from human participants, addressing challenges in scalability, cost, recruitment, retention, and ethical considerations.

However, this anthropomorphic perspective introduces significant methodological risks, particularly when researchers employ naive applications that rely exclusively on prompt engineering without adequate consideration of underlying model limitations, training biases, and domain-specific validation requirements (Larooij and Törnberg, 2025). The quality and representativeness of training datasets become critically important as grounding mechanisms, especially for socially sensitive tasks where cultural nuance, contextual understanding, and authentic human judgment remain central to meaningful analysis. While earlier concerns focused on detecting artificial or malicious content, contemporary LLMs produce increasingly sophisticated outputs that superficially mimic human communication patterns (Crothers et al., 2023). This evolution makes validation more critical yet paradoxically more challenging: the better LLMs become at generating plausible content, the more crucial it becomes to understand where and how they diverge from authentic human behavior.

Our paper addresses a fundamental question at the intersection of natural language processing and computational social science: **Can current LLMs reliably replicate authentic human social media behavior patterns when tasked with user modeling applications?** This question becomes particularly pressing given the growing reliance on synthetic data in computational social science (Burgard et al., 2017), where the assumption of authentic human-like generation underpins the validity of research findings. We approach this question through a systematic comparison between genuine X content and synthetic posts generated through both prompt-based and fine-tuned approaches, examining linguistic discrepancies across multiple analytical dimensions.

1.1. Research Questions and Hypotheses

We investigate three primary research questions using a self-collected German and English X dataset:

RQ1 To what extent do LLM-generated social media posts exhibit detectable linguistic patterns that distinguish them from authentic human content across quantitative, morphological, and semantic dimensions?

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 45–56
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

RQ2 How does fine-tuning on domain-specific social media data improve the linguistic authenticity of generated content compared to prompt-based generation approaches?

RQ3 Can machine learning classifiers reliably distinguish between human and synthetic social media content, and what features prove most discriminative?

Building on empirical evidence from related work on LLM limitations in social simulation (Liu et al., 2022; Hershcovich et al., 2022; Münker et al., 2026), we hypothesize that while LLMs can produce individually plausible social media posts, systematic analysis reveals consistent linguistic signatures that enable reliable detection of synthetic content. Furthermore, we anticipate that fine-tuned models will show reduced but still detectable deviation patterns compared to prompt-based approaches. Related research confirms that fine-tuned models outperform prompt-based approaches in social simulations (Lin, 2024) and text annotation tasks (Alizadeh et al., 2025) in human-LLM alignment.

1.2. Our Contributions

Our work makes three primary contributions to the language resources and evaluation community:

1. We publish a history-conditioned Reply Prediction Dataset for X content (NRP-X dataset), comprising human posts with corresponding synthetic generations using both prompt-based and fine-tuned approaches across English and German languages. (Sec. 3.1)
2. We present a multi-dimensional evaluation framework combining multiple layers of quantitative linguistics analysis to assess human-machine linguistic alignment. (Sec. 3.2)
3. We conduct a comparison of encoder (tf-idf, static dense, transformer) and feature (see above) combinations for detecting the synthetic content. (Sec. 3.3)

2. Background

2.1. LLMs as Human Simulacra

The emergence of Large Language Models has fundamentally transformed computational social science research, with contemporary studies increasingly positioning LLMs as human simulacra (Park et al., 2023) capable of simulating complex user behaviors through sophisticated text-based interaction (Larooij and Törnberg, 2025; Münker et al., 2026). This paradigm shift offers compelling advantages including cost reduction, ethical compliance, and enhanced scalability for large-scale behavioral studies (Pérez et al., 2023; Thapa et al., 2025).

However, empirical validation reveals significant limitations in the authenticity of LLM-generated social behavior. Studies demonstrate systematic biases in the diversity of political (Liu et al., 2022; Münker, 2025b) and cultural (Hershcovich et al., 2022; Münker, 2025a) positions represented in current LLMs. These limitations challenge the prevailing assumption that large language models (LLMs) can function as reliable human proxies, especially when researchers implement applications that depend solely on prompt engineering without sufficient attention to the model's inherent limitations, training biases, and the need for domain-specific validation.

The anthropomorphic perspective introduces methodological risks that become especially problematic in applications requiring nuanced social understanding. While individual LLM-generated texts may appear plausible, systematic analysis often reveals consistent linguistic signatures that distinguish synthetic from authentic content. This detectability gap has important implications for the ecological validity of LLM-based simulations in social research contexts, where the assumption of authentic human-like generation underpins the validity of research findings.

2.2. Synthetic Content Detection in Social Media

The field of synthetic content detection has evolved significantly alongside advances in generation capabilities. Traditional approaches to misinformation detection on social media platforms target artificial or malicious content from regular users (Yang et al., 2019) and develop comprehensive bot detection systems (Hayawi et al., 2023). However, the current generation of LLMs produces increasingly sophisticated outputs that closely mimic human communication patterns, making detection more challenging and validation more critical.

Recent advances in AI-generated content detection (Chong et al., 2023; Abburi et al., 2024) reveal that even sophisticated generation techniques exhibit systematic linguistic patterns across multiple dimensions. These patterns manifest in quantitative features (complexity, readability, lexical diversity), morphosyntactic structures (part-of-speech distributions, syntactic complexity), and semantic distributions (topic diversity, emotion patterns, sentiment biases). The persistence of these linguistic signatures across different generation approaches suggests fundamental limitations in current language modeling techniques for authentic social media simulation.

Our work motivates a shift toward multi-dimensional evaluation frameworks that capture the full spectrum of linguistic differences between human and synthetic content. Surface-level plausibility assessments prove insufficient for validating LLM-generated social media content, necessitating comprehensive protocols that examine linguistic authenticity across quantitative, morphological, and semantic dimensions simultaneously. We build upon these methodological foundations while introducing a novel history-conditioned dataset and systematic comparison of detection approaches, addressing the critical gap between generation capability and authentic behavioral replication.

3. Methods

3.1. Data: Authentic vs. Synthetic

Collection/Preprocessing Our final dataset is based on two raw data dumps – English and German – collected from X. The sets are collected around keywords concerning the political discourses in the US and Germany during the first half of 2023. The samples contain two types of content: a) Tweets (posts) and (b) replies from X users towards these tweets (DE: $3,381,111$, EN: $7,790,741$).

First, we group all first-order replies with the tweets to which they are responding, creating tweet-reply pairs that preserve conversational context. We then reorganize these samples by user, resulting in subsets containing each user's complete reply history along with the original tweets they responded to.

Next, we apply two preprocessing steps to ensure data quality. First, we remove tweet-reply pairs containing URLs (images, GIFs, and links), as these cannot be properly processed by the LLM and the classifiers. Second, we remove users with the highest reply frequencies (DE: 5%, EN: 1%; Quotas result in max DE: 24, EN: 21 samples per user) and split the remaining users into train and test sets. This ensures our analysis captures the model's ability to learn generalizable linguistic styles across the user population, rather than memorizing patterns of individual users.

Transformation We construct a history-conditioned reply prediction Task (Münker et al., 2026; Schwager et al., 2026), using the native instruction-completion format of instruction-LLMs: three tweet-reply pairs as "history", along with a fourth tweet for the model to respond to. We add a system prompt: *You are a social media user responding to conversations. Keep your replies consistent with your previous writing style and the perspectives you have expressed earlier.* This conditions the LLM by presenting the tweet-reply history as if it had already generated those replies during prior turns.

This approach offers three advantages: (1) the model learns from authentic behavioral patterns without hand-crafted features encoding response characteristics; (2) It allows synthetic sample generation without further training only by prompting (3) during fine-tuning, the withheld fourth reply serves as the supervised target.

Fine-Tuning We fine-tune Qwen3 8B (Yang et al., 2025) for each language variant using supervised learning with loss computed exclusively on the last assistant responses. Both training datasets are sub-sampled to 5000 examples. Training uses a warm-up ratio of 0.1 and single-epoch optimization with otherwise default hyperparameters (e.g., learning rate of $2\mathrm{e}{-5}$). Each model trains for approximately 100 minutes on an NVIDIA L40S GPU with 48GB of VRAM.

Generation We generate a single synthetic reply per test prompt using both the base and fine-tuned Qwen3 8B models. Generation uses Qwen3's default sampling parameters (`temperature`: 0.6, `top_k`: 20, `top_p`: 0.9) with a maximum output length of 200 tokens. No post-generation filtering is applied. All model outputs are retained regardless of length, coherence, or formatting.

Published NRP-X Dataset The published dataset *(GitHub repository, see Sec. 3.4)* serves as the foundation for all subsequent analyses. It consists of 1000 samples per language, each containing: `prompt` (three historical tweet-reply pairs plus fourth tweet in chat completion format), `authentic reply` (ground truth from test users), and two generated columns `base model reply` and `ft model reply` produced by applying the generation procedure described above to the base and fine-tuned Qwen3 8B models respectively. As a scientific artifact, the dataset serves three potential usages: 1) improving the next reply prediction task given our proposed metrics, 2) developing additional metrics to analyze the LLM-human alignment further, and 3) improving synthetic content detection classifiers.

3.2. Evaluation: Levels of Alignment

Quantitative Features We implement the complete NeLa feature suite (Horne et al., 2018) through a modular extraction pipeline spanning five linguistic dimensions: complexity, style, bias, affect, and moral reasoning patterns. The system extracts linguistic profiles including type-token ratios, average sentence length, lexical diversity measures, readability scores (Flesch-Kincaid (Kincaid et al., 1975), Gunning Fog (Gunning, 1968)), and character-level complexity metrics.

Morphosyntactic Extraction Using the spaCy processing pipeline (Montani et al., 2023), we extract comprehensive linguistic annotations including part-of-speech tag distributions following Universal Dependencies standards (De Marneffe et al., 2021), named entity recognition patterns across 18 standard categories (PERSON, ORG, GPE, DATE, etc.), dependency relation frequencies, and syntactic complexity measures. Our implementation computes frequency-normalized distributions for both POS categories and NER labels, incorporating lexical diversity metrics and average sentence length measurements.

Semantic Classification We employ the Tweet-Eval benchmark (Barbieri et al., 2020) through pre-trained transformer-based classifiers to evaluate content across multiple semantic dimensions. Our pipeline integrates three specialized models: `tweet-topic-21-multi` (Antypas et al., 2022) for topic classification, `twitter-RoBERTa-base-emotion` (Camacho-Collados et al., 2022) for emotion detection, and `twitter-RoBERTa-base-sentiment` (Barbieri et al., 2020) for sentiment analysis.

Cluster-based Similarity Utilizing the state-of-the-art instruction-following embedding model Qwen3 (Zhang et al., 2025), we compute semantic similarity distributions within and across content categories. Through cluster analysis using PCA dimensionality reduction (Pearson, 1901) and Affinity Propagation (Frey and Dueck, 2007), we analyze the proportion of clusters per content category.

Feature-Vector Distance Computation To quantify linguistic alignment between human and synthetic content, we implement a distance-based similarity metric. For each corpus C and linguistic feature set F, we compute normalized feature vectors through the following procedure:

1. Extract mean feature scores for each corpus-feature combination:

$$\bar{f}_C^i = \frac{1}{|C|} \sum_{d \in C} F_i(d)$$

   where d represents sample in corpus C and F_i denotes the i-th feature in F.

2. Construct corpus vectors: $\mathbf{v}_C = [\bar{f}_C^1, \ldots, \bar{f}_C^{|F|}]$

3. Compute pairwise cosine similarity between two corpus vectors defined as $s(\mathbf{v}_{C_1}, \mathbf{v}_{C_2})$.

3.3. Validation: Detecting Synthetics

As a downstream validation task, we implement a comparison of detection approaches spanning the spectrum from traditional sparse representations to modern dense embeddings. We concatenate the above-described features with the following text embeddings to investigate if these features improve the identification of synthetically generated examples.

Encoding Approaches

TF-IDF Sparse term frequency-inverse document frequency vectorization (Ramos et al., 2003) with uni-gram features and lowercase normalization for, baseline, traditional bag-of-words representation.

FastText Dense 300-dimensional word vectors (Joulin et al., 2017) using spaCy's `en_core_web_lg` and `de_core_news_lg` model, aggregated through mean pooling for efficient semantic representation.

Qwen3 Embedding : State-of-the-art instruction-following embeddings using the Qwen/Qwen3-Embedding-8B model (Zhang et al., 2025) with a specialized authorship detection prompt: *"Instruct: Find tweets with similar authorship patterns (human vs. AI-generated) based on writing style, vocabulary choice, and content structure"*. We choose Qwen3 as it shows benchmark-leading performance in text classification tasks relative to its parameter count (Pan et al., 2025; Heseltine, 2025).

Feature Combination Strategy To investigate the complementary nature of different representation types, we systematically evaluate all possible combinations of encoding approaches and extracted features, creating hybrid representations that capture multiple linguistic perspectives simultaneously.

Classification Model We utilize XGBoost (eXtreme Gradient Boosting) (Chen and Guestrin, 2016) as our classification algorithm. We select XGBoost for its promising performance on heterogeneous feature combinations, robust handling of different feature scales, and interpretability through feature importance analysis.

3.4. Reproducibility and Code Availability

All experimental procedures, statistical analyses, and model training protocols are implemented using open-source tools including scikit-learn (Buitinck et al., 2013), spaCy (Montani et al., 2023), Transformers Reinforcement Learning (TRL) (von Werra et al., 2020) and Sentence Transformers (Reimers and Gurevych, 2019). Complete code implementations, experimental configurations, and final datasets are available through the following repository: https://github.com/cl-trier/TWON-NRP-X-Dataset

4. Results

Our results reveal systematic linguistic differences between human and synthetic content across all analytical dimensions, with fine-tuned models consistently showing superior alignment to human content compared to prompt-based approaches. These findings address our three research questions through complementary lenses: similarity analysis (RQ1 and RQ2) and classification performance (RQ3).

4.1. Quantitative Linguistics Analysis

Table 1 presents the calculated similarity scores between corpus subsets across all feature extraction approaches. The results show a consistent hierarchy of alignment, with fine-tuned models (F) showing highest similarity to human original content (O), followed by moderate alignment between original and prompt-based content (P), while prompt-based and fine-tuned models exhibit the lowest mutual similarity.

Feat./Lang.	$s(O,P)$	$s(O,F)$	$s(P,F)$
Quantitative Features (NeLa)			
German	0.7908	0.8048	0.6995
English	0.8408	0.8957	0.8410
Morphosyntactic Extraction (SpaCy)			
German	0.9498	0.9748	0.9437
English	0.9423	0.9816	0.9357
Semantic Classification (TweetEval)			
German	0.9695	0.9874	0.9832
English	0.9745	0.9819	0.9786
Cluster-based Similarity			
German	0.8016	0.9713	0.7435
English	0.8977	0.9620	0.8323

Table 1: Comparison of the calculated similarity s between the corpora subsets human original (O), synthetic only prompted (P) and synthetic fine-tuned (F) across German and English on all features described in section 3.2. A higher value indicates a more aligned model behavior.

Quantitative Features The NeLa features reveal substantial differences in linguistic complexity and style patterns. For German, similarity between original and fine-tuned content reaches 0.8048, significantly higher than the 0.6995 similarity between prompt-based & fine-tuned approaches. English demonstrates even stronger alignment patterns, with original & fine-tuned similarity achieving 0.8957, while original-prompt similarity reaches 0.8408.

Morphosyntactic Extraction The Morphosyntactic analysis reveals the highest overall similarity scores across all approaches. German shows a high alignment between original & fine-tuned content (0.9748), with prompt-based models achieving 0.9498 similarity to original content. However, detailed examination reveals that prompt-based models exhibit distinctive usage patterns, particularly in coordinating (CCONJ) and subordinating conjunctions (SCONJ) (Figure 1).

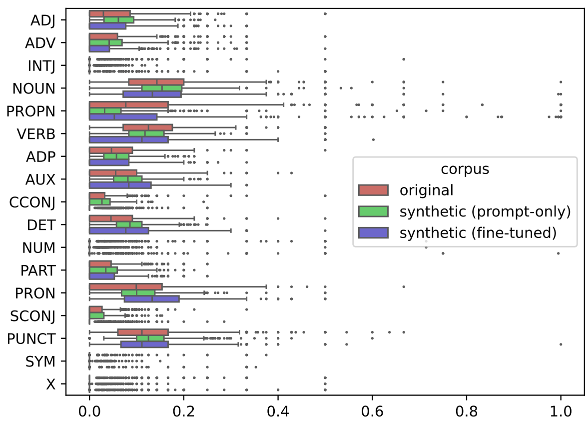


Figure 1: Locality, spread and skewness (x-axes) of each POS category (y-axes) for the English corpus split into subsets.

Semantic Classification Semantic classification shows the most consistent alignment across all model types, with similarity scores exceeding 0.97 in all comparisons. German achieves the highest alignment between prompt and fine-tuned models (0.9832), while English shows marginally lower but still substantial similarity (0.9786). Despite these high similarity scores, qualitative analysis reveals that prompt-based models generate more topically diverse content and exhibit significantly higher proportions of positive emotion classifications compared to human content (Figure 2).

Cluster-based Similarity Embedding-based cluster analysis reveals the most pronounced differences between generation approaches. Fine-tuned models achieve high alignment with original content (German: 0.9713, English: 0.9620), while

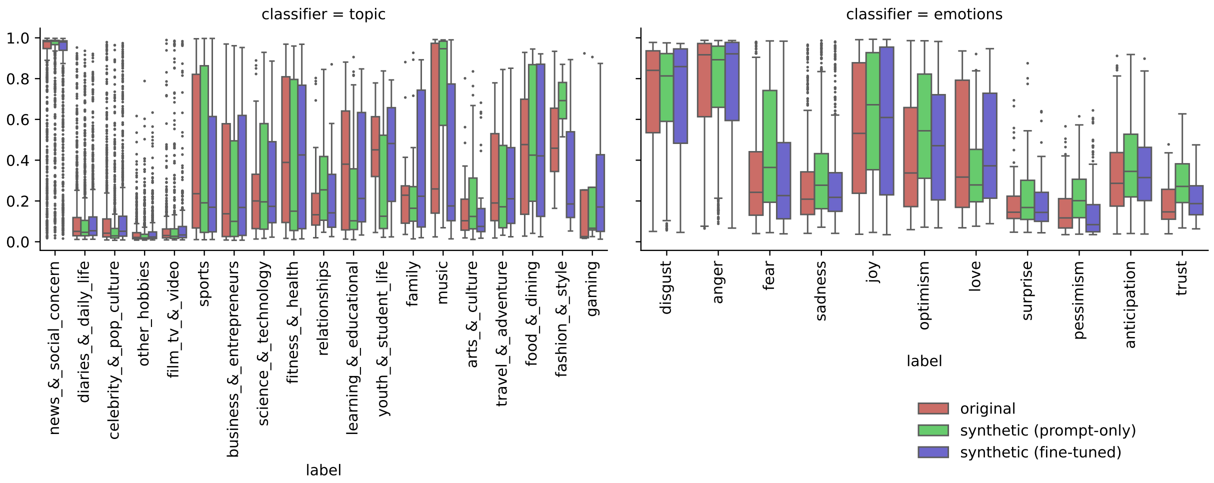


Figure 2: Locality, spread and skewness (y-axes) of the TweetEval topic and emotion classifier (x-axes) for the English corpus split into subsets.

prompt-based models show notably lower similarity to both original content and fine-tuned variants. The substantial gap between original & prompt similarity (German: 0.8016, English: 0.8977) and original & fine-tuned similarity demonstrates that semantic distributional properties are particularly sensitive.

4.2. Validation Task

Table 2 presents the classification results for distinguishing between human original (O), synthetic fine-tuned (F), and synthetic prompted (P) content across various feature combinations and encoding approaches. The results consistently demonstrate that prompt-based synthetic content (P) achieves the highest detection accuracy, while fine-tuned content (F) proves most challenging to distinguish from human original content (O).

Feature Combination Performance The most effective approach combines tf-idf, fastText embeddings, and one or more extracted features (TweetEval, SpaCy, NeLa), achieving macro $F1$ scores of 0.7301 for German and 0.6972 for English. Notably, prompt-based content consistently achieves the highest individual $F1$ scores across all feature combinations (German: $0.8297 - 0.8510$, English: $0.6725 - 0.8163$).

Encoding Approach Analysis Modern embedding approaches show competitive but not superior performance compared to traditional methods. The Qwen embedding model alone achieves moderate performance (German $F1$: 0.6365, English $F1$: 0.6549), while fastText embeddings demonstrate strong baseline performance (German $F1$: 0.7007, English $F1$: 0.6240). Surprisingly, simple tf-idf representations prove remarkably effective, particularly for prompt-based content detection (German: 0.8333, English: 0.7800).

Feat./Lang.	$F1(O)$	$F1(F)$	$F1(P)$	avg
tf–idf + fastText + {TweetEval, SpaCy, NeLa}				
German	0.6666	0.6938	0.8297	0.7301
English	0.6534	0.6382	0.8000	0.6972
tf–idf + fastText + Qwen				
German	0.6336	0.6476	0.8510	0.7107
English	0.5531	0.6136	0.6725	0.6240
Qwen				
German	0.5800	0.5849	0.7446	0.6365
English	0.5544	0.5940	0.8163	0.6549
fastText				
German	0.6734	0.6407	0.7878	0.7007
English	0.5858	0.6136	0.6725	0.6240
tf–idf				
German	0.6400	0.5961	0.8333	0.6898
English	0.5200	0.5000	0.7800	0.6000

Table 2: Results of the detection task for the individual $F1$ scores per class, human original (O), synthetic only prompted (P) and synthetic fine-tuned (F), and the macro average across German and English on a selected range of feature combinations with XGBoost as classifier.

5. Discussion

5.1. Implications for Computational Social Science

Our findings reveal fundamental challenges for the ecological validity of LLM-based simulations in social research contexts. While current generation techniques can produce individually plausible social media posts, systematic analysis reveals consistent patterns that distinguish synthetic from authentic content across multiple linguistic dimensions. The detection accuracies achieved in our validation task, particularly for prompt-based content, indicate persistent linguistic signatures that compromise the authenticity of LLM-generated social media discourse.

This detectability gap has important implications for applications in computational social science, where researchers increasingly rely on LLMs as human proxies for behavioral studies. The systematic differences we observe in quantitative features, morphosyntactic patterns, and semantic distributions suggest that naive deployment of LLMs for social simulation may introduce systematic biases that compromise research validity. The observation that even fine-tuned models, while substantially improved, still exhibit detectable patterns in classification tasks suggests that the challenge extends beyond simple technical optimization to fundamental questions about the nature of human-like generation.

These findings align with broader concerns about the anthropomorphism of AI systems (Salles et al., 2020) and highlight the necessity for validation protocols when deploying LLMs in social research contexts (Møller and Aiello, 2024). The consistent performance hierarchy observed across all feature extraction approaches, with fine-tuned models showing highest alignment to human content, followed by prompt-based models, while the two synthetic approaches exhibit lowest mutual similarity, provides empirical evidence for the complexity of achieving authentic human simulation.

5.2. Linguistic Authenticity and Model Limitations

The systematic differences we observe across quantitative linguistics, morphosyntactic patterns, and semantic distributions point to inherent limitations in current language modeling approaches. Our analysis reveals that prompt-based models exhibit distinctive linguistic signatures, including more complex sentence structures (evidenced by coordinating and subordinating conjunction usage patterns), more topically diverse content, and significantly higher proportions of positive emotion classifications compared to human content.

Particularly concerning is the cluster-based similarity analysis, which shows the most pronounced differences between generation approaches. The substantial gaps between original & prompt similarity and original & fine-tuned similarity demonstrate that semantic distributional properties are particularly sensitive to generation method. These findings suggest that LLMs may be systematically biased toward producing "ideal" rather than authentic communication, potentially missing the natural variation, errors, and stylistic inconsistencies that characterize genuine human social media discourse (Thapa et al., 2025).

The cross-linguistic consistency of these patterns across English and German corpora strengthens the generalizability of our findings, indicating that the observed limitations are not language-specific artifacts but reflect fundamental characteristics of current language modeling approaches.

5.3. Methodological Considerations for LLM Deployment

The superior performance of fine-tuned models compared to prompt-based approaches across all similarity metrics provides strong evidence for the importance of domain adaptation in social media generation tasks. Fine-tuned models consistently achieve higher similarity scores with human content compared to prompt-based models. However, the persistence of detectable patterns even after fine-tuning, suggests that current adaptation techniques may be insufficient for achieving true linguistic authenticity (Münker et al., 2026).

The effectiveness of different encoding approaches in our validation task reveals important insights about the nature of synthetic content detection. The surprising performance of traditional tf-idf representations, particularly for prompt-based content detection, suggests that surface-level lexical patterns remain highly discriminative despite the sophistication of modern language models. The superior performance of hybrid approaches combining tf-idf, fastText embeddings, and extracted linguistic features demonstrates that multiple representational perspectives are necessary to capture the full spectrum of linguistic differences between human and synthetic content.

6. Conclusion

Our paper has examined a fundamental question about the viability of LLMs as human simulacra in computational social science: can current generation techniques produce social media content that reliably replicates authentic human linguistic behavior? Through systematic analysis of a novel history-conditioned dataset spanning English and German X content, we provide evidence-based answers to three interconnected research questions.

6.1. Research Questions

RQ1: Linguistic Pattern Detection Our results demonstrate that LLM-generated social media posts exhibit systematic and detectable linguistic patterns across quantitative, morphological, and semantic dimensions. The similarity analysis reveals that, while individual synthetic posts may appear plausible, aggregate patterns consistently deviate from human norms. Most notably, prompt-based models show distinctive signatures in morphosyntactic complexity, with systematic differences in conjunction usage patterns indicating artificially complex sentence structures compared to human originals. Semantic analysis reveals systematic biases toward positive emotion classifications and increased topical diversity compared to authentic human content.

RQ2: Fine-tuning versus Prompt-based Approaches Fine-tuned models consistently outperform prompt-based approaches across all similarity metrics, achieving substantially higher alignment with human content. However, even fine-tuned models remain distinguishable from human content in classification tasks, particularly through cluster-based similarity analysis where the most pronounced differences emerge. This finding confirms that training models with human data for concrete, well-defined tasks consistently outperforms general prompt-based usage approaches, aligning with findings from concurrent work demonstrating the limitations of generic prompting strategies (Münker et al., 2026).

RQ3: Machine Learning Detection Capability Our validation task demonstrates reliable classification performance across multiple encoding approaches and feature combinations. The highest performing hybrid approach (tf-idf + fastText + extracted features) achieves macro $F1$ scores of 0.7301 (German) and 0.6972 (English), with particularly strong detection rates for prompt-based content ($F1 > 0.8$ across multiple configurations). Surprisingly, traditional tf-idf representations prove remarkably effective, suggesting that surface-level lexical patterns remain highly discriminative despite advances in generation sophistication.

6.2. Recommendations for Responsible LLM Deployment

Based on our findings, we propose specific guidelines for the responsible deployment of LLMs in social applications:

Mandatory Validation Protocols Researchers employing LLMs for social simulation must implement comprehensive validation protocols that assess linguistic authenticity across multiple dimensions rather than relying on surface-level plausibility assessments. Our multi-dimensional evaluation framework provides a template for such validation, combining quantitative linguistics analysis, morphosyntactic profiling, semantic classification, and distributional similarity measures.

Domain-Specific Fine-tuning Requirements Our results confirm that fine-tuned models consistently outperform prompt-based approaches for social media generation tasks across all similarity metrics. However, fine-tuning alone proves insufficient to achieve complete linguistic authenticity, as evidenced by persistent detectability in classification tasks. This suggests that domain adaptation should be considered a minimum requirement rather than a sufficient solution.

Multi-dimensional Evaluation Standards The complementary nature of different linguistic analysis approaches in our study demonstrates that single-metric evaluation is insufficient for assessing generation quality. Researchers should adopt multi-layered evaluation frameworks that capture quantitative features, morphosyntactic patterns, semantic distributions, and embedding-based similarity measures simultaneously.

6.3. Future Directions

Our findings open several relevant directions for future research. First, investigating the temporal stability of linguistic signatures as generation techniques continue to evolve will be essential to understand the longevity of current detection methods and to develop robust evaluation frameworks. Second, examining domain transfer across different social media platforms beyond X will help establish the generalizability of these linguistic signature patterns across diverse communication contexts with varying discourse norms and constraints.

Third, exploring adversarial training approaches specifically designed to reduce detectability while maintaining content quality and authenticity represents a promising direction for improving generation fidelity. Such approaches could inform the development of more sophisticated LLMs that better capture the natural variation, errors, and stylistic inconsistencies characteristic of genuine human discourse on social media. Finally, developing more nuanced evaluation metrics that capture subtle aspects of human communication patterns beyond current similarity measures could provide deeper insights into the fundamental challenges of achieving truly human-like text generation.

The cross-linguistic consistency of our findings across English and German corpora suggests that these challenges transcend language-specific artifacts and reflect fundamental limitations in current language modeling approaches.

Limitations

Our analysis focuses on X data collected during the first half of 2023, which may not generalize to other social media platforms or communication contexts with different discourse norms and constraints. The temporal dimension of our dataset may not capture evolving generation capabilities as LLM technology continues to advance rapidly. Additionally, our current framework focuses on English and German languages, and expanding the analysis to include morphologically richer languages, tonal languages, and non-European linguistic families would strengthen the cross-linguistic validity of these findings. Beyond these core limitations, we acknowledge several methodological considerations.

Analysis Framework Our linguistic analysis framework, while comprehensive across quantitative, morphosyntactic, and semantic dimensions, does not capture complex discourse quality metrics such as argumentation coherence, irony detection, or cultural nuance recognition. The focus on individual post generation rather than multi-turn conversational dynamics limits our understanding of how synthetic content would perform in sustained social interactions and community discussions.

Validation Experiments Our detection validation experiments, while demonstrating reliable classification performance, are limited to the specific LLM architectures and fine-tuning approaches employed in this study. The rapid evolution of language models means that newer generation techniques may exhibit different linguistic signatures than those captured in our analysis. Additionally, our evaluation framework relies primarily on automated feature extraction and classification metrics, which may not capture subtle qualitative differences that human evaluators would detect.

Single Model Architecture Our results are based exclusively on Qwen3 8B, which represents only a single model architecture and size configuration. The observed linguistic patterns and detection accuracies may vary considerably across different model families, model sizes, quantization approaches within the same base model, and model versions. This architectural specificity limits the generalizability of our findings to the broader landscape of available LLMs.

Reply Prediction Task The history-conditioned reply prediction task relies on only three prior tweet-reply pairs as context, which may provide sparse predictive signal for capturing individual user behavior patterns and writing styles. This limited historical context may not fully represent the complexity and variation present in users' broader communication patterns, potentially affecting both the fine-tuning quality and the authenticity of generated content.

German vs. English The German and English datasets differ substantially in their collection contexts, temporal distribution, and underlying discourse characteristics. These systematic differences make direct cross-linguistic performance comparisons not recommended, as observed variations may reflect dataset-specific properties rather than fundamental linguistic or modeling differences. Each language corpus should be interpreted within its own context rather than as directly comparable benchmarks.

Ethical Considerations

As is typical for AI methods, the modeling approach presented in this paper is a dual-use technology. While behavior-based user modeling and synthetic content generation are primarily intended for computational social science research and platform safety applications, the findings can also be used to develop more sophisticated manipulation techniques or improve the convincingness of synthetic social media content for malicious purposes.

Privacy and Consent Considerations A significant ethical concern in our study involves the use of real user data from X to train models that replicate individual behavior patterns. While our dataset consists of publicly available posts from political discourse and replies from regular users, the individuals whose data we used did not provide explicit informed consent for their communication patterns to be learned and replicated by generative models. This raises important questions about digital privacy rights, even when dealing with publicly posted content.

Potential for Misuse The detection methodologies developed in this work, while intended to improve synthetic content identification, could potentially be used adversarially to develop more sophisticated generation techniques that evade detection. The detailed analysis of linguistic signatures across quantitative, morphosyntactic, and semantic dimensions provides a road-map for improving synthetic content quality, which could enhance both legitimate applications and malicious use cases.

Broader Implications The development of increasingly sophisticated user modeling and synthetic content generation capabilities raises broader questions about the boundaries of acceptable research practices in computational social science.

As these technologies advance, the research community must carefully balance the scientific value of realistic behavioral simulation against the privacy rights and dignity of individuals whose data enables such research, while considering the potential societal impacts of increasingly convincing synthetic social media content.

Acknowledgments

We thank Simon Werner and Christoph Hau for our constructive discussions. This work is supported by TWON (project number 101095095), funded by the European Union under the Horizon framework (HORIZON-CL2-2022-DEMOCRACY-01-07); and EMO-SCI (project number 569156112) funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Bibliographical References

Harika Abburi, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2024. Toward robust generative ai text detection: Generalizable neural model. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1651–1656. IEEE.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Jan Pablo Burgard, Jan-Philipp Kolb, Hariolf Merkle, and Ralf Münnich. 2017. Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 11(3):233–244.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Alicia Tsui Ying Chong, Hui Na Chua, Muhammed Basheer Jasser, and Richard TK Wong. 2023. Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features. In *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pages 205–210. IEEE.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Robert Gunning. 1968. *The technique of clear writing*. McGraw-Hill Book Company, New York.

Kadhim Hayawi, Susmita Saha, Mohammad Mehedy Masud, Sujith Samuel Mathew, and Mohammed Kaosar. 2023. Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*, 35(12):8903–8918.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. Association for Computational Linguistics.

Michael Heseltine. 2025. Comparing large language models for text classification: Model selection across tasks, texts, and languages.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, pages 235–238.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report, Naval Technical Training Command, Millington*.

Maik Larooij and Petter Törnberg. 2025. Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. *arXiv preprint arXiv:2504.03274*.

Haocheng Lin. 2024. Designing domain-specific large language models: The critical role of fine-tuning in public opinion simulation. *arXiv preprint arXiv:2409.19308*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Anders Giovanni Møller and Luca Maria Aiello. 2024. Prompt refinement or fine-tuning? best practices for using llms in computational social science tasks. *arXiv preprint arXiv:2408.01346*.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. spacy: Industrial-strength nlp. *Zenodo*.

Simon Münker. 2025a. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. In *Proceedings of 0 th Moral and Legal AI Alignment Symposium*, page 61.

Simon Münker. 2025b. Political bias in llms: Unaligned moral values in agent-centric simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

Simon Münker, Nils Schwager, and Achim Rettinger. 2026. Don't trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1141–1151. Association for Computational Linguistics.

Yuan Pan, Guocong Feng, Kaitian Huang, and Chunmei Zhang. 2025. Qwen3-powered log classification for improved soc decision-making. In *2025 8th International Conference on Computer Information Science and Application Technology (CISAT)*, pages 651–655. IEEE.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pages 1–22.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Jaime Pérez, Mario Castro, and Gregorio López. 2023. Serious games and ai: Challenges and opportunities for computational social science. *IEEE Access*, 11:62051–62061.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

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.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.

Nils Schwager, Simon Münker, Alistair Plum, and Achim Rettinger. 2026. Towards simulating social media users with LLMs: Evaluating the operational validity of conditioned comment prediction. In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 208–221. Association for Computational Linguistics.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Quid est VERITAS?
A Modular Framework for Archival Document Analysis

Leonardo Bassanini[1], Ludovico Biancardi[1], Alfio Ferrara[2], Andrea Gamberini[3], Sergio Picascia[4*], Folco Vaglienti[3]

[1]Università degli Studi di Milano, University Library Service
Via Santa Sofia, 7/9 - 20122 Milano (Italy)
leonardo.bassanini@unimi.it, ludovico.biancardi@unimi.it

[2]Università degli Studi di Milano, Department of Literary Studies, Philology and Linguistics
Via Festa del Perdono, 7 - 20122 Milano (Italy)
alfio.ferrara@unimi.it

[3]Università degli Studi di Milano, Department of Historical Studies
Via Festa del Perdono, 7 - 20122 Milano (Italy)
andrea.gamberini@unimi.it, folco.vaglienti@unimi.it

[4]Università degli Studi di Milano, Department of Computer Science
Via Celoria, 18 - 20133 Milano (Italy)
sergio.picascia@unimi.it

Abstract

The digitisation of historical documents has traditionally been conceived as a process limited to character-level transcription, producing flat text that lacks the structural and semantic information necessary for substantive computational analysis. We present VERITAS (Vision-Enhanced Reading, Interpretation, and Transcription of Archival Sources), a modular, model-agnostic framework that reconceptualises digitisation as an integrated workflow encompassing transcription, layout analysis, and semantic enrichment. The pipeline is organised into four stages—Preprocessing, Extraction, Refinement, and Enrichment—and employs a schema-driven architecture that allows researchers to declaratively specify their extraction objectives. We evaluate VERITAS on the critical edition of Bernardino Corio's *Storia di Milano*, a Renaissance chronicle of over 1,600 pages. Results demonstrate that the pipeline achieves a 67.6% relative reduction in word error rate compared to a commercial OCR baseline, with a threefold reduction in end-to-end processing time when accounting for manual correction. We further illustrate the downstream utility of the pipeline's output by querying the transcribed corpus through a retrieval-augmented generation system, demonstrating its capacity to support historical inquiry.

Keywords: vision language models, document layout analysis, historical document digitisation, digital humanities

1. Introduction

The digitisation of historical documents has long been conceived as a process whose primary objective is the faithful transcription of textual content into machine-readable form. Under this paradigm, the output is flat text: a character sequence suitable for keyword search but lacking the structural, semantic, and relational information necessary for substantive scholarly inquiry. This limitation stems from an entrenched architectural assumption that semantic enrichment constitutes a separate, downstream task. In practice, this separation means that the majority of digitised collections never receive such enrichment, as the resources required to revisit already-processed materials and execute enrichment pipelines are rarely available. The consequence is a widening gap between the volume of digitised material and the volume that is genuinely searchable, interoperable, and amenable to computational analysis.

We argue that overcoming this gap requires reconceptualising digitisation itself: rather than treating semantic enrichment as optional post-processing, it should be an integral component of the digitisation workflow, producing structured, semantically annotated digital objects. The feasibility of this shift is enabled by recent advances in foundation models, particularly vision-language models (VLMs) and large language models (LLMs). Unlike traditional OCR systems, which operate at the character level with no awareness of document semantics, VLMs can jointly process visual and textual information, performing layout analysis, transcription, and content interpretation within a unified inference pass. LLMs, in turn, can be leveraged for downstream semantic tasks previously requiring dedicated, task-specific systems. Together, these technologies make it viable to construct pipelines traversing the full arc from raw image to semanti-

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 57–66
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

cally enriched digital object in a single automated workflow.

In this paper, we present VERITAS (Vision-Enhanced Reading, Interpretation, and Transcription of Archival Sources), a modular framework that operationalises this integrated approach. VERITAS is organised into four sequential stages, each constituting a self-contained module with well-defined inputs and outputs. The framework is model-agnostic: individual components can be substituted as newer models become available without architectural modifications. A schema-driven extraction process allows researchers to configure the target data structure according to their analytical objectives. We evaluate VERITAS on Bernardino Corio's *Storia di Milano*, a Renaissance chronicle of over 1,600 pages, demonstrating substantial improvements over a commercial OCR baseline in both transcription accuracy and processing efficiency. Furthermore, we illustrate downstream utility by submitting the complete transcription to a retrieval-augmented generation (RAG) system and posing historically motivated research questions.

2. Related Work

The automated processing of historical documents has attracted growing attention at the intersection of computer science and the digital humanities. We review the relevant literature across two areas: transcription technologies and integrated digitisation pipelines.

2.1. Transcription of Historical Documents

The transcription of historical documents has evolved through several technological paradigms. Commercial OCR engines such as Tesseract (Smith, 2007) have long served as default tools; however, their reliance on character-level pattern matching renders them ill-suited to the degraded image quality, irregular layouts, and archaic typographic conventions of historical materials. Deep learning substantially improved capabilities for such documents. Transkribus (Kahle et al., 2017; Muehlberger et al., 2019), developed within the EU-funded READ project, became one of the most widely adopted platforms for automatic text recognition in the humanities. In parallel, eScriptorium (Kiessling et al., 2019) extended recognition to non-Latin and bidirectional writing systems. While both platforms significantly lowered the barrier of entry, they remain primarily interactive tools requiring manual ground-truth preparation, model training, and post-hoc correction.

Generative AI has opened new avenues for historical document transcription. One line of research explores LLMs as post-OCR correctors: Thomas et al. (2024) demonstrated that instruction-tuned Llama 2 models can achieve a 54.5% reduction in character error rate on 19th-century British newspapers, though subsequent studies reported mixed results in multilingual contexts (Kanerva et al., 2025; Boros et al., 2024). A more recent line of research has shifted focus to direct transcription through vision-language models (VLMs), which jointly process visual and textual information end-to-end without a separate OCR stage. Several evaluations demonstrate that VLMs can match or surpass dedicated OCR systems: Humphries et al. (2024) reported character error rates of 5–7% on 18th–19th century English manuscripts, Kim et al. (2025) found general-purpose VLMs outperforming traditional OCR on historical tabular documents, and Levchenko (2025) benchmarked 12 multimodal LLMs on 18th-century Russian texts, highlighting both promise and pitfalls such as *over-historicisation*. On the specialised side, CHURRO (Semnani et al., 2025), a 3B-parameter VLM fine-tuned on 155 historical corpora spanning 46 language clusters, demonstrated that targeted training yields accuracy superior to commercial alternatives at a fraction of the cost. VERITAS builds on this latter paradigm, employing VLMs as its primary extraction mechanism while retaining flexibility for LLM-based refinement.

2.2. Integrated Digitisation Pipelines

While significant progress has been made on individual components, fewer efforts have addressed their integration into end-to-end pipelines. The OCR-D project (Neudecker et al., 2019) developed a modular framework for OCR processing of historical printed documents in German libraries, emphasising interoperability through standardised formats, though its scope does not extend to semantic enrichment. The DAHN project (Chiffoleau, 2024) proposed a TEI-centred pipeline comprising six stages, leveraging eScriptorium for HTR and TEI Publisher for dissemination, representing an important step but remaining oriented towards digital scholarly editions rather than flexible data extraction. At larger scales, SocFace (Boillet et al., 2024) demonstrated end-to-end pipelines for French censuses, and Constum et al. (2024) proposed an approach for handwritten Parisian marriage records. These efforts, however, tend to be tightly coupled to specific document genres.

VERITAS extends this body of work in several respects. Unlike interactive platforms such as Transkribus and eScriptorium, it operates as a fully automated, configurable pipeline. Unlike OCR-focused frameworks such as OCR-D, it extends beyond transcription to encompass semantic enrichment, entity linking, and structured data indexing. Its model-agnostic architecture provides flexibility absent from existing designs. Finally, by demonstrating downstream utility for LLM-assisted scholarly inquiry (Section 5), we address a gap where evaluation typically stops at transcription accuracy without assessing usability for substantive research.

3. Methodology

VERITAS (Vision-Enhanced Reading, Interpretation, and Transcription of Archival Sources) is designed to transform raw historical documents into structured, semantically enriched, machine-readable data. By leveraging VLMs and LLMs, the pipeline extends digitisation beyond character-level transcription towards comprehensive document understanding. The architecture is model-agnostic, allowing practitioners to substitute models optimised for particular languages or document types without architectural modifications. Combined with schema-driven extraction, this ensures broad applicability across diverse archival collections.

The pipeline comprises four sequential stages—Preprocessing, Extraction, Refinement, and Enrichment—each a self-contained module with well-defined inputs and standardised outputs (Figure 1). Individual components can be activated, bypassed, or substituted according to project requirements. Throughout this section, we illustrate operations using pages from the case study in Section 4: the critical edition of Bernardino Corio's *Storia di Milano* (1978), a Renaissance chronicle of over 1,600 pages.

3.1. Preprocessing

The Preprocessing stage transforms heterogeneous raw inputs into a standardised format suitable for automated analysis through three operations.

Data Conversion. The pipeline accepts input formats commonly encountered in archival research, e.g. digitised PDFs, individual page scans (TIFF, JPEG, PNG), and photographs captured with handheld devices, and converts them into a uniform representation, ensuring that all subsequent pipeline stages operate on a homogeneous, high-quality image format. Each input is rendered as a high-resolution raster image normalised to a consistent colour space and resolution.

Image Enhancement. Optional computer vision techniques may be applied to the normalised images: deskewing (correcting rotational misalignment), denoising (removing noise, stains, or scanning artifacts), binarisation (converting to black-and-white to increase text-background contrast), and page detection (isolating the document area from extraneous background elements). The specific enhancement operations applied are configurable based on the quality and characteristics of the source materials. Figure 2 shows the result of these operations conducted on an image from the proposed case study.

Schema Definition. This operation constitutes the critical interface between scholarly intent and machine-processable output: analytical requirements are formalised into a structured specification determining what information the pipeline extracts and in what form. The schema, typically expressed in JSON Schema, defines the semantic fields to populate for each document element, and determines the tools employed in subsequent stages. Its design is a collaborative exercise between domain experts, who articulate driving research questions, and technical personnel, who translate these into formal specifications. Below is an illustrative example.

```
{
    "type": "object",
    "properties": {
        "bbox": {
            "type": "array",
            "items": { "type": "integer" }
    },
        "category": {
            "type": "string",
            "enum": ["title", "text", "header",
                    "footnote", "figure", "table"]
        },
        "text": { "type": "string" },
        "speaker": { "type": "string" },
        "date": { "type": "string" },
        "place": { "type": "string" },
        "entities": {
            "type": "array",
            "items": {
                "type": "object",
                "properties": {
                    "mention": { "type": "string" },
                    "type": {
                        "type": "string",
                        "enum": ["person", "
                            institution", "place"]
                    }}}}},
    "required": ["bbox", "category", "text"]
}
```

3.2. Extraction

The Extraction stage constitutes the core of the pipeline, performing the transformation of visual information into machine-readable structured data. This stage is designed with high modularity, offering three distinct processing paths that can be selected based on project requirements, available computational infrastructure, and the complexity of the target extraction tasks.

- **Specialised Vision-Language Models**. Compact, resource-efficient VLMs optimised for document analysis, achieving high accuracy in layout detection, text localisation, reading order determination, and transcription. Their advantage lies in computational efficiency and transcription fidelity; however, they operate according to a fixed output schema and cannot

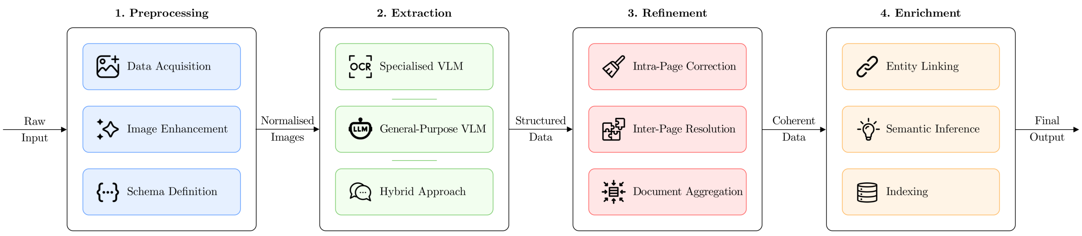


Figure 1: VERITAS architecture diagram illustrating the four stages and their inputs and outputs.

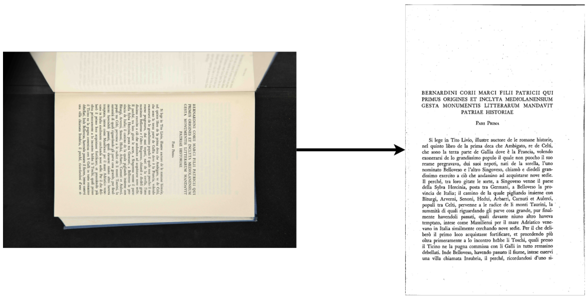


Figure 2: Preprocessing operations applied to a page from the critical edition of Corio's *Storia di Milano*. Left: the original colour scan, captured in landscape orientation with portions of the adjacent page visible. Right: the result after rotation correction, page detection, grayscale conversion, and adaptive thresholding.

accommodate arbitrary user-defined instructions.

- **General-Purpose Vision-Language Models**. Large-scale VLMs capable of processing both visual and textual information. While potentially exhibiting marginally lower performance on specialised document analysis benchmarks compared to the previous models, their strength resides in flexibility: users provide natural language instructions to guide extraction towards specific information needs beyond conventional transcription.

- **Multi-Step Hybrid Approach**. Combines both approaches through a two-phase process: a specialised VLM performs foundational layout detection and transcription, then a general-purpose LLM or VLM refines and enhances extraction according to user-specified instructions. This path suits complex documents requiring both accurate transcription and sophisticated semantic interpretation.

The selection among these paths is determined by the interplay of several factors, including the available computational resources, the quality and complexity of the source documents, and the depth of semantic analysis required. For projects prioritising transcription accuracy with minimal computational overhead, the first path offers an efficient solution. For projects requiring flexible, instruction-guided extraction or semantic inference, the second and third path are more appropriate. Figure 3 contrasts the predefined output of a specialised VLM against the more flexible output of a general-purpose VLM performing semantic inference.

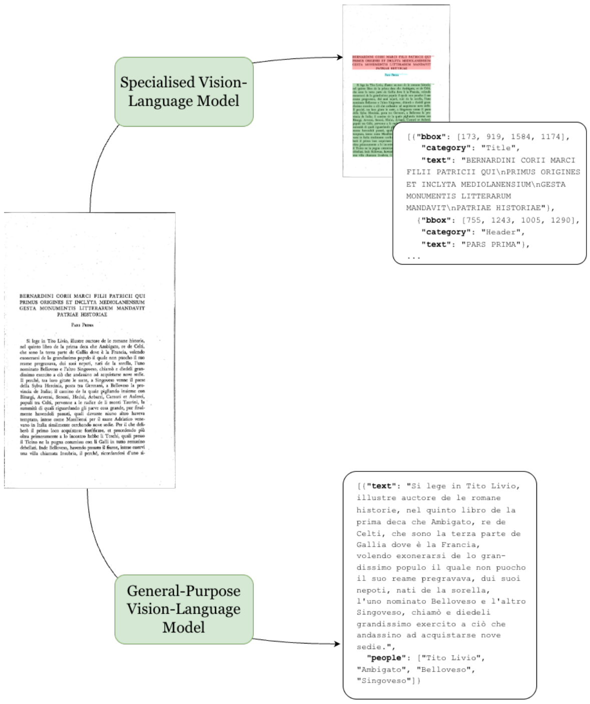


Figure 3: The difference in output between the Specialised Vision-Language Model and the General-Purpose Vision-Language Model processing paths.

3.3. Refinement

The Refinement stage transforms the collection of page-level extraction outputs into a single, coherent document representation through a series of cleaning, consolidation, and aggregation operations.

Intra-Page Correction. This operation addresses artifacts and inconsistencies within individual page outputs. Typical corrections include the reconstruction of hyphenated words split across line breaks, normalisation of typographic conventions (e.g., unifying quotation mark styles, standardising whitespace), and validation of the extracted data against the defined schema to ensure structural integrity.

Inter-Page Resolution. This operation resolves dependencies and continuities that span page boundaries. In historical documents, content units frequently extend across multiple pages, e.g a newspaper article beginning on one page and concluding on another, or a parliamentary speech spanning several folios. Inter-page resolution ensures that such fragmented content is correctly identified and linked, and that metadata is propagated appropriately across page breaks.

Document Aggregation. The refined page-level outputs are merged into a unified data structure representing the complete source document. This aggregation produces a single coherent object that serves as input to the subsequent Enrichment stage.

3.4. Enrichment

The Enrichment stage enhances the value of the structured data by connecting it to external knowledge sources, obtaining additional semantic insights, and preparing it for storage and downstream analysis.

Entity Linking. Named entities identified within the transcribed text—such as persons, organisations, and locations—are disambiguated and linked to canonical identifiers in external knowledge bases (e.g., Wikidata, VIAF, or domain-specific authority files). This process not only resolves ambiguities (e.g., distinguishing between individuals sharing the same name) but also situates the document within a broader knowledge graph, enabling cross-referencing and relational queries.

Semantic Inference. This operation leverages LLMs to perform advanced analytical tasks and infer information not explicitly present in the source text. Depending on the research objectives, such tasks may include topic classification, sentiment analysis, event extraction, temporal reasoning, named entity recognition for domain-specific entity types, or abstractive summarisation. The schema-driven architecture of the pipeline allows these inferred annotations to be systematically incorporated into the structured output. These operations can be also performed with general-purpose VLMs during the Extraction stage.

Indexing. The indexing operation formats the enriched data according to established standards appropriate to the target research community and loads it into suitable storage systems. For humanities scholarship, standard serialisation formats such as XML-TEI (Text Encoding Initiative) facilitate interoperability with existing digital humanities infrastructures. For quantitative analysis, tabular formats (e.g., CSV) or integration with database systems (e.g., relational databases, full-text search engines, vector or graph databases) enable efficient querying and statistical processing.

Figure 4 shows different examples of Enrichment operations: through semantic inference we can extract named entities, e.g. people and places; an encoder can compute the embedding of a textual element to index it for semantic search; we can match the entities within the text with external knowledge bases.

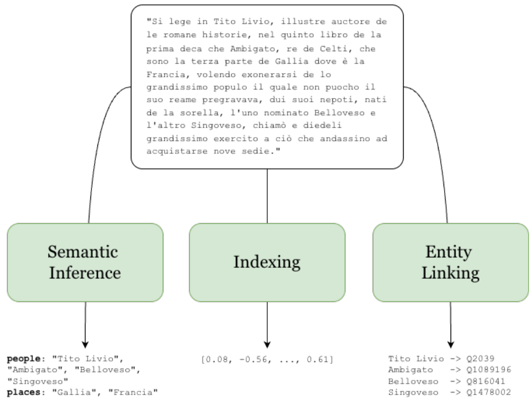


Figure 4: Examples of Enrichment operations.

4. Evaluation

To assess the effectiveness of the VERITAS pipeline, we conducted an empirical evaluation on a representative historical document collection. This evaluation focused on two primary dimensions: (i) transcription accuracy, measured through standard OCR quality metrics, and (ii) computational efficiency, assessed via processing time analysis. Additionally, we evaluated the pipeline's capacity for accurate layout analysis and element extraction.

4.1. Case Study

The evaluation was conducted on the *Storia di Milano* (*History of Milan*), a Renaissance chronicle authored by Bernardino Corio (1459–c. 1519). The digitisation of this work was undertaken as part of a broader interdisciplinary project at our university, bringing together historians, computational linguists, and computer scientists with the shared objective of making this primary source accessible for large-scale computational analysis. Corio's work constitutes one of the most significant sources for the history of the Duchy of Milan, spanning from antiquity to the late fifteenth century. For this study, we employed the critical edition curated by Anna Morisi Guerra, published by UTET (Turin) in 1978 as part of the *Classici della Storiografia* series (Corio, 1978). The edition comprises two volumes totalling 1,688 pages.

Although this critical edition employs modern typography, thereby avoiding the challenges posed by historical typefaces or manuscript hands, it nonetheless presents a rich variety of difficulties for automated processing that motivated its selection as an evaluation testbed. The text alternates between vernacular and modern Italian, requiring the extraction system to handle two distinct languages within a single document flow. The page layouts are heterogeneous, comprising dense prose, extensive footnotes, illustrative figures, tables, indices, and front matter, each demanding correct identification and classification by the layout analysis component.

4.2. Evaluation Methodology

Ground Truth Construction. A subset of 100 pages was selected from the corpus using stratified sampling to ensure adequate representation of the diverse page layouts present in the edition (e.g., pages with dense prose, pages with extensive footnotes, title pages, and pages containing illustrations or tables). Three domain experts, working from a shared set of transcription guidelines, manually transcribed these pages to establish a reference ground truth. Additionally, the structural elements identified by the vision-language model (text, headers, footnotes, etc.) were manually verified and corrected by the same annotators to provide ground truth for layout analysis evaluation.

Baseline. To contextualise the performance of the VERITAS pipeline, we compared it against ABBYY FineReader, a widely adopted commercial OCR solution that is also the standard digitisation tool currently employed by our university library. The software was executed with default configuration settings, which is appropriate given that the critical edition employs modern typefaces that do not require specialised historical document profiles.

Pipeline Configuration. For this evaluation, the VERITAS pipeline was configured as follows. In the preprocessing stage, several image normalisation operations were applied to address artifacts introduced during digitisation. The original scans were captured in landscape orientation with portions of adjacent pages partially visible in the frame; furthermore, the scans were acquired in colour. To prepare the images for extraction, we applied rotation correction to restore portrait orientation, employed an object detection model to identify and isolate the region of interest corresponding to the target page (Boillet et al., 2021), and converted the images to grayscale. Additionally, adaptive thresholding techniques were applied to enhance text-background contrast and improve content legibility. For the extraction stage, the pipeline was configured to use Path A (Specialised Vision-Language Models), employing a document-specialised VLM, specifically, `dots.ocr` (Li et al., 2025), followed by minimal post-processing operations consisting of end-of-line hyphenation correction and output format normalisation (removal of markdown artefacts).

Metrics. Transcription quality was assessed using two standard metrics: Word Error Rate (WER), defined as the minimum number of word-level insertions, deletions, and substitutions required to transform the predicted transcription into the ground truth, normalised by the total number of words in the reference; and Character Error Rate (CER), analogously defined at the character level, providing a finer-grained assessment of transcription fidelity. Both metrics were computed at the corpus level, i.e., over the concatenated text of all evaluated pages, to avoid potential bias introduced by page-length variability. To isolate substantive transcription errors from superficial formatting differences, we report results under two conditions: Raw (no normalisation) and Normalised (text converted to lowercase with punctuation removed). We also evaluated element extraction computing the F1 Score: an extracted element was considered a true positive if it matched a ground-truth element in both spatial location (bounding box overlap) and semantic label; false positives comprised spurious detections or incorrect label assignments; false negatives represented missed elements.

4.3. Results

Table 1 presents the corpus-level transcription error rates for both systems. The results demonstrate that the VLM-based approach employed in VERITAS consistently outperforms the commercial baseline across all metrics and normalisation conditions. Under normalised evaluation, VERITAS achieves a WER of 1.1% and a CER of 0.7%, representing relative improvements of 67.6% and 50.0%, respectively, compared to ABBYY FineReader. The performance gap is even more pronounced under raw evaluation conditions.

	ABBYY		**VERITAS**	
Text	WER	CER	WER	CER
Raw	0.142	0.031	**0.055**	**0.013**
Normalised	0.034	0.014	**0.011**	**0.007**

Table 1: Corpus-level transcription error rates computed over the concatenated text of all 100 evaluated pages. Lower values indicate better performance.

Regarding processing efficiency, while the VLM-based approach incurs substantially higher computational cost per page when processed individually (46 seconds versus 4 seconds for ABBYY), modern inference frameworks enable efficient concurrent processing through automatic request batching. In our experimental configuration, we employed vLLM (Kwon et al., 2023), a high-throughput serving framework that dynamically batches incoming requests and optimises GPU memory utilisation. Using one third of the memory allocation of a single Nvidia H100 NVL 94Gb GPU, vLLM's continuous batching mechanism reduced the effective per-page processing time to 0.89 seconds, a 4.5$\times$ improvement over the commercial solution. Beyond automated processing time, however, the practical impact of the pipeline must also account for the manual effort required to correct its output. The domain experts recorded the time needed to correct the ABBYY FineReader transcriptions, yielding a mean correction time of 2 minutes and 15 seconds per page. Extrapolated to the full 1,688-page corpus, this amounts to approximately 63 hours of manual post-correction labour. Since VERITAS reduces the normalised WER by 67.6% relative to ABBYY, a proportional reduction in correction effort can be reasonably assumed, bringing the estimated per-page correction time to approximately 44 seconds and the projected corpus-level total to roughly 20 hours, a saving of over 40 hours of expert labour. Combining automated processing and manual correction, the total estimated time for producing a verified transcription of the full edition decreases from approximately 65 hours with ABBYY to approximately 21 hours with VERITAS, a threefold reduction in end-to-end effort. These gains should, however, be interpreted in light of the computational profile of the selected extraction model. While concurrent inference substantially lowers effective per-page latency, VLM-based processing remains more demanding than conventional OCR in terms of GPU memory, serving infrastructure, and energy consumption.

Finally, the element extraction performance of the VLM-based approach confirms its reliability for layout analysis, with the model achieving an F1 Score of 0.966. These results indicate that the specialised VLM reliably identifies and correctly classifies the structural elements present in the document pages, a capability essential for the subsequent refinement and enrichment stages of the pipeline, as accurate element extraction enables proper content aggregation and semantic annotation.

5. Downstream Application

The preceding evaluation demonstrates that the VERITAS pipeline achieves high-fidelity transcription and reliable layout analysis. However, the ultimate value of a digitisation framework for the humanities resides not merely in the accuracy of its output, but in the degree to which that output can support substantive scholarly inquiry. To illustrate this potential, we conducted an exploratory study in which the complete transcribed text of the *Storia di Milano*, produced by the VERITAS pipeline, was ingested into a retrieval-augmented generation (RAG) system, and a set of historically motivated research questions was posed to the model.

5.1. Setup

On top of the indexing results of the Enrichment phase of VERITAS, we developed a dedicated RAG pipeline tailored to the structure of Corio's chronicle. The 1,688 digitised pages were processed by a custom ingestion module that annotates each page with temporal and structural metadata and groups them into chronologically coherent chunks aligned with the chronicle's year markers and chapters divisions. The resulting corpus comprises approximately 1,331 content chunks and 200 footnotes, embedded using BAAI/bge-m3[1] (Chen et al., 2024) and indexed in a ChromaDB collection with cosine distance.

At query time, an embedding-based router classifies each question as either *specific* (targeting events, persons, or dates) or *general* (spanning themes, style, or interpretive questions). Specific queries trigger year-filtered semantic search with footnote augmentation, while general queries employ Maximal Marginal Relevance (MMR) reranking to ensure temporal and thematic diversity across retrieved chunks. Retrieved passages are then fed to a generative model, GLM-4.7-Flash[2] (et al., 2025), with prompt templates that enforce grounding in the source material and instruct the model to match the query language.

A panel of historians formulated a set of research questions spanning diverse analytical dimensions: factual retrieval, interpretive analysis, thematic synthesis, prosopographic reconstruction, and cross-referencing of events and actors. These questions were designed to reflect the types of inquiry that scholars would naturally pursue when working with this primary source. From this set, we selected three representative questions for detailed discussion, each exemplifying a distinct mode of historical inquiry.

5.2. Representative Queries

Factual Entity Extraction. The question *"Who were the ducal secretaries and chancellors during*

[1] `https://huggingface.co/BAAI/bge-m3`

[2] `https://huggingface.co/zai-org/GLM-4.7-Flash`

the Sforza era?" requires the model to identify specific individuals and their institutional roles from mentions dispersed across the chronicle. The system produced a well-organised response grounded in the retrieved passages. It correctly identified Cicco Simonetta as *general secretary*, citing the chronicle's account of his formal appointment following the death of Galeazzo Maria Sforza in 1477, and described his sweeping administrative authority over both domestic and foreign affairs. The response further retrieved Giovanni Francesco Marliano, appointed as jurist and governor during Ludovico Sforza's departure from Italy in 1499, as well as peripheral figures such as Bernardino Curtio and his brother Iacopo, named prefect and captain of the Milanese fortresses during the Duke's illness in 1489. Notably, the system also reconstructed the institutional reorganisation into two senates described in the chronicle: one for civil affairs in the Corte dell'Arenga and another for state deliberations in the castle, where Simonetta and his associates exercised decisive influence. This response demonstrates the system's capacity to aggregate factual information scattered across hundreds of pages into a structured outcome, a task that would require considerable manual effort if conducted through traditional close reading alone.

Interpretive Stance Detection. The question *"Does the author reveal his political sympathies?"* demands a qualitatively different analytical operation: the model must synthesise evidence of authorial bias across the entire work and articulate an interpretive judgment. The system's response identified a multi-layered political stance that evolves over the course of the chronicle. It recognised Corio's explicit loyalty to the Sforza dynasty. The response detected an early apologetic posture, exemplified by the author's praise of Francesco Sforza's restoration of Milan's fortifications and his justification of ducal authority as a bulwark against popular disorder. However, the system also identified a progressive disenchantment in the later portions of the work: Corio's depiction of Giovanni Galeazzo Maria as a ruler corrupted by ministerial avarice, and his characterisation of *hybris* as the root cause of the Sforza downfall, reveal a capacity for self-critical judgment regarding his own patrons. This response illustrates the system's ability to move beyond literal extraction toward historical interpretation, identifying ideological tensions within the source.

Thematic Synthesis and Causal Reasoning. The question *"Which epidemics does the author record and what sanitary measures were taken to limit contagion?"* requires the model to identify a recurring thematic thread across several centuries of narrative and, for each instance, link the event to any associated policy response. The system identified multiple epidemic events, most prominently the *peste acerrima* of 1485, which Corio describes as having driven him into rural retreat and which directly motivated the composition of the chronicle, and the pestilence of 1450, reported to have caused approximately thirty thousand deaths and severely disrupted the Jubilee. For each event, the model provided contextual details drawn from the source text, including Corio's autobiographical account of fleeing to the countryside. Regarding sanitary measures, the response correctly noted that the chronicle reflects predominantly reactive and informal responses: rural withdrawal as a form of quarantine, obligations placed on rulers such as Emperor Henry VII to maintain urban infrastructure at their own expense (bridges, roads), and the role of civic assemblies (*Credentia*) in coordinating crisis management. This response demonstrates the system's capacity for multi-hop reasoning: extracting thematically related passages distributed across the chronicle and synthesising them into a coherent analytical account.

5.3. Discussion

The exploratory results presented above suggest that RAG-based LLM systems, when provided with high-quality transcriptions produced by the VERITAS pipeline, can serve as effective tools for preliminary historical analysis. However, while these results are encouraging, they should be interpreted with appropriate caution. The responses have not been subjected to systematic validation, and RAG-based systems remain susceptible to hallucination, particularly when queries require inference beyond what is explicitly stated in the source material. A rigorous evaluation of factual accuracy and interpretive validity, conducted in collaboration with domain historians, constitutes an essential direction for future work. Notwithstanding these limitations, this demonstration highlights the potential of integrating high-fidelity transcription pipelines with LLM-based querying systems to lower the barrier of entry for large-scale historical document analysis, enabling scholars to formulate and explore research hypotheses across extensive corpora more efficiently than traditional manual methods would allow.

6. Conclusion

We have presented VERITAS, a modular, model-agnostic framework that reconceptualises historical document digitisation as an integrated process encompassing transcription, structural analysis, and semantic enrichment within a unified pipeline. The framework's schema-driven architecture allows researchers to declaratively specify their extraction objectives, ensuring that the pipeline's outputs are tailored to the analytical needs of diverse scholarly communities.

Our evaluation on the critical edition of Corio's *Storia di Milano* demonstrates that a VLM-based extraction pipeline can substantially outperform a commercial OCR baseline, achieving a 67.6% relative reduction in word error rate under normalised conditions, while concurrent inference reduces effective per-page processing time by a factor of 4.5. When accounting for the manual correction effort that remains indispensable in any digitisation workflow, these improvements translate into an estimated threefold reduction in end-to-end processing time for the complete 1,688-page corpus. The downstream application of the pipeline's output through a RAG-based system further illustrates that high-fidelity, structured transcriptions can directly support substantive historical inquiry, from factual entity extraction to interpretive stance detection.

Several directions for future work remain. First, a systematic evaluation of the Enrichment stage, including entity linking accuracy and the reliability of LLM-based semantic inference on historical texts, is needed to validate the full pipeline beyond its transcription capabilities. Second, the downstream RAG-based analysis presented here is exploratory; a rigorous assessment of factual accuracy and interpretive validity, conducted in collaboration with domain historians, is essential. Third, we intend to evaluate VERITAS on document collections that pose greater palaeographic challenges, such as manuscript sources and early printed books with non-standard typefaces, to assess the generalisability of the approach. Finally, we plan to release the framework as an open-source toolkit to facilitate adoption and community-driven extension across the digital humanities.

Bibliographical References

Mélodie Boillet, Solène Tarride, Yoann Schneider, Bastien Abadie, Lionel Kesztenbaum, and Christopher Kermorvant. 2024. The socface project: Large-scale collection, processing, and analysis of a century of french censuses. In *Document Analysis and Recognition - ICDAR 2024*, pages 57–73, Cham. Springer Nature Switzerland.

Mélodie Boillet, Christopher Kermorvant, and Thierry Paquet. 2021. Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2134–2141.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Floriane Chiffoleau. 2024. Keeping it open: A tei-based publication pipeline for historical documents. *Journal of the Text Encoding Initiative*, Issue 15.

Thomas Constum, Lucas Preel, Théo Larcher, Pierrick Tranouez, Thierry Paquet, and Sandra Brée. 2024. End-to-end information extraction in handwritten documents: Understanding paris marriage records from 1880 to 1940.

Bernardino Corio. 1978. *Storia di Milano*. Classici della Storiografia. Unione Tipografico-Editrice Torinese, Torino. Critical edition of *Historia patria*.

GLM Team et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models.

Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2024. Unlocking the archives: Using large language models to transcribe handwritten historical documents.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. 2025. OCR error post-correction with LLMs in historical documents: No free lunches. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 38–47, Tallinn, Estonia. University of Tartu Library, Estonia.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on*

Document Analysis and Recognition Workshops (ICDARW), volume 2, pages 19–19.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Maria Levchenko. 2025. Evaluating llms for historical document ocr: A methodological framework for digital humanities.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. dots.ocr: Multilingual document layout parsing in a single vision-language model.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frederic Kaplan, Florian Kleber, Roger Labahn, Eva Maria Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro Héctor Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster, and Konstantinos Zagoris. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Clemens Neudecker, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann, and Elisa Herrmann. 2019. Ocr-d: An end-to-end open source ocr framework for historical printed documents. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 53–58, New York, NY, USA. Association for Computing Machinery.

Sina J. Semnani, Han Zhang, Xinyan He, Merve Tekg"urler, and Monica S. Lam. 2025. CHURRO: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Do we still need corpora and corpus analysis platforms? Discourse analysis in times of LLMs

Julia Krasselt, Dolores Lemmenmeier-Batinić, Philipp Dreesen

Zurich University of Applied Sciences, Institute of Language Competence
Theaterstrasse 17, 8400 Winterthur
{krss, leme, dree}@zhaw.ch

Abstract

Corpus-based discourse analysis investigates the linguistic construction of societally shared knowledge by iterating between quantitative pattern detection and qualitative interpretation in large text collections. Large Language Models (LLMs) promise to lower practical barriers to such work (e.g., natural-language querying, qualitative coding), yet they also introduce risks that are especially consequential in discourse-analytic settings, where fluent summaries can encourage ungrounded interpretation. This position paper argues that integrating LLMs into corpus analysis platforms is appropriate only insofar as it remains compatible with three epistemic premises of corpus research: (1) transparency of the data basis and traceability of analytical operations; (2) interpretability as evidence-constrained sense-making; and (3) seriality and patternedness as distributional structure and variation. In this opinion paper, we contribute a platform-oriented requirements perspective that translates these premises into design constraints for tool-calling/RAG-style integration, and we outline implementation directions that treat LLMs as an interaction layer over inspectable corpus retrieval and platform-based analysis.

Keywords: corpus-based discourse analysis, corpus platforms, large language models (LLMs)

1. Background and Motivation

Corpus-based discourse analysis approaches the study of societally shared knowledge by iterating between quantitative pattern detection and qualitative interpretation of language use (lexical choices, argumentative patterns, stance-taking, etc.) across large text collections (cf. Baker, 2023; Baker & McEnery, 2015; Bubenhofer, 2009). Over the past two decades, this approach has been adopted across the humanities and social sciences, including public health sciences, political science, and media studies (e.g., (Grimmer & Stewart, 2013; Krasselt et al., 2022; O'Halloran, 2010), and it also informs applied communication tasks such as discourse-informed message design and stakeholder-oriented communication (Cooren, 2015).

This development has been supported by advances in corpus infrastructure. Curated corpora and platform ecosystems (web-based and local) have made large-scale discourse analysis accessible beyond corpus linguistics. Tools such as Sketch Engine (Kilgarriff et al., 2014), AntConc (Anthony, 2024), and #LancsBox X (Brezina & Platt, 2025) as well as initiatives such as ParlaMint (Erjavec et al., 2023), the Leipzig Corpus Collection (Goldhahn et al., 2012) and Swiss-AL (Krasselt et al., 2023) exemplify this expansion of accessible data and methods.

At the same time, effective corpus-based discourse analysis remains demanding because it is inherently iterative. Robust studies move between quantitative indicators and qualitative inspection—for example, by relating collocation profiles to concordance evidence (Baker, 2023)—and require consequential decisions about discourse modelling, corpus construction, analytical settings, and criteria of interpretive relevance. Difficulties often arise where statistical outputs must be connected to defensible discourse claims and where multiple analytical steps need to be integrated into a coherent, documentable workflow (cf. also McEnery & Brezina, 2022).

Large Language Models (LLMs) may reduce some of these frictions and are beginning to shape expectations about how corpus resources can be accessed and explored by researchers (Brezina, 2025). They can support natural interaction with corpora, assist with query formulation, and guide users through analytical options (e.g., Anthony, 2025; AI integration on english-corpora.org). However, current research demonstrates substantial risks, particularly when generative AI is used for qualitative analysis. Studies report that semantic categorisation of keywords is often generic – especially when items are presented without context – shows only marginal overlap with human-produced categorisations, and may even introduce fabricated assignments during the categorisation process. Reproducibility is an additional concern, given the non-deterministic behaviour of general-purpose models (Curry et al., 2024; Gillings et al., 2024; Morgan, 2023).

Incentives for speed and simplification are even stronger outside academia. Democratic societies depend on the formation of public opinion, which makes the analysis of public meaning-making processes attractive not only for research but also for public-facing monitoring and communication. Organizations such as political parties, public authorities, associations, and NGOs therefore have strong motivations to seek fast "insights" into what can be said, by whom, and with which effects. A current risk is that LLMs will be adopted for these tasks because they generate fluent, plausible-sounding accounts on demand, while concealing data choices and interpretive steps. In contrast, corpus-based discourse analysis provides suitable data, methods and tools for

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 67–72
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

producing more accountable analyses of public meaning-making.

Against this background, we return to the question raised in the title: Do we still need corpora and corpus analysis platforms? We argue for a conditional integration of LLM-based workflows into corpus platforms for discourse analysis. Here, *conditional* means that such integration must remain compatible with the epistemic logic of corpus analysis. By *corpus platforms* we mean integrated, access-controlled environments that stabilise corpus definitions and metadata, support concordance-level inspection, and log platform-side computations and parameters. We propose three epistemic premises: (1) corpus workflows must ensure transparency of the data basis and traceability of analytical operations; (2) interpretation must be supported by inspectable concordance-level evidence and triangulation across measures; and (3) claims must be grounded in seriality and patternedness, evidenced by distributions and structured variation, for example across time, genres, and languages. Building on these premises, we discuss which forms of LLM integration can support corpus-based discourse analysis while remaining anchored in accountable procedures and transparent evidence.

2. Epistemic premises of corpus analysis

2.1 Transparency and Traceability

In corpus analysis, transparency and traceability are essential for justifying interpretations and claims based on empirical evidence (Baker, 2023: 225; Bednarek et al., 2024). Transparency concerns the data basis: it should be clear which corpus (or which parts of it) were analysed and on what textual instances a claim rests. Traceability concerns the process: analytical steps and settings (from preprocessing to statistical measures) should be documented so that results can be reconstructed when the same procedures are applied again. Together, these requirements support interpretability as making sense of patterns in light of a research question (see Section 2.2).

A central advantage of corpora is that they are explicitly delimited and enriched with metadata (e.g., time, genre, speaker, region), enabling controlled subsetting and auditing. This differs from LLMs, whose training data and selection principles are generally not inspectable at the level of individual documents and whose outputs do not provide provenance for the instances that would support a claim (Heersmink et al., 2024). Consequently, LLM responses cannot function as evidence on their own.

Transparency and traceability translate into platform-level documentation of corpus modelling, processing, and analysis. Platforms should record corpus composition and metadata, key processing choices (e.g., deduplication, tokenization, tagging/lemmatization, language identification), and analytical parameters (e.g., measures, thresholds, window sizes, dispersion metrics). Because tools often present aggregated outputs (e.g., keyword or collocation lists), transparency also requires drill-down to concordance lines and contextual evidence.

LLM integration introduces specific risks for transparency. Interfaces may silently reformulate queries or apply undocumented defaults, and generated summaries can remain detached from the empirical evidence if they are not explicitly linked to the retrieved data and the relevant quantitative outputs. For this reason, LLM functionality should be constrained by corpus retrieval and platform-side calculation: if implemented, summaries should be based on computed results, and any claim should link back to concordance-level evidence and distributional information so that users can reconstruct how it was derived.

2.2 Interpretability

In corpus-based discourse analysis, interpretability is a central epistemic requirement because analyses substantiate claims about how shared understandings are distributed and become prevalent in society (Keller, 2024). Since such objects are inseparable from social context and may inform public debate or institutional practice, it is not sufficient that an analysis yields plausible statements; interpretations must be justified in relation to a research question and constrained by the available evidence.

Interpretation denotes the methodological step of relating distributional observations (e.g., collocation profiles, concordance patterns, shifts over time) to an analytical focus and, where relevant, to an applied problem. What counts as salient and how a pattern is understood depends on the question and theoretical perspective; accordingly, there is rarely a single “correct” reading. Interpretability in discourse analysis is thus supported by practices that keep such justification empirically and analytically constrained. This includes contextualisation (reading patterns against relevant co-text and situational knowledge), testing whether an observation is stable or driven by particular sources or periods, and triangulation across alternative operationalisations and measures (Baker, 2023: 44; Bednarek, 2009; Marchi & Taylor, 2009). Equally important is attention to structured variation: differences across genres, time periods, or speaker groups are often analytically central and should remain visible rather than being collapsed into a single narrative.

LLM-assisted analysis introduces specific risks at this interpretive stage. Models may hallucinate examples or explanations not supported by the corpus, and fluent summaries can invite narrative overreach (Ji et al., 2023). Conversational interaction may also amplify confirmation dynamics and encourage premature closure, smoothing analytically consequential variation across time, genres, or groups.

The issue of interpretability also becomes evident when collecting and evaluating ethically sensitive data or data that is protected by copyright, for example. Discourse corpora often contain copyrighted material and potentially sensitive data (e.g., political extremism), the processing of which requires specialist knowledge. Corpus platforms can provide secure research environments and transparent, documented data-handling procedures that generic chat interfaces layered on top of an LLM typically do not. In sum, the interface is changing, but the need for accountable, corpus-based infrastructures for discourse analysis remains.

2.3 Seriality, Patternedness, and Distributional Grounding

In corpus-based discourse analysis, the epistemic goal is not the description of isolated events but the identification of seriality and patternedness in public discourse (Foucault, 1981, 1982). Discourses become socially powerful not because a statement occurs once, but because formulations, evaluations, and argumentative moves recur across texts and sediment into shared assumptions about how the world can or should be understood. This repetition structures what appears normal, plausible, or sayable in society. From a corpus perspective, seriality becomes observable as patterned language use (e.g., recurring lexical choices, frames, metaphors, actor configurations, or stance profiles) that can be shown to be stable or systematically shifting across time, contexts, and communities.

Importantly, discourse-analytic seriality is not simply frequency. A pattern is discourse-relevant when it exceeds the idiosyncrasies of single authors, outlets, or formats and when it appears as a structured distribution across the discursive field. This includes, for example, patterns that are dispersed across sources rather than concentrated in a few texts; patterns that recur across genres while taking genre-specific forms; or patterns that differentiate speaker groups, political camps, or linguistic communities. Seriality therefore combines recurrence with structured variation.

Corpora and corpus platforms support the identification of seriality by making such distributional questions testable. They allow researchers to delimit the discursive space through corpus design and metadata, detect candidate patterns via aggregated measures (e.g., keywords, collocations, association profiles), and inspect instances through concordance evidence and contextual reading. Combined with transparency and traceability, this enables analysts to evaluate whether an observation is genuinely serial by checking dispersion across sources, stability across subcorpora or time periods, and sensitivity to operationalisation (e.g., alternative queries, reference corpora, or window settings).

LLM-assisted workflows pose particular risks to establishing seriality and patternedness, especially when fluent, abstractive summaries become the primary analytic output. Such summaries can smooth heterogeneity and obscure analytically relevant variation. Majority bias may privilege dominant frames and paraphrases, masking marginal positions. In addition, LLM outputs often neglect dispersion and distribution, making it difficult to distinguish broadly shared patterns from event-driven spikes or source-specific effects. In the worst case, the epistemic target shifts from demonstrating seriality as distributional structure to producing plausible accounts of “what the discourse is about.” Conversely, when LLMs are used to support corpus-anchored tasks (e.g., schema-driven, custom annotation, candidate retrieval, or consistency checks across coded instances, cf. Yu et al., 2024) they may strengthen rather than weaken the establishment of patterned variation, provided results remain traceable and are validated through distributional analysis.

3. Implications for the integration of LLMs into platforms for corpus-based discourse analysis

3.1 Criteria for LLM integration

Building on the epistemic premises outlined above, we propose the following criteria for LLM integration into corpus platforms for discourse analysis:

Outputs must remain corpus-grounded and evidence-linked. LLM assistance should operate on an explicitly defined corpus and maintain a clear separation between model text and corpus evidence; analytic statements should link to inspectable contexts (concordances/documents) and the relevant computed outputs.

Workflows must be traceable and reversible. Platforms should log and export executed queries, preprocessing state, parameters, and statistical measures, and make any automated query reformulations or defaults explicit and reversible rather than silently applied.

Computation must remain platform-side. Quantitative results (e.g., keywords, collocations, dispersion) should be computed by the platform; the LLM may guide exploration and explanation but must not substitute for, or fabricate, analytical results.

Interpretation must be supported without narrative smoothing. LLM interaction should encourage contextual inspection, counterevidence, and alternative framings, while preserving structured variation and enabling checks of dispersion and stability across time, genres, groups, and languages.

Hallucination resilience is required. When the available evidence is insufficient, the system should signal these limits and prompt further retrieval and inspection rather than producing confident, ungrounded claims.

3.2 Implementation directions

These criteria matter because corpus-based discourse analysis advances by iteratively testing candidate patterns against textual evidence—a process that is time-consuming and cognitively demanding. Consider studies of argumentative structures in thematic discourses (so-called *topoi*, Wengeler, 2012). Identifying topoi typically involves inspecting keywords, n-grams, and collocates, reading concordances (and, where necessary, full texts), and iteratively building an annotation scheme until higher-order categories stabilise (Kalwa, 2013). The process produces many false positives, since only a small subset of retrieved items is relevant to the analytical aim. LLMs seem to offer a shortcut by answering questions like "Which topoi appear in discourse X?" even before a corpus is delimited. Yet such outputs are not discourse analysis but plausible summaries that bypass corpus definition, distributional testing, and evidence inspection.

Corpus platforms remain central in the LLM era because they stabilise the data basis, document preprocessing and statistical procedures, and preserve access to concordance-level evidence and structured variation. A robust integration strategy therefore treats the LLM primarily as an interaction layer, while retrieval and computation are executed by specialized platform functions and returned in inspectable form. This division of labour is increasingly reflected in tool-calling architectures (e.g., the Model Context Protocol, MCP and related function/tool-calling approaches), which standardise how models invoke external tools and incorporate their outputs (Anthropic, 2024; Hou et al., 2026).

Accordingly, two implementation directions are particularly compatible with the criteria outlined above. First, LLMs can be used to post-process platform results via predefined prompt templates— for instance to sort, filter, cluster, or categorise concordance lines or aggregated outputs (cf. Davies, 2025), while preserving links to evidence and keeping operations reversible. Second, Retrieval-Augmented Generation (RAG, cf. Lewis et al., 2020; Gao et al., 2023) can enable natural-language interaction with user-defined corpus slices by retrieving relevant passages or documents first and using them as grounded context for the model's response. In contrast to "pure" generation, RAG and tool-calling workflows make the evidential basis explicit and allow outputs to be tied back to retrievable sources. In both cases, the decisive requirement is that AI support remains constrained by corpus retrieval and platform-provided results, and that outputs preserve traceability, evidence access, and distributional visibility rather than replacing them with narrative closure.

4. Conclusion

Do we still need corpora and corpus platforms for discourse analysis in the age of LLMs? We argue that we do – because LLMs change the interface to text analysis without removing the epistemic requirements that make discourse-analytic claims accountable. Corpus-based discourse analysis relies on (i) transparency of the data basis and traceability of analytical operations, (ii) interpretability as evidence-constrained sense-making supported by contextualisation and triangulation, and (iii) the demonstration of seriality and patternedness as structured distributions and variation. These premises define what counts as a defensible discourse-analytic result.

The societal stakes sharpen this point. Discourses shape shared assumptions and orders of speech, influencing what becomes sayable and legitimate. This makes discourse analysis a particularly sensitive domain: LLMs can generate fluent, plausible accounts while obscuring data choices and interpretive steps, and are therefore likely to be adopted for public-facing "insights" into opinion formation and communicative dynamics. Without corpus-grounded procedures, such uses risk replacing evidence-based analysis with persuasive narrative.

For these reasons, corpus platforms may become more rather than less relevant by providing orientation amid fast access to unreliable knowledge. Their distinctive contribution is curated, documented corpora and robust views of distributions and structured variation with inspectable concordance evidence. The criteria proposed here translate these premises into design requirements: LLMs may support interaction and exploration, but results must remain corpus-grounded, workflows reconstructable, and variation visible. Tool-calling architectures point to a division of labour in which the LLM acts as an interaction layer while specialized platform tools handle retrieval and quantitative analysis.

5. Bibliographical References

Anthony, L. (2024). *AntConc* (Version 4.3.1)

[Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Anthony, L. (2025). Integrating AI technology into corpus-based language learning through *ChatAI*. *Computer Assisted Language Learning*, 1–19. https://doi.org/10.1080/09588221.2025.2589747

Anthropic. (2024). *Introducing the Model Context Protocol*. https://www.anthropic.com/news/model-context-protocol

Baker, P. (2023). *Using Corpora in Discourse Analysis* (2nd edition). Bloomsbury Academic. https://doi.org/10.5040/9781350083783

Baker, P., & McEnery, T. (Eds). (2015). *Corpora and Discourse Studies. Integrating Discourse and Corpora*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137431738

Bednarek, M. (2009). Corpora and Discourse: A Three-Pronged Approach to Analyzing Linguistic Data. In M. Haugh, K. Burridge, J. Mulder, & P. Peters (Eds), *Selected proceedings of the 2008 HCSNet workshop on designing the Australian national corpus: Mustering languages*.

Bednarek, M., Schweinberger, M., & Lee, K. K. H. (2024). Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory*, *20*(3), 539–566. https://doi.org/10.1515/cllt-2023-0104

Brezina, V. (2025). Corpus linguistics and AI: #LancsBox X in the context of emerging technologies. *International Journal of Language Studies*, *19*(2). https://doi.org/10.5281/ZENODO.15250820

Brezina, V., & Platt, W. (2025). *#LancsBox X* [Computer software]. Lancaster University. http://lancsbox.lancs.ac.uk.

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. De Gruyter.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

Davies, Mark (2025). Comparing the predictions of Large Language Models to actual corpus data. (White papers). English-Corpora.org. https://www.english-corpora.org/ai-llms/

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., De Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., De Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Foucault, M. (1981). The order of discourse. In R. Young (Ed.), *Untying the text: A post-structuralist reader* (pp. 51–78). Routledge & Kegan Paul.

Foucault, M. (1982). *The archaeology of knowledge*. Pantheon Books.

Gillings, M., Kohn, T., & Mautner, G. (2024). The rise of large language models: Challenges for Critical Discourse Studies. *Critical Discourse Studies*, 1–17. https://doi.org/10.1080/17405904.2024.2373733

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Heersmink, R., De Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, *26*(3), 41. https://doi.org/10.1007/s10676-024-09777-3

Hou, X., Zhao, Y., Wang, S., & Wang, H. (2026). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ACM Transactions on Software Engineering and Methodology*, 3796519. https://doi.org/10.1145/3796519

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kalwa, N. (2013). *Das Konzept 'Islam': Eine diskurslinguistische Untersuchung*. De Gruyter.

Keller, R. (2024). *The Sociology of Knowledge Approach to Discourse: Foundations, Concepts and Tools for a Research Programme*. Springer International Publishing. https://doi.org/10.1007/978-3-031-55114-7

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Krasselt, J., Dreesen, P., Fluor, M., & Rothenhäusler, K. (2023). Swiss-AL. Korpus und Workbench für mehrsprachige digitale Diskurse. In M. Kupietz & T. Schmidt (Eds), *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022* (pp. 127–142). Narr Francke Attempto. 10.24053/9783823396024

Krasselt, J., Robin, D., Fadda, M., Geutjes, A., Bubenhofer, N., Suzanne Suggs, L., & Dratva,

J. (2022). Tick-Talk: Parental online discourse about TBE vaccination. *Vaccine*, *40*(52), 7538–7546. https://doi.org/10.1016/j.vaccine.2022.10.055

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Marchi, A., & Taylor, C. (2009). *If on a winter's night two researchers…: A challenge to assumptions of soundness of interpretation*. https://cris.unibo.it/handle/11585/792146

McEnery, T., & Brezina, V. (2022). *Fundamental Principles of Corpus Linguistics* (1st edn). Cambridge University Press. https://doi.org/10.1017/9781107110625

Morgan, D. L. (2023). Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods*, *22*, 16094069231211248. https://doi.org/10.1177/16094069231211248

O'Halloran, K. (2010). How to use corpus linguistics in the study of media discourse. In A. O'Keeffe & M. McCarthy (Eds), *The Routledge Handbook of Corpus Linguistics* (pp. 563–576). Routledge.

Wengeler, M. (2012). *Topos und Diskurs: Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960-1985)*. Niemeyer. https://doi.org/10.1515/9783110913187

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu

GaelEval: Benchmarking LLM Performance for Scottish Gaelic

Peter Devine,[1] William Lamb,[1] Beatrice Alex,[1] Ignatius Ezeani,[2] Dawn Knight,[3] Mícheál J. Ó Meachair,[4] Paul Rayson,[2] Martin Wynne[5]
[1]University of Edinburgh, [2]Lancaster University, [3]University of Cardiff,
[4]Dublin City University, [5]University of Oxford
{pdevine2, w.lamb, b.alex}@ed.ac.uk, {i.ezeani, p.rayson}@lancaster.ac.uk,
knightd5@cardiff.ac.uk, micheal.omeachair@dcu.ie, martin.wynne@ling-phil.ox.ac.uk

Abstract

Multilingual large language models (LLMs) often exhibit emergent 'shadow' capabilities in languages without official support, yet their performance on these languages remains uneven and under-measured. This is particularly acute for morphosyntactically rich minority languages such as Scottish Gaelic, where translation benchmarks fail to capture structural competence. We introduce **GaelEval**, the first multi-dimensional benchmark for Gaelic, comprising: (i) an expert-authored morphosyntactic MCQA task; (ii) a culturally grounded translation benchmark and (iii) a large-scale cultural knowledge Q&A task. Evaluating 19 LLMs against a fluent-speaker human baseline ($n = 30$), we find that Gemini 3 Pro Preview achieves 83.3% accuracy on the linguistic task, surpassing the human baseline (78.1%). Proprietary models consistently outperform open-weight systems, and in-language (Gaelic) prompting yields a small but stable advantage (+2.4%). On the cultural task, leading models exceed 90% accuracy, though most systems perform worse under Gaelic prompting and absolute scores are inflated relative to the manual benchmark. Overall, GaelEval reveals that frontier models achieve above-human performance on several dimensions of Gaelic grammar, demonstrates the effect of Gaelic prompting and shows a consistent performance gap favouring proprietary over open-weight models.

Keywords: benchmarking, multilingual evaluation, large language models, morphologically rich languages, Scottish Gaelic

1. Introduction

Although most large language models (LLMs) officially support a small fraction of the approximately 7,000 human languages spoken worldwide, they display emergent 'shadow' capacities in many more. For instance, OpenAI advertises support for 59 languages in ChatGPT,[1] none of which belong to the Celtic family (e.g. Irish, Welsh and Scottish Gaelic). Despite this, the system processes and generates text in every Celtic language. The distinction between official and *de facto* support raises a methodological challenge: as coverage expands and model varieties diversify, establishing robust evaluation frameworks becomes crucial for both official languages and the minority languages they nevertheless represent.

The current evaluation landscape is markedly skewed. High-resource languages like English benefit from a self-reinforcing ecosystem of training corpora and mature benchmarks. In contrast, low-resource languages suffer from sparse training data (Joshi et al., 2020) and little or no evaluation resources (Romanou et al., 2024). Even where benchmarks exist, they rarely include human baselines (Assadi et al., 2025), making it impossible to ascertain whether a model's output follows a given community's sociolinguistic norms or not. Furthermore, the English-dominance of these models risks processing the world's cultural-linguistic mosaic through an Anglocentric lens. Without objective measurement, academics and language communities alike cannot determine if an LLM should be explored or eschewed.

Scottish Gaelic ('Gaelic') epitomises these challenges. Ranking 104th in Common Crawl accessibility,[2] Gaelic occupies the digital margins, yet it possesses a rich morphosyntax that defies the only benchmarks available for it: the surface-level translation based FLORES-200 (Goyal et al., 2022) and BritEval (BritLLM, 2026). These benchmarks do not capture whether a model is truly 'Gaelic-conversant' or merely performing a high-dimensional translation of English concepts.

To address this gap, we present **GaelEval**: a targeted, multi-dimensional evaluation suite that moves beyond surface equivalence toward deeper morphosyntactic and culturally grounded competence. Our framework includes three distinct tasks:

1. **Linguistic Competence**: A multiple-choice question answering (MCQA) task comprising 120-questions and probing fine-grained grammatical and idiomatic usage.

2. **Translation**: A rigorous assessment using BLEU and chrF metrics against hand-translated gold labels.

[1] https://help.openai.com/en/articles/8357869-how-to-change-your-language-setting-in-chatgpt. Accessed 22 Feb 2026.

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages. Accessed 21 Feb 2026.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 73–85
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3. **Cultural Understanding**: A culturally grounded Q&A task (1,087 questions) derived from pedagogical content produced by fluent speakers.

We evaluate 19 contemporary LLMs (14 proprietary; 5 open-weight), providing the first systematic comparison of LLM performance for Scottish Gaelic. Gemini 3 Pro Preview leads overall and surpasses the fluent-speaker baseline on the linguistic competence task.

Our principal contributions are:

- **GaelEval**, the first multi-dimensional benchmark for Scottish Gaelic, spanning morphosyntax, translation and culturally grounded knowledge;
- the first human baseline for Gaelic LLM evaluation ($n = 30$);
- evidence that Gemini 3 Pro Preview exceeds the fluent-speaker mean on a controlled morphosyntactic task;
- a consistent aggregate advantage for in-language (Gaelic) prompting for the morphosyntactic task; and
- quantification of the performance gap between frontier proprietary and open-weight models in a minority-language setting.

In what follows, we review related work (§2), describe our design and evaluation methodology (§3), present empirical results (§4) and conclude, with proposed directions for future work (§5).

2. Related Work

Multilingual LLM Evaluation Frameworks Large-scale multilingual benchmarks are central to evaluating LLM capabilities across languages. MMLU (Hendrycks et al., 2020) introduced a widely used multiple-choice framework for knowledge-intensive reasoning in English. Global MMLU (Singh et al., 2025) extended this paradigm cross-lingually, largely via translation of English-source materials. FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2024) expanded coverage to 200+ languages, including Gaelic, but evaluates only machine translation (MT). BritEval (BritLLM, 2026) consists of 3 major English benchmarks translated into 4 languages from Britain and Ireland, including Gaelic. XTREME-UP (Ruder et al., 2023) incorporates additional low-resource tasks (e.g., transliteration, OCR), while INCLUDE (Romanou et al., 2024) departs from translation-based design by constructing question answering benchmarks from native regional exam materials.

Many multilingual benchmarks rely heavily on translation or adaptation from English-centric datasets. While valuable, this approach under-represents language-specific morphosyntax, culturally grounded knowledge, and idiomatic usages that resist direct translation (e.g. that the colour of grass in Gaelic is *gorm* 'lit. blue', not *uaine* 'lit. green'). For morphologically rich languages such as Gaelic, translation-based evaluation also is unlikely to capture fine-grained inflectional contrasts or edge cases that distinguish structural competence from superficial word recognition. To our knowledge, beyond BritEval, FLORES-200 and related benchmarks (e.g., SIB-200; Adelani et al., 2024), no large-scale Gaelic evaluation suite exists.

Low-Resource and Morphologically Rich Language Evaluation Recent work addresses the challenges of evaluating LLMs on low-resource and morphologically complex languages, including tokenisation and pattern extrapolation (Xia et al., 2025). IndicGenBench (Singh et al., 2024) covers 29 Indic languages using human-curated parallel data; AfriQA (Ogundepo et al., 2023) introduces question answering for African languages; and TurkBench (Toraman et al., 2026) evaluates Turkish across 21 subtasks. Xia et al. (2025) further propose a cross-lingual benchmark spanning Cantonese, Japanese and Turkish, combining human evaluation with automated metrics across diverse tasks. Irish-BLiMP evaluates LLMs on Irish linguistic knowledge using 1020 minimal pairs and provides a human baseline (McGiff et al., 2025). Collectively, these efforts highlight the need to evaluate LLMs on morphologically rich, low-resource languages. We extend this line of work by directly assessing model competence in Gaelic morphosyntax and non-compositional usage.

Culturally and Linguistically Informed Evaluation A growing literature argues that linguistic competence cannot be evaluated independently from cultural knowledge. Tao et al. (2024) document systematic bias toward English-speaking contexts in ostensibly multilingual LLMs. Relatedly, multilingual models have been shown to process non-English inputs through English-dominant representational pathways (Papadimitriou et al., 2023; Wendler et al., 2024), raising concerns about whether these systems encode language-specific structures or just rely on Anglocentric priors.

Recent benchmarks increasingly integrate cultural and linguistic evaluation. For example, ProverbEval (Azime et al., 2025) assesses Ethiopian languages (and English) through proverb interpretation, requiring both morphosyntactic competence and culturally grounded reasoning. Knowledge-grounded benchmarks similarly test community-specific factual knowledge in domains such as food, holidays and social practices (Myung et al., 2025). Importantly, Myung et al. (2025) show that in-language prompting benefits medium- and high-resource languages, whereas low-resource languages often perform better under English prompting. This resource-sensitive pattern motivates our evaluation under both English and Gaelic prompt conditions to test whether similar asymmetries arise for Gaelic.

LLM-Assisted Benchmark Construction LLM-assisted benchmark generation offers a practical solution when human-curated datasets are scarce (Perez et al., 2023; Anwar et al., 2026). Prior work has used LLMs to extract cultural knowledge from large corpora such as C4 (Nguyen et al., 2023) and TikTok (Shi et al., 2024), derive culturally grounded Q&A from web scrapes (Wang et al., 2024) and Wikipedia (Fung et al., 2024), and generate multilingual evaluation data across 13 languages (Zhao et al., 2025). Although LLMs typically perform worse on low-resource languages, raising concerns about synthetic benchmark quality, manual analysis of 10,000 generated instructions in 13 Indic languages found over 99.7% to be of high or moderate quality (Chitale et al., 2025).

In our setting, automated generation was required for scale. To reduce associated risks, we applied structured filtering and answerability scoring (§3.1.3), discarding items below predefined thresholds. While not a substitute for native-speaker validation, this procedure provides a systematic safeguard against noise and incoherence.

3. Benchmark Design and Evaluation Methods

In this section, we describe the design of **GaelEval** and our evaluation methods. Unlike translation-based frameworks such as FLORES-200, GaelEval integrates an expert-designed morphosyntactic MCQA task with culturally grounded Gaelic-source texts for translation and Q&A evaluation.

3.1. Tasks

3.1.1. Linguistic Competence

We define *linguistic competence* as the ability to select grammatically and idiomatically appropriate forms in controlled morphosyntactic contexts. The 120-item MCQ set was designed by a Gaelic domain expert, who used a recent grammar (Lamb, 2024) to identify grammatical edge cases (e.g. long-distance relativisation) and constructions resistant to literal translation from English.[3] Items span 16 grammatical categories (Table 1), with nominal morphology the largest (17 items; 14.6%). The design prioritised breadth across grammatical domains (e.g. case marking, agreement, idiomatic conventions) over depth within individual micro-phenomena (e.g. the feminine singular genitive), while keeping the task manageable for human participants. The MCQs were not publicly available prior to evaluation, ensuring zero-shot conditions.

Category	**N**
Nominal morphology	17
Adjectives	11
Verbal noun cores	12
Formulaic expressions	10
Questions and tags	10
Prepositions	9
Pronouns and anaphor resolution	9
Tense Aspect Modality (TAM) system	7
Impersonals and passives	7
Adverbials	5
Conjunctions and particles	5
Relative clauses	5
Clefts and focussing expressions	4
Colours	3
Determiners	3
Numerals	3
Total	**120**

Table 1: Distribution of MCQs across principal grammatical categories.

Each question contained a single gap and was designed to have one unambiguous correct answer. For example, the following question asks for the feminine singular basic definite form of the noun *fuil* 'blood', where b. is the correct answer.

```
Chunnaic Màiri ____.
('Mary saw ____.')
a. am fuil
b. an fhuil
c. am fhuil
d. na fala
```

Distractors were constructed to be linguistically plausible and, as much as possible, attested in contemporary usage. This ensured that a successful choice required grammatical discrimination rather than mere lexical recognition. The placement of the correct answer was varied during preparation, but the task was issued in a fixed order across all iterations to ensure consistency for humans and models.

Human participants were recruited via convenience sampling on social media (Facebook and

[3] Our approach is inspired by an unpublished Irish-language MCQA study by Joseph McInerney.

	FLUENT		NON-FLUENT		
Onset	**Near-/Native**	**Adv**	**Upp-Int**	**Int**	**Sum**
0–4	12	0	0	0	12
12–18	2	2	1	0	5
18–30	7	5	3	1	16
30+	2	0	0	0	2
Total	**23**	**7**	**4**	**1**	**35**

Table 2: Participant distribution by age of first exposure to Gaelic (Onset), self-reported proficiency level ($N = 35$) and fluency grouping. 'Fluent' combines near-/native and advanced.

LinkedIn), yielding a voluntary, non-representative sample of Gaelic speakers ($N = 35$; $n = 30$ after filtering). The task was administered online via Qualtrics. Following informed consent and instructions in English, for accessibility, responses were collected anonymously along with self-reported proficiency and age of acquisition ("Onset"). Proficiency levels were: Near-/Native, Advanced (fluent in most contexts), Upper intermediate (comfortable in most discussions), and Intermediate (everyday conversational ability).

For benchmarking, we combined the Near-/Native and Advanced groups as 'Fluent' to approximate stable adult competence. While native speakers typically acquire Gaelic in childhood, advanced and near-native speakers often receive formal instruction. This may increase familiarity with the prescriptive grammatical forms targeted in the MCQA.

The distribution of participants by onset and proficiency is shown in Table 2. Approximately one third (12/35) identified as Near-/Native and reported acquiring Gaelic before age five, consistent with socialisation in a Gaelic-speaking home. We refer to this group as 'native' in Table 5.

Mean task duration was 45.7 minutes (median = 33.75; range = 12.5–165.2). Questions were presented in fixed order, and participants were required to select a single response per item. Missing responses were scored as incorrect; submissions with more than 10 missing items were excluded.

Following task administration, a few questions were flagged as potentially dialect-sensitive or admitting multiple acceptable responses. After review in light of documented Gaelic morphological variation (Adger, 2010; Iosad and Lamb, 2020), three items (IDs 12, 21, 48) were excluded. All reported results, therefore, are based on 120 questions versus the original 123.

We input the MCQs to the models listed in Table 3 (see §3.2) with each item evaluated in a single-turn call. For each call, the prompt comprised a fixed system instruction and a user message containing the question sentence with a single blank and the full list of answer options (see §6.1). Models were instructed to return only the text of the correct option (e.g. `b. an fhuil`), without explanation or additional punctuation, and short examples were included to enforce this format.

Decoding parameters were left at model defaults. Responses were scored by exact string match after trimming whitespace; outputs beginning with `Error:` were logged as API failures. To mitigate transient failures and rate limits, calls were retried with exponential back-off (up to five attempts), and per-item outputs and correctness flags were stored in JSONL format.

3.1.2. Culturally Relevant Translation

To construct the translation task, we collected parallel English and Gaelic transcripts of the Gaelic learning podcast *An Litir Bheag* ('The Little Letter') from LearnGaelic.scot. As these transcripts are professionally translated, we treat them as gold references for English–Gaelic evaluation. Produced by a fluent Gaelic speaker for intermediate and advanced learners, the episodes frequently address culturally salient topics, making them a relevant resource for assessing LLM performance on MT.

We downloaded all available episodes of *An Litir Bheag* with both English and Gaelic transcripts at the time of writing (episodes 154–1076). Five episodes were excluded due to failed mp3 downloads. (We had initially intended to compile a parallel corpus including Gaelic audio to support ASR evaluation.) The final dataset comprises 918 English–Gaelic parallel transcripts.

On manual inspection, we found errors in how the podcast producers had published some of the podcasts, and so we performed filtering to ensure general data quality. First, we found that some transcripts had their identifying language reversed (i.e. Gaelic vs English, and vice-versa), so we applied a text language identification model, OpenLID v2 (Burchell et al., 2023), to all transcripts and manually removed those that showed switched languages.

We also identified cases where transcripts did not correspond to the correct podcast or failed to align as parallel pairs. To detect such mismatches systematically, we translated all Gaelic transcripts into English using GPT-5.2 and computed BLEU scores against the paired English versions using SacreBLEU (Post, 2018). Episodes were sorted by lowest BLEU score to identify likely mismatches. Following automated filtering and manual inspection, 10 episodes were removed, yielding a final dataset of 908 parallel transcripts.

Finally, we downloaded episode subject metadata from the LearnGaelic.scot website and aligned it with the episode transcripts. This enabled us to

Open-weight: DeepSeek R1, GLM 4.7, GPT OSS 120B, GPT OSS 20B, Llama 4 Maverick. **Closed-weight:** Claude Haiku 4.5, Claude Opus 4.6, Gemini 2.5 Flash, Gemini 3 Flash Preview, Gemini 3 Pro Preview, GPT-4.1, GPT-4.1 Mini, GPT-4.1 Nano, GPT-4o, GPT-4o Mini, GPT-5, GPT-5 Mini, GPT-5 Nano, GPT-5.2

Table 3: Models evaluated in GaelEval.

classify each episode according to one of the following thematic categories: Folklore, Gaelic language, History, Nature, Pastimes, People or Places.

3.1.3. Q&A Task on Cultural Understanding

Alongside our parallel transcripts, we also generated a MCQ set to assess the LLMs' level of Gaelic understanding and knowledge. We first instructed GPT-5.2 to rate the cultural significance of each episode's transcripts from 1-5 (system messages are detailed in the Appendix 6.2). For this task, we removed any episode rated less than 4 to filter transcripts unrelated to general knowledge of Gaelic culture (e.g. autobiographic episodes). Following this procedure, we maintained 713 episodes out of the original 908.

We then instructed GPT-5.2 to generate between 1 and 10 general knowledge-style Q&A pairs per episode based on the episode transcripts in Gaelic, alongside English translations of each question and answer. This yielded 6,802 Q&A pairs.

Occasionally, generated questions referred to transcript-specific details despite instructions to avoid contextual dependence, rendering them unsuitable as stand-alone general knowledge items. We therefore used GPT-5.2 to assign an 'answerability' score (1–5) to each Gaelic question and its English translation, where 5 denotes a fully self-contained general-knowledge item and 1 indicates contextual dependence. Questions scoring below 4 in either language were excluded. This filtering yielded a final set of 1,087 questions drawn from 440 unique episodes, a subset of which was manually verified by a Gaelic domain expert.

Finally, we re-input the transcripts to GPT-5.2 together with the generated questions and answers, instructing it to produce three plausible distractors per item. English translations were also generated for each distractor. This yielded 1,087 multiple-choice questions, each comprising one correct answer and three distractors. Answer options were randomly shuffled and labelled A–D, to prevent positional bias and ensure consistent single-letter responses from models.

3.2. LLM Models Evaluated

We evaluated 19 models (Table 3) across three tasks: linguistic competence, translation and cultural understanding, spanning both open- and closed-weight systems. Models were accessed via the OpenAI, Anthropic, Google AI Studio and Together AI endpoints, with batch processing used to reduce costs. Responses were constrained to a predefined JSON schema (single key–value pair) to minimise explanatory output preceding the answer.

3.3. Evaluation Metrics

For the MCQA tasks, we report accuracy, defined as the percentage of outputs that both conformed to the required JSON schema (see Section 3.2) and contained the correct answer. For translation, we report BLEU (Papineni et al., 2002) and chrF (Popović, 2015).

4. Results

4.1. Linguistic Competence Task

As detailed in §3.1.1, we deployed a 120-item MCQ set to assess models' Gaelic linguistic competence. We also administered the same task to human participants ($N = 35$). Table 4 reports model accuracy under two prompt conditions (Gaelic and English system messages: see §6.1), alongside the performance of fluent speakers ($n = 30$); we exclude intermediate learners to approximate stable adult competence.

Gemini 3 Pro Preview achieves the strongest overall results, scoring 83.3% under Gaelic prompting and 80.0% under English prompting. Its Gaelic-prompted score is significantly higher than the fluent-speaker mean of 78.1% (95% CI: 73.9%-82.2%; $p < 0.05$). This contrasts with comparable work on Irish, where fluent speakers ($n = 3$) outperformed all evaluated LLMs (McGiff et al., 2025). Among OpenAI's models, GPT-5 performs best (69.2% Gaelic; 71.7% English), while Claude Opus 4.6 leads the Anthropic systems (59.2% Gaelic; 50.8% English), though it remains more than 24 percentage points below Gemini 3 Pro Preview under Gaelic prompting.

Among open-weight systems, DeepSeek R1 (45.0% Gaelic; 42.5% English) and Llama 4 Maverick (45.0% Gaelic; 40.0% English) performed best, yet remained approximately 38 percentage points below Gemini 3 Pro Preview. Both models were only modestly above the 25% chance baseline, consistent with results reported for other morphologically rich low-resource languages (Etxaniz et al., 2024; McGiff et al., 2025). This gap may reflect differences in training data scale and composition between open- and closed-weight systems. A plausible factor in Google's favour is its long-standing support for Scottish Gaelic in Google Translate (since 2016) (BBC, 2016), and the associated accumulation of Gaelic data.

We observe a modest aggregate advantage for Gaelic prompting. Averaged across models (excluding humans), Gaelic system messages outperform English ones by 2.4 percentage points (46.8% vs. 44.4%). Although the effect varies by model family, the overall pattern indicates a small but consistent in-language advantage. This pattern is interesting in light of Myung et al. (2025), who find that in-language prompting benefits medium- and high-resource languages but not typically low-resource ones. At the same time, the advantage is small and contrasts to our findings for the cultural understanding task (see §4.3, Table 7).

Notwithstanding the small sample size (see §3.1.1), advanced speakers – most of whom reported acquiring Gaelic between ages 18 and 30 – outperformed native speakers on this task (82.6% vs 70.9%; Table 5). Feasibly, some tested phenomena involve formal registers or infrequent edge cases more likely to be acquired through structured learning than everyday usage. Within our sample, self-reported fluency therefore does not fully align with prescriptive grammatical knowledge.

Within model families, we observe patterns consistent with known scaling-law behaviour. Performance improvements have been shown to follow predictable power-law trends as training compute, dataset size and parameter count increase proportionately (Kaplan et al., 2020; Hoffmann et al., 2022). Table 4 reflects this tendency: more recent flagship models generally outperform earlier versions within the same family (e.g. Gemini 3 Pro: 83.3% vs Gemini 2.5 Flash: 61.7% in the Gaelic condition). Similar intra-family gains are observed among OpenAI and Anthropic models.

Smaller 'Mini' and 'Nano' variants – designed to reduce parameter count and inference cost through compression and distillation – consistently underperform their corresponding flagship models (e.g. GPT-5: 69.2% vs GPT-5 Mini: 47.5% under Gaelic prompting), reflecting the expected trade-off between efficiency and representational capacity. While recent work shows that smaller models can achieve strong performance in low-resource contexts when specifically adapted (Etxaniz et al., 2024), our results indicate that, without such adaptation, reduced-capacity variants perform well below large proprietary models for Gaelic.

Overall, under Gaelic prompting, Gemini 3 Pro Preview demonstrates linguistic competence in Gaelic exceeding that of fluent speakers, with GPT-5's English-prompted performance slightly under the human baseline. Although the extent to which this task generalises to broader real-world text production remains unclear, we observe comparable performance patterns in the translation and cultural understanding tasks discussed below, supporting its construct validity.

Model / Group	Gaelic	English	Δ G-E
Gemini 3 Pro Prev	**83.3**	**80.0**	3.3
Gemini 3 Flash Prev	79.2	77.5	1.7
Humans (Fluent)	n/a	78.1	–
GPT-5	69.2	71.7	-2.5
Gemini 2.5 Flash	61.7	62.5	-0.8
Claude Opus 4.6	59.2	50.8	8.3
GPT-4o	50.0	46.7	3.3
Claude Haiku 4.5	47.5	43.3	4.2
GPT-5 Mini	47.5	41.7	5.8
DeepSeek R1	45.0	42.5	2.5
Llama 4 Maverick	45.0	40.0	5.0
GPT-5.2	43.3	40.0	3.3
GPT-4.1	42.5	42.5	0.0
GPT-5 Nano	36.7	30.8	5.8
GPT-4o Mini	34.2	30.0	4.2
GPT OSS 120B	33.3	27.5	5.8
GLM 4.7	32.5	26.7	5.8
GPT-4.1 Nano	30.0	31.7	-1.7
GPT OSS 20B	29.2	22.5	6.7
GPT-4.1 Mini	24.2	27.5	-3.3
Mean (models)	46.8	44.4	2.4

Table 4: MCQA accuracy (%) under Gaelic and English prompting. Δ indicates Gaelic minus English performance. Human baseline shown for fluent speakers (n=30).

Finally, Figure 1 disaggregates accuracy by grammatical category (cf. Table 1), comparing the fluent-speaker baseline with the nine highest-performing models. Two contrasts are evident. First, models underperform humans on interactional and idiomatic material – most clearly in questions/tags (Human = 0.85; Gemini 3 Pro = 0.50; GPT-5 = 0.30) and formulaic expressions (Human = 0.82; top models ≈ 0.40–0.60). Second, several models match or exceed the human mean in more structural domains, including determiners, pronouns and anaphor resolution, and relative clauses (e.g. REL: Human = 0.53 vs. Gemini 3 Pro ≈ 0.80). Given uneven category sizes, these differences are indicative rather than definitive. Nevertheless, if replicated with a larger, balanced MCQA, the pattern would suggest that current models handle codified morphosyntax more reliably than interactional discourse, with implications for conversational AI and CALL targeting everyday Gaelic.

4.2. Culturally Relevant Translation

Translation metrics are reported in Table 6. We observe consistent score degradation when translating into Gaelic (en→gd) versus English (gd→en).

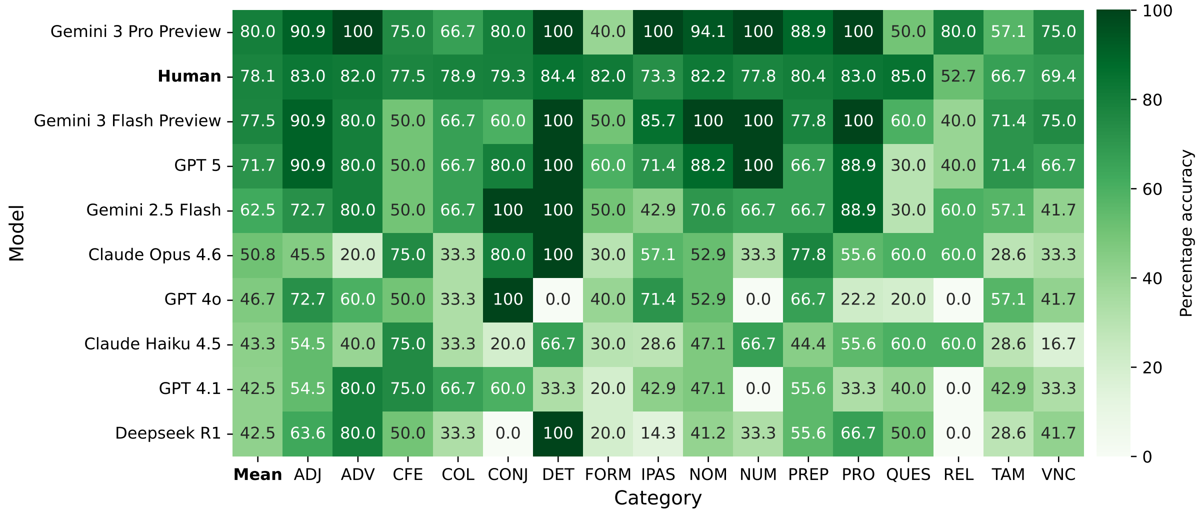


Figure 1: Linguistic competence accuracy by grammatical category for the human baseline and the top nine most-performant models (0–1 scale; darker = higher accuracy). ADJ = adjectives; ADV = adverbials; CFE = clefts and focussing expressions; COL = colours; CONJ = conjunctions and particles; DET = determiners; FORM = formulaic expressions; IPAS = impersonals and passives; NOM = nominal morphology; NUM = numerals; PREP = prepositions; PRO = pronouns and anaphor resolution; QUES = questions and tags; REL = relative clauses; TAM = Tense-Aspect-Modality system; VNC = verbal noun cores. Means are cross-category and so differ from those in Table 4. English prompting conditions used.

Level	**N**	**Mean**	**Max**	**Min**
Near-/Native	23	76.7	98.3	53.3
Native only	12	70.9	85.0	53.3
Advanced	7	**82.6**	97.5	63.3
Upper Intermediate	4	73.1	79.2	67.5
Intermediate	1	41.7	41.7	41.7
Fluent	30	78.1	98.3	53.3

Table 5: Human accuracy (%) on the 120-item MCQA by self-reported proficiency. 'Fluent' includes near-/native and advanced speakers.

	en → gd		**gd → en**	
	BLEU	**CHRF**	**BLEU**	**CHRF**
Gemini 3 Flash Preview	**71.47**	**79.07**	71.75	77.38
Gemini 3 Pro Preview	65.56	78.57	73.41	**78.20**
Gemini 2.5 Flash	65.53	74.72	74.36	77.58
GPT-4.1	65.41	74.62	66.60	72.77
Deepseek R1	62.12	71.84	**75.28**	77.73
GPT-5.2	61.19	72.16	71.02	76.18
GPT-5	56.94	70.27	70.27	75.82
Claude Opus 4.6	53.16	69.34	70.37	77.38
GPT-4o	52.39	67.83	68.73	74.80
GPT-5 Mini	49.19	60.93	69.58	73.34
Llama 4 Maverick	42.93	62.30	73.49	75.74
GPT-5 Nano	42.56	55.54	61.78	67.41
GPT-4.1 Mini	38.90	58.01	69.09	75.48
Claude Haiku 4.5	38.44	59.27	72.32	73.92
GPT OSS 120B	37.79	54.74	56.23	63.01
GPT-4o Mini	34.02	56.84	69.01	73.29
GPT-4.1 Nano	33.93	50.35	63.31	69.04
GPT OSS 20B	0.00	0.00	48.21	53.38
GLM 4.7	0.00	0.00	0.00	0.00

Table 6: BLEU and CHRF scores for both English to Gaelic (en→gd) and Gaelic to English (gd→en) translation tasks.

This corroborates prior work showing that generation in a low-resource language imposes greater representational demands than comprehension mapped back to a high-resource language such as English (Goyal et al., 2022).

Gaelic to English MT largely tests how well a model projects low-resource inputs into English-dominant latent spaces and then generates in English. Conversely, English to Gaelic MT demands active production of morphologically complex and culturally-grounded forms. This asymmetry is especially pronounced in GPT 5 Mini, which drops over 20 BLEU points across directions. Gemini 3 Flash Preview is the exception, maintaining near symmetry (71.47 BLEU en→gd; 71.75 BLEU gd→en), suggesting a more balanced multilingual pre-training distribution.

Directional asymmetry is particularly pronounced among open-weight models. DeepSeek R1 attains the highest BLEU score for gd→en translation (75.28), marginally surpassing leading proprietary systems, but drops to 62.12 BLEU for en→gd generation. This pattern suggests that while mid-tier open-weight models can parse and comprehend Gaelic, they lack the generative capacity to produce morphologically fluent output. The effect is most extreme in GPT OSS 20B, which achieves 48.21 BLEU for gd→en yet collapses to 0.00 BLEU for en→gd, failing to generate valid Gaelic within the

Model	Gaelic	English	Δ G-E
Gemini 3 Flash Prev	**91.35**	**91.63**	-0.28
Gemini 3 Pro Prev	91.17	90.98	0.19
GPT-5	85.28	88.50	-3.22
Gemini 2.5 Flash	79.85	83.44	-3.59
Claude Opus 4.6	79.48	83.53	-4.05
GPT-5 Mini	71.48	80.04	-8.56
GPT-5.2	66.70	70.65	-3.95
GPT-4.1	66.24	74.70	-8.46
GPT-4o	63.85	73.41	-9.56
GLM 4.7	63.75	72.40	-8.65
Llama 4 Maverick	59.06	70.01	-10.95
DeepSeek R1	58.33	75.53	-17.20
GPT-5 Nano	54.37	70.29	-15.92
GPT OSS 120B	52.81	70.56	-17.75
GPT-4.1 Mini	47.93	65.59	-17.66
GPT-4o Mini	43.05	65.04	-21.99
GPT OSS 20B	42.41	57.04	-14.63
Claude Haiku 4.5	40.29	47.84	-7.55
GPT-4.1 Nano	34.87	51.61	-16.74
Mean	63.75	73.30	-9.55

Table 7: Accuracy (%) on the cultural knowledge Q&A task under Gaelic and English prompting. Δ indicates Gaelic minus English performance.

required JSON format.

We also note that GLM 4.7 was not able to form a single correctly formatted JSON response to any of our translation requests. This indicates that prompting some LLMs to process long passages of low-resource languages may degrade their ability to perform more basic tasks, such as JSON formatting.

Finally, divergence between word-level BLEU and character-level chrF highlight the challenges posed by Gaelic morphology. Gaelic's rich inflectional morphology means a model may retrieve the correct lemma yet miss the surface form that BLEU requires. Accordingly, chrF scores are consistently higher and less variable, particularly for en→gd translation, indicating difficulty with morphological realisation.

4.3. Q&A Task on Cultural Understanding

Table 7 reports accuracy on the Gaelic- and English-prompted versions of the cultural Q&A task. Gaelic-prompted performance ranges from 34.87% (GPT-4.1 Nano) to 91.35% (Gemini 3 Flash Preview), indicating substantial variation in cultural knowledge and reasoning. The strongest models – Gemini 3 Flash Preview, Gemini 3 Pro Preview, and GPT-5 – perform consistently well across both languages, with less than a 4-point gap between conditions.

In contrast to the linguistic competence task (see Table 4), most models perform worse under Gaelic prompting (mean Δ = -9.55), with the disparity widening among weaker systems. For example, GPT OSS 20B declines from 57.04% (English) to 42.41% (Gaelic), while GPT-4.1 Nano drops from 51.61% to 34.87%, approaching chance. This pattern aligns with prior findings that LLM representations for low-resource languages are weaker than for English (Goyal et al., 2022; Romanou et al., 2024; Singh et al., 2025).

Comparing these results with the hand-curated Linguistic Competence task reveals strong rank correlation across benchmarks. The top three models – Gemini 3 Flash Preview, Gemini 3 Pro Preview, and GPT-5 – retain the same ordering, and mid-tier open-weight systems (e.g., DeepSeek R1, Llama 4 Maverick) and smaller distilled variants exhibit similar relative positions.

However, absolute scores are inflated on the synthetically generated cultural QA task. The strongest models exceed 90% accuracy, whereas the linguistic competence MCQ peaks at 83.3%. This suggests that LLM-generated benchmarks may create easier evaluation conditions while nevertheless preserving relative performance rankings.

For low-resource languages lacking curated datasets, this result is encouraging: frontier models can generate synthetic cultural benchmarks from relevant authentic text, enabling scalable comparative evaluation. Although such datasets are not reliable as absolute measures of capability – scores may over-estimate fluency – they provide a low-cost and practical heuristic when creating a manual dataset would be prohibitive.

5. Conclusion

This paper introduces **GaelEval**, the first multi-dimensional benchmark for Scottish Gaelic, combining expert-authored morphosyntactic MCQA, culturally grounded translation and large-scale cultural knowledge evaluation. Across 19 contemporary LLMs, we observe variation in Gaelic competence, with proprietary systems outperforming open-weight models by a large margin. Notably, Gemini 3 Pro Preview exceeds the fluent-speaker baseline on the linguistic competence task, while translation and cultural Q&A results reveal directional asymmetries and consistent performance gaps between English- and Gaelic-prompted conditions.

Category-level analysis on the linguistic competence task suggests that current models handle codified morphosyntactic structure more reliably than interactional or formulaic usage. At the same time, rank correlation across the three tasks indicates that synthetic, LLM-generated cultural benchmarks can provide reliable evaluation signals, even if absolute scores may be inflated.

For minority languages lacking established evaluation infrastructure, GaelEval demonstrates that rigorous, language-specific benchmarking is both feasible and necessary. Future work will expand human validation, balance per-category item counts and incorporate further assessment of fluency and sociolinguistic appropriateness. Robust evaluation remains essential if 'shadow' LLM competence in low-resource languages is to be understood, trusted and responsibly deployed.

Data and Code Availability

To preserve the integrity and reusability of GaelEval as a zero-shot evaluation benchmark, the underlying dataset is not being released at this time. Public release likely would expose the data to web scraping bots, leading to its inclusion in future model pre-training corpora and compromising the benchmark's effectiveness. To support ongoing evaluation without exposing the data, a Gaelic LLM leaderboard is planned at `https://eist.ac.uk`. All code for prompt construction, API interaction, and evaluation will be released in the following public repository upon publication, ensuring transparency and reproducibility: `https://github.com/Peter-Devine/gaeleval`.

Ethical Considerations

Institutional ethical approval for this research was sought on 27 January 2026 and granted on 9 February 2026 by the Ethics Officer of the first author's host institution. The study involved minimal risk to participants and did not collect personally identifiable information beyond self-reported proficiency and age of acquisition.

Although no substantial personal or social risks are associated with this work, we acknowledge the environmental costs associated with large-scale model inference. While the evaluation tasks reported here required limited compute, the broader ecological impact of LLM development remains an important consideration.

Limitations

For the linguistic task, the human sample is small and non-representative (35 participants; 30 after filtering) and the MCQA is relatively short (c.f. McGiff et al., 2025) and unbalanced across categories. GPT-5.2 was used for data generation and filtering and was also evaluated, which may inflate its cultural Q&A performance due to self-preference bias (Wataoka et al., 2024; Xu et al., 2025). Moreover, the cultural Q&A benchmark is entirely LLM-generated, although a subset was reviewed by a Gaelic domain expert.

Finally, the evaluation focuses on linguistic and cultural accuracy rather than general reasoning or mathematical ability, as commonly assessed in English-language benchmarks. Models may exhibit different reasoning performance when operating in a low-resource language setting.

Acknowledgements

This work was carried out by members of the CLARIN Knowledge Centre for Digital Resources for the Languages in Ireland and Britain (DR-LIB) as part of the project 'Unlocking AI for the Languages in Britain and Ireland' project, funded by EPSRC (project number UKRI3181).

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

David Adger. 2010. Gaelic morphology. In Moray Watson and Michelle Macleod, editors, *The Edinburgh Companion to the Gaelic Language*, pages 283–303. Edinburgh University Press, Edinburgh.

Muhammad Raheel Anwar, Shah Khalid, Saied Alshahrani, Hafiz Syed Muhammad Bilal, and Mohammed Aldawsari. 2026. MCQs generation with large language models: A survey of methodologies, evolution, and open research issues. *IEEE Access*, 14:10991–11018.

Adnan El Assadi, Isaac Chung, Roman Solomatin, Niklas Muennighoff, and Kenneth Enevoldsen. 2025. Hume: Measuring the human-model performance gap in text embedding tasks.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6265–6281, Albuquerque, New Mexico. Association for Computational Linguistics.

BBC. 2016. Google translate introduces 13 new languages including scots gaelic and sindhi.

BritLLM. 2026. BritLLM: Freely available large language models for UK languages and use-cases. Wayback Machine archive of `https://llm.org.uk/`. Archived January 21, 2026. Available at: `https://web.archive.org/web/20260121124627/https://llm.org.uk/`.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Pranjal A Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. 2025. UPDESH: Synthesizing grounded instruction tuning data for 13 indic. *arXiv preprint arXiv:2509.21294*.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972.

Yi R Fung, Chenkai Sun, Jae Doo, Ruining Zhao, and Heng Ji. 2024. No culture left behind: Massively multi-cultural knowledge acquisition & LM benchmarking on 1000+ sub-country regions and 2000+ ethnolinguistic groups. *arxiv*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10.

Pavel Iosad and William Lamb. 2020. Dialect variation in scottish gaelic nominal morphology: A quantitative study. *Glossa*, 5(1):1–31.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

William Lamb. 2024. *Scottish Gaelic: A Comprehensive Grammar*. Routledge, Oxon.

Josh McGiff, Khanh-Tung Tran, William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan, Hoang D Nguyen, and Nikola S Nikolov. 2025. Irish-BLiMP: a linguistic benchmark for evaluating human and language model performance in a low-resource setting. *arXiv preprint arXiv:2510.20957*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pages 1907–1917.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. 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.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: an online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):1–9.

Çağrı Toraman, Ahmet Kaan Sever, Ayse Aysu Cengiz, Elif Ecem Arslan, Görkem Sevinç, Mete Mert Birdal, Yusuf Faruk Güldemir, Ali Buğra Kanburoğlu, Sezen Felekoğlu, Osman Gürlek, et al. 2026. TurkBench: A benchmark for evaluating Turkish large language models. *arXiv preprint arXiv:2601.07020*.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy Chen. 2024. CRAFT: Extracting and tuning cultural instructions from the wild. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*.

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.

Chengxuan Xia, Qianye Wu, Hongbin Guan, Sixuan Tian, Yilun Hao, and Xiaoyu Wu. 2025. Evaluating modern large language models on low-resource and morphologically rich languages: A cross-lingual benchmark across Cantonese, Japanese, and Turkish. *arXiv preprint arXiv:2511.10664*.

Wenda Xu, Sweta Agrawal, Vilém Zouhar, Markus Freitag, and Daniel Deutsch. 2025. Deconstructing self-bias in LLM-generated translation benchmarks. *arXiv preprint arXiv:2509.26600*.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A Hedderich. 2025. MAKIEval: A multilingual automatic Wikidata-based framework for cultural awareness evaluation for LLMs. *arXiv preprint arXiv:2505.21693*.

6. Appendix

6.1. Linguistic Competence Task: System Messages

English You are a knowledgeable assistant that can answer all kinds of questions. Please select the correct option. Output ONLY the letter of the correct option, without any additional explanation or punctuation.

Examples:

> What colour is the sky? ['A. blue', 'B. yellow', 'C. green']. Return ONLY 'A'
>
> Which of these countries is in Africa? ['A. Germany', 'B. Mexico', 'C. Nigeria']. Return ONLY 'C'

Gaelic Is e cuidiche fiosrachail a tha annad agus is urrainn dhut a h-uile seòrsa ceist a fhreagairt. Tagh an romhainn cheart. Na cuir a-mach ach litir na freagairte ceirte, as aonais mìneachadh no pongachadh sam bith eile.

Eisimpleirean:

> Dè 'n dath a th' air an iarmailt? ['A. gorm', 'B. buidhe', 'C. uaine']. Na cuir a-mach ach 'A'
>
> Cò 'n dùthaich às na leanas a tha ann an Afraca? ['A. A' Ghearmailt', 'B. Meagsago', 'C. Nigèiria']. Na cuir a-mach ach 'C'

6.2. Transcript Scoring System Message

You are a Scottish Gaelic article scoring assistant. Given a Gaelic article transcript and its English translation, give a score between 1 and 5 for the cultural relevancy of the article to Gaelic culture. If the article contains no material that relates to Gaelic culture, then give a score of 1. If the article is full of information on important aspects of Gaelic culture, then give a score of 5. For articles somewhere in-between these two, then give a score most fitting the content.

6.3. Cultural Understanding QA system message

You are a Scottish Gaelic question and answer generating assistant. Given a Gaelic article transcript and its English translation, write between 1 and 10 question in Gaelic about the content within the article. The questions should test the answerer's knowledge of Gaelic culture in some way, using only the article as the factual basis for the question and answer. The answers to the questions should not be easily guessed from the question. Only include as many questions as you are able to make out of the content of the article. Each question should be written so that it makes total sense in isolation and can be answered by someone knowledgeable on the subject without reading the article. Make sure the questions are self contained. You may introduce people, things, places etc. from the article in each question if that helps make the question understandable without reading the article. Do not refer to entities contextually - always use a persons name rather than using 'the man', for example, where possible. Each question can be as long as you would like but should be answerable in less than 10 words. Write the questions and answers in Gaelic and also write an English translation of each.

6.4. Answerability System Message

You are a Scottish Gaelic and English question scoring assistant. Given a question in Gaelic and its English translation, give a score between 1-5 on the self-contained answerability of both questions. Give a score of 5 if the question is a good general knowledge question, is self-contained, is not contextual dependent, and can be answered purely from using knowledge of Gaelic culture. Give a score of 1 if the question is contextual dependent, refers to implicit information outside of general knowledge (e.g. 'the man' rather than 'Robert the Bruce'), or is otherwise badly written. For questions somewhere in-between these two, then give a score most fitting the content. Evaluate the question in both languages on an individual basis, evaluating the wording of the question purely in that language.

6.5. Distractor System Message

You are a Scottish Gaelic distractor generating assistant. Given a Gaelic article transcript and its English translation, as well as a question and answer about the content within the article, write three distractors for the answer. The distractors should be incorrect answers to the question. The distractors should also be plausible while remaining wrong answers. If the correct answer is written in a particular style or format, make sure the distractors also follow this style or format. Remember that there may be multiple correct answers to the original question, so make sure that your three distractors are all completely INCORRECT answers. Write the distractors in Gaelic and also write an English translation for each.

Argumentation through Discourse Relations and Subjectivity: Introducing FreCaDiS, a French Multi-Genre Corpus

Joanna Blochowiak[1,2], Cristina Grisot[2]

[1]Université catholique de Louvain, [2]Universität Zürich,
Institut Langage et Communication, Louvain-la-Neuve, Belgium; CLARIN-CH, Zürich, Switzerland
{joanna.blochowiak, cristina.grisot}@uzh.ch

Abstract

This paper addresses a crucial yet understudied issue in argumentation studies: the distinction between explanations and justifications, and their interaction with subjectivity. Building on insights from Bex and Walton (2016), who highlight the importance of not conflating explanations with arguments, we propose a corpus-based approach to operationalize this distinction in French. We present FreCaDiS (***Fre**nch Corpus of **Ca**usal Connectives, **Di**scourse Relations, and **S**ubjectivity*), a novel corpus of French texts annotated for explanatory and justificatory discourse relations and their perceived subjectivity. FreCaDiS comprises excerpts of 2–3 sentences drawn from five distinct genres—SMS, online discussions, blogs, press, and contemporary literature—spanning informal to formal registers. Specifically, we focus on sentences introduced by the connectives *parce que* and *car* ("because") and annotate them along two dimensions: (i) discourse relation (explanation vs. justification) and (ii) subjectivity (subjective vs. objective). The corpus was annotated by three independent human annotators using complementary approaches: a holistic, an intuitive method for subjectivity and a guided, operationalized method for discourse relations. FreCaDiS provides a rich resource for the study of argumentation, causal discourse, causal connectives, and subjective interpretation in French and can support future work in computational argument mining, discourse analysis, and NLP applications.

Keywords: explanation, justification, subjectivity, causal connectives, argumentation, human annotations, different text genres, French

1. Introduction

The automatic identification of argument components and structures has become a central topic in NLP. Recent work has explored various architectures and methods for argument structure learning (Wei et al., 2024), domain adaptation and robustness (Ruiz-Dolz et al., 2024), and mining arguments in less conventional domains (Liu et al., 2024). These approaches have advanced the state of the art but often treat arguments as a homogeneous category, typically focusing on claims and premises without distinguishing whether a premise functions as an explanation or a justification.

Research in argumentation mining has proposed different frameworks for automatically discovering argumentative structures in texts. These approaches generally aim to identify all arguments in a document, determine their relations, and represent their overall structure—often as trees or graphs (e.g., Marcu 2000; Mochales Palau & Moens 2009; Buckingham Shum et al. 2003). Such work draws on linguistic, rhetorical, and computational theories to detect propositions that function argumentatively, without assessing their logical validity. However, these broad frameworks tend to overlook the finer-grained linguistic cues that distinguish between argumentative and non-argumentative discourse relations.

By contrast, recent linguistic and discourse-analytic studies highlight the need to align computational models with such distinctions. For example, Ding et al. (2024) link argumentative structure to cohesion in student essays, showing that rhetorical and discourse cues carry important information for argument mining.

Our work builds on this insight by focusing on a specific subset of discourse relations—those introduced by causal connectives, such as the French *parce que* and *car*—and by operationalizing the distinction between explanatory and justificatory uses. Unlike most argumentation mining approaches that seek to reconstruct entire argumentative networks, our approach targets a single, linguistically grounded pattern: discourse relations expressed with causal connectives that are potential indicators of argumentation. Beyond discourse relations, we also annotate subjectivity, indicating whether each discourse segment is perceived by humans as subjective or objective.

This dual focus enables us to address two closely related questions: (i) how to empirically distinguish between arguments and explanations in authentic discourse, and (ii) how subjectivity interacts with these distinctions. To address these questions, we construct a new French corpus, FreCaDiS, annotated with these features across five different text genres.

The rest of the paper is organized as follows. We first discuss the importance of distinguishing between arguments and explanations, with a focus on discourse relations and causal connectives (Section 2). Next, we examine the complexity of the notion of subjectivity and

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 86–94
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

present the rationale for our holistic approach to it (Section 3). We then describe our annotation studies (Section 4). Finally, we outline the main conclusions and contributions of our research (Sections 5 and 6).

2. Arguments, Discourse Relations, and Causal Connectives

The distinction between arguments and explanations has long been recognized as an important issue in the study of discourse, reasoning, and argumentation (van Eemeren et al., 1996; Walton, 2004; Bex and Walton, 2016). As Bex and Walton (2016: 55) rightly observe: "it would be a fundamental error to criticize an argument as falling short of standards for a rational argument, when what was put forward was actually an explanation." Yet, despite its theoretical importance, this distinction remains insufficiently operationalized in NLP tasks: different automatic approaches to arguments detection have different level of granularity (sentence-, two sentences-, paragraph-level, etc.) but most of them tend to treat explanatory and justificatory structures uniformly, often subsuming both under broad categories of argumentative structures (Mochales Palau and Moens, 2011; Lawrence and Reed, 2020; Lippi and Torroni, 2016).

Disentangling explanations and justifications is challenging because they both rely on similar linguistic markers (such as causal connectives), and there are usually few additional cues that can help distinguish them. Yet, their conflation is problematic precisely because one of them—explanation—is not strictly argumentative, whereas the other—justification—typically is. To address this gap, the present study proposes a systematic way to distinguish between explanatory and justificatory uses in French, focusing specifically on clauses introduced by the causal connectives *parce que* and *car* ("because"). While these connectives are generally grouped under the umbrella of causal connectives in linguistic and discourse studies (Sweetser, 1990; Sanders et al., 1992; inter alia), their discourse function can vary and goes well beyond mere cause. Specifically, in explanatory uses, the subordinate clause presents an external cause for an event whereas in justificatory uses, it conveys a reason linked to the speaker's stance, intention, or evaluation, thus aligning more closely with argumentative discourse (Blochowiak et al., 2020). Although it has been argued that *car* tends to favor justificatory uses, whereas *parce que* is more often associated with explanatory ones (e.g., Groupe Lambda-L,1975 and subsequent works in various traditions), both connectives can technically introduce either type of clause. In contemporary French, particularly among younger speakers, their distribution appears to be less strictly specified and more interchangeable (Author1 and Author2 2022).

The examples below illustrate how the same causal connectives, *parce que* and *car*, can introduce clauses that serve these different pragmatic functions in discourse.

(1) a. Le chat est monté sur la table *car/parce qu*'il sentait le poulet.
'The cat climbed onto the table *because* it smelled the chicken.'
b. Le chat peut rester sur la table *car/parce que* c'est son anniversaire.
'The cat can stay on the table *because* it's its birthday.'

(2) a. Le chien aboie *car/parce que* le facteur est là.
'The dog is barking *because* the mailman is here.'
b. Le facteur est là *car/parce que* le chien aboie.
'The mailman is here *because* the dog is barking.'

(3) a. Je suis tombé de la chaise *car/parce qu*'elle s'est cassée.
'I fell off the chair because it broke.'
b. Je suis tombé de la chaise *car/parce que* je voulais faire rire tout le monde.
'I fell off the chair because I wanted to make everyone laugh.'

In (1a), the subordinate clause introduced by the connective provides an external cause—the smell of the chicken—that explains why the cat climbed onto the table. By contrast, in (1b), the connective introduces a justificatory reason: the speaker provides a normative rationale (because it's its birthday), not an actual cause. Similarly, in (2a), the causal connective expresses a straightforward causal relation between the mailman's arrival and the barking of the dog. Interestingly, in (2b) a simple causal interpretation is not possible: the dog's barking does not somehow mysteriously cause the mailman's arrival. Rather, in (2b) the clause introduced by the connective provides a justification for the belief expressed in the first clause—the speaker infers the mailman's arrival from the barking of the dog. Finally, in (3a), the connective introduces a physical explanation (the chair broke), whereas in (3b), it signals an intentional or motivational justification: the speaker fell deliberately to make others laugh.

Even though the sentences above are constructed examples, they reflect genuine interpretative ambiguities in natural discourse. In authentic texts, identifying whether a sentence containing a causal connective is explanatory or justificatory often requires context-sensitive interpretation, making annotation and automatic processing non-trivial. This is precisely why distinguishing between explanation and justification is crucial in argument mining: without such a distinction, systems risk misclassifying non-argumentative explanations as arguments, thereby introducing noise into downstream reasoning, stance detection, and discourse structure modeling (Cabrio and Villata, 2018; Green, 2018).

3. Subjectivity and Its Perception

Another dimension that has received increasing attention is the role of subjectivity in arguments and their perception. Wachsmuth et al. (2024) argue that judgments about argument quality are inherently subjective and context-dependent, especially in the era of instruction-following large language models. Traditional methods for creating such corpora typically focus on detecting subjective lexical items (Wiebe and Riloff, 2005; Riloff and Wiebe, 2003), often derived from predefined lexicons (Das and Sagnika, 2020; Yu and Hatzivassiloglou, 2003; Villena-Román et al., 2015). However, these lexicon-based strategies face well-known limitations, as they tend to depend heavily on domain- and language-specific features (Pang and Lee, 2004).

Several recent resources and annotation efforts have tackled subjectivity more explicitly at the sentence-level, such as the NewsSD-ENG corpus (Antici et al., 2024), which provides subjectivity annotations for English news articles. In their guidelines, they provided annotators with a general definition of subjectivity, according to which: “A sentence is considered subjective when it is based on—or influenced by—personal feelings, tastes, or opinions. Otherwise, the sentence is considered objective.” A series of more specific indications has also been provided, such as: “A sentence is subjective if it explicitly reports the personal opinion of its author.” These definitions perfectly align with the traditional position on the topic of subjectivity, such as the classical one by Benveniste, who sees it as: “the capacity of the speaker to posit himself as subject” (Benveniste, 1966).

However, this type of approach seems to focus on only one aspect of subjectivity, namely by treating subjectivity as equivalent to expressing opinions, while equating objectivity with stating facts. In this view, explanations, which concern facts, are considered inherently objective, whereas justifications, which concern opinions, are seen as subjective. This position is not wrong, but it overlooks an important point: at another level, subjectivity also influences how humans interpret discourse relations more broadly. Specifically, the explanation of a fact can be perceived as more objective or more subjective, while the justification of an opinion can likewise be interpreted as more or less subjective, depending on the choice of words, tone of voice, stance, genre, context, background knowledge, and many other factors. For instance, our example (1a), repeated here as (4a) for convenience, is a simple explanation of a fact that can clearly be considered an objective explanation. However, its slight modification in (4b) does not change the type of discourse relation it expresses—it remains an explanation. What changes is the use of the expressive word *bloody*, which makes it more subjective. In our approach, this would be classified as a subjective explanation.

(4) a. The cat climbed onto the table because it smelled the chicken.
b. The bloody cat climbed onto the table because it smelled the chicken.

Like explanations, justifications can also vary in their perceived degree of subjectivity. Some of them provide a basis for a claim in a more subjective way, some in a more objective way, as in the examples below.

(5) a. The mailman is here because the dog is barking.
b. The mailman is here because this neurotic dog is barking again.

In (5a), the justification relies on an external and observable event: the dog is barking. The speaker uses it to support the inference that the mailman is present. The inferential link between the dog’s barking and the mailman’s arrival is presented in a neutral way, without any additional evaluative language. This is therefore an instance of objective justification. In (5b), the type of discourse relation remains unchanged—the utterance still provides a justification for the conclusion that the mailman is here. What changes is the introduction of evaluative language through the phrase *this neurotic dog* and the attitudinal marker *again*. These elements convey the speaker’s stance and emotional attitude toward the dog, which makes the justification more subjectively framed.

In sum, the key for our study is to approach subjectivity in a more nuanced way—going beyond the simple divide between fact description and opinion expression—and to examine how it intersects with explanatory discourse relations (which are not inherently argumentative) and justificatory discourse relations (which often are), as well as how these relations are perceived in terms of subjectivity.

4. Resource Development

A central contribution of this research is the development of a new corpus FreCaDiS: a French corpus of causal connectives annotated for discourse relations and subjectivity. The FreCaDiS corpus consists of five text genres in French, ranging from informal texts such as SMS and online discussions to more formal ones such as blogs and press, as well as contemporary literature. This variation is relevant because different genres are associated with distinct degrees of subjectivity, discourse conventions, and argumentative strategies, which allows us to observe how explanatory and justificatory relations are realized and perceived across contexts.

Informal genres such as SMS and online discussions are typically rich in personal stance, implicit reasoning, and evaluative language. Blogs, situated between informal and formal registers, often blend personal opinions with factual information, offering hybrid contexts where explanations and justifications can co-occur. Press texts, by contrast, generally favor explicit and structured argumentation on the one hand and are expected to display more objective explanatory and justificatory patterns. Finally, literary texts provide stylistically and narratively rich examples, including non-prototypical expressions of stance. This cross-genre design enables us to investigate how explanation, and justification are distributed across these diverse communicative contexts and how they interact with subjectivity.

4.1 Corpus

The corpus consists of a total of 1081 excerpts (52527 words): 200 from blogs, 220 from contemporary literature, 221 from web discussions, 205 from SMS and 215 from written press. The excerpts were randomly collected through search using “car” and “parce que” as keywords in Sketch Engine (blogs and web discussions), the UCL Corpus (contemporary literature), the Belgian SMS Corpus (Cougnon 2012) and the Le Monde Corpus (year 2012). The excerpts were then randomly selected among all search results in each of the listed sources with the aim to have half of the excerpts with *car* and half with *parce que*.

Each excerpt consists of 1–2 sentences containing either *parce que* or *car*. Each excerpt was annotated along two dimensions: (i) subjectivity (subjective vs. objective) in Annotation Study 1 and (ii) discourse relation (explanation vs. justification) in Annotation Study 2. The annotation was carried out by three human judges, who independently annotated each excerpt for both features. This dual annotation design enables us to examine not only the distribution of explanatory and justificatory relations across different genres but also how subjectivity interacts with these discourse functions.

4.2 Annotation Study 1

In Annotation Study 1, three independent judges were recruited to annotate the corpus excerpts. All annotators were female native speakers of French with a background in linguistics[1] and were compensated for their participation. The study followed a non-guided annotation approach (Author 1 et al., 2020), which is based on the assumption that listeners and readers routinely—and often implicitly—evaluate the degree of subjectivity in the information they process.

Annotators were presented with excerpts containing causal connectives and were asked to decide whether the information conveyed by the sentence appeared more subjective or more objective. Importantly, no explicit definitional criteria for subjectivity or objectivity were provided. This deliberate choice was intended to activate the annotators’ intuitive interpretive layer, allowing for a more natural and holistic assessment of subjectivity.

Annotators were therefore instructed to rely on their overall impression and contextual understanding of each sentence, rather than on predefined diagnostic features. This design makes it possible to capture a gradient and context-sensitive notion of subjectivity that is closer to how it is perceived in everyday language use.

4.3 Annotation Study 2

In Annotation Study 2, we implemented a guided annotation approach to classify sentences containing the causal connectives *parce que* and *car* as instances of either explanation or justification. The same three annotators who participated in Study 1 were recruited for this second task. After a two-week interval, they were re-contacted and introduced to the new annotation procedure. To minimize potential carry-over effects, annotators were asked to delete all materials related to the first study from their devices prior to beginning the new task.

In contrast to the intuitive, non-guided approach used in Annotation Study 1, the Annotation Study 2 provided explicit operational definitions and clear guidelines to support consistent annotation. Annotators were given the following definitions. Explanation answers the question of why or how something happened, or why something is the case. It specifies the cause(s) of an event or state of affairs. Justification addresses why a claim, decision, or action is valid, correct, desirable, or appropriate. It provides reasons, evidence, or arguments to support an assertion, belief, or stance.

Following this presentation, annotators were provided with detailed written guidelines defining

[1] As pointed out by an anonymous reviewer, argument mining is often influenced by the reader’s disciplinary background. This raises the question of whether one would expect different agreement patterns if the task were performed by annotators with different backgrounds. We tested this assumption in a pilot study by recruiting seven annotators via Prolific (2 male, 5 female; native speakers of French; age range 20–51; age at completion of formal education ranging from 19 to 26; with diverse disciplinary backgrounds). We obtained similar results and agreement rates, suggesting that task performance is not substantially affected by individual differences among annotators.

the two discourse relations and were given the opportunity to ask clarifying questions to ensure a shared understanding of the criteria.

Based on these definitions, the annotators' task was to determine, for each sentence containing the causal connective *parce que* or *car*, whether it represented an explanatory or justificatory discourse relation.

4.4 Results

Inter-annotator agreement was measured with Cohen's Kappa, which corrects for agreement that could happen by chance. Its scores range from less than 0 (poor agreement) to 0.8-1 (almost perfect agreement), with intermediary values between 0.01-0.20 as slight agreement, 0.21-0.4 (fair agreement), 0.41-0.6 (moderate agreement), 0.61-0.80 (substantial agreement) (Landis & Koch, 1977). Nevertheless, more recently Grisot (2017) demonstrates using the examples of several types of semantic and pragmatic features and several annotators that inter-annotator agreement scores are strongly influenced by the nature of the linguistic information being annotated: semantic (encoded) information tends to yield higher agreement rates, whereas pragmatic (context-dependent) information often results in lower agreement. Grisot's approach offers a quantitative method for assessing linguistic meaning beyond theoretical intuition.

4.4.1 Discourse Relations

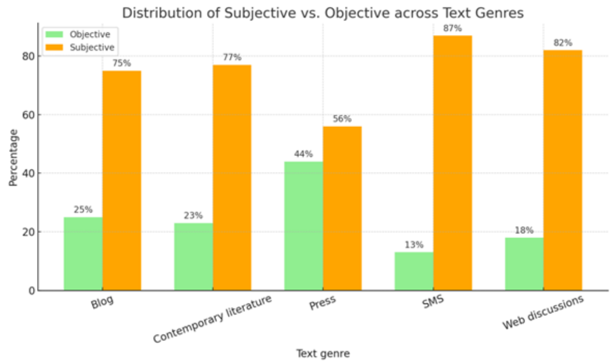
The average agreement among the three annotators was of 70% (i.e., 69% for the first pair of annotators, 66% for the second pair and 75% for the third pair). This average agreement rate corresponds to an average kappa score of 0.27 (i.e., 0.27 for the first pair of annotators, 0.2 for the second pair of annotators and 0.36 for the third pair). These low scores were thus expected due to the pragmatic and context-dependent nature of distinguishing between explanations and justifications, and for this reason, the annotations are accepted as reliable. Furthermore, the categorization of each corpus excerpt as explanation or justification corresponded to the categorization made by the majority (i.e., 2 out of 3, and by 3 out 3 annotators). As such, no corpus excerpt was discarded from further analysis.

The corpus consists of 832 justifications (77%) and 249 explanations (23%). In this set of data, justifications and explanations are distributed as follows in the five corpora:

Figure 1 Frequency of subjective and objective relations in the five corpora

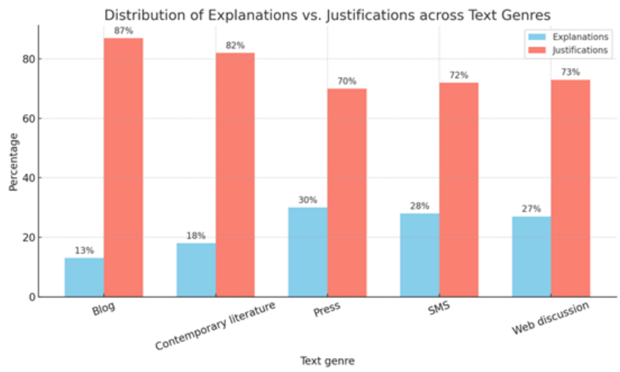
The results show an overall pattern across the five corpora according to which there are less explanations than justifications. Nevertheless, the frequency of explanations is significantly lower in blogs than in press ($\chi^2(4) = 27.83$, $p < .001$).

4.4.2 Subjectivity



The average agreement among the three annotators was of 70% (i.e., 69% for the first pair of annotators, 66% for the second pair and 75% for the third pair). This average agreement rate corresponds to an average kappa score of 0.27 (i.e., 0.27 for the first pair of annotators, 0.2 for the second pair of annotators and 0.36 for the third pair). As before, these low scores were thus expected due to the pragmatic and context-dependent nature of subjectivity. Furthermore, the categorization of each corpus excerpt as subjective or objective corresponded to the categorization made by the majority.

The corpus consists of 814 (75.4%) excerpts perceived as subjective and 266 perceived as objective (24.6%). In this set of data, subjective and objective relations are distributed as follows



in the five corpora:

Figure 2 Frequency of justifications and explanations in the five corpora

The results reveal an overall pattern according to which there are more subjective relations than objective ones in blogs, contemporary literature, SMS and web discussions. In contrast, in press data, the frequency of objective relations is significantly higher than in the other four corpora, whereas in SMS the frequency of objective relations is significantly lower than in the other corpora ($\chi^2(4) = 62.9$, $p < .001$).

4.4.3 At the Intersection of Discourse Relations and Subjectivity

The first step in our analysis is to examine how, in general, discourse relations correlate with the intuitive perception of subjectivity. More specifically, we investigate the question of how explanations and justifications are perceived as more subjective or more objective in each of the five corpora.

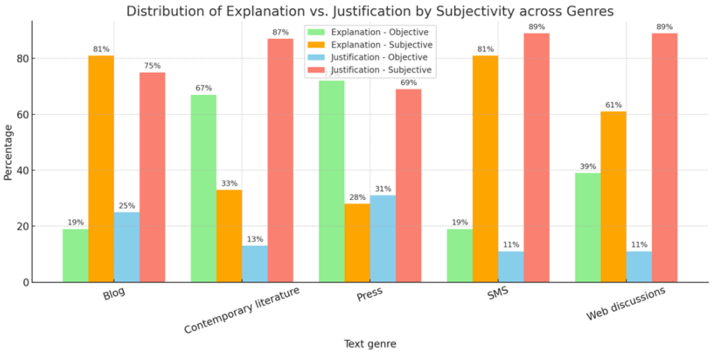


Figure 3 Frequency of explanations and justifications perceived as more subjective or more objective in the five corpora

The distribution of subjective and objective discourse segments reveals a clear predominance of subjectivity across all registers, with particularly high proportions in SMS (87%), web discussions (82%), blogs (75%), and contemporary literature (77%). Even in press texts, where objectivity is expected to play a central role, more than half of the segments (56%) are subjective.

These findings suggest that subjective stance is not limited to overtly personal or informal registers but is a pervasive feature of discourse more generally. The variation across registers reflects their communicative functions: interactive and expressive genres naturally foreground subjective expression, while more informational genres balance subjective and objective content to a greater degree.

5. Conclusions

This paper introduced FreCaDiS, a new French corpus designed to explore the intersection between causal discourse relations, argumentation, and subjectivity across five text genres. By focusing on the causal connectives *parce que* and *car*, we operationalized the theoretical distinction between explanation and justification, and complemented it with human judgments of subjectivity using both guided and non-guided annotation procedures. This dual approach provides an empirically grounded resource for investigating how subjectivity interacts with argumentative structures in natural language.

6. Applications

Beyond its theoretical relevance for the study of discourse and argumentation, FreCaDiS offers a wide range of applications for natural language processing (NLP) and AI research. By combining discourse relations (explanation vs. justification) with explicit subjectivity annotations across five distinct text genres, the corpus provides a valuable resource for advancing computational modeling of argumentation and subjectivity in natural language.

From an applied perspective, FreCaDiS can support several key NLP tasks. In stance detection and subjectivity analysis, its explicit encoding of subjective versus objective discourse enables the training and evaluation of models that more accurately identify subjective statements, opinions, or evaluative language. It also facilitates research on rhetoric and stance-taking, allowing models to capture subtle linguistic cues of perspective and bias. In discourse relation classification, the corpus operationalizes a fine-grained distinction between explanatory and justificatory uses of causal relations—an aspect often neglected in existing resources. This opens new possibilities for language models that distinguish between argumentative and non-argumentative causal relations, improving downstream applications such as text understanding, summarization, and question answering. In future research, this dataset could be used to create a contrastive benchmark to examine whether instruction-following LLMs exhibit patterns similar to those of standard NLP classifiers and human annotators.

This could be done in a manner similar to Escouflaire et al. (2024), who assessed the performance of both humans and LLMs—specifically the state-of-the-art models at the time, GPT-3.5 (Brown et al., 2020) and a French fine-tuned CamemBERT transformer (Martin et al., 2019)—in choosing between *parce que* and *car* after these connectives had been removed from the original excerpts (with minimal context, i.e., a single sentence). In their study, they used a test set of 420 French segments containing either *car* or *parce que*, drawn from two genres: journalistic texts (*Le Monde*) and SMS messages from the SMS4science corpus (Fairon et al., 2006). The target connective was masked and surrounding context was removed. CamemBERT was fine-tuned on 10,000 additional sentences (5,000 per connective) extracted from SMS4science messages and Belgian news articles. Across three connective prediction experiments, human native speakers showed low accuracy and agreement, GPT-3.5 performed inconsistently and generally worse than humans, and the fine-tuned CamemBERT model achieved the best results (66.7% accuracy). Text genre consistently affected the performance of both humans and models.

With respect to FreCaDiS, because it spans five genres ranging from informal (SMS, web discussions) to formal (press, blogs, literature), it can contribute to register- and genre-aware NLP and AI. Specifically, it can enable the development of systems that adapt to different communicative contexts, improving performance in tasks such as genre classification, style transfer, and domain adaptation. While the text genre distinction does not appear to strongly affect human annotators when distinguishing between explanation and justification, an open question is whether this also holds for computational models, or whether text genre-specific models will remain necessary for SSH-related tasks. As we saw, previous work suggests that text genre can indeed be an important differentiating factor in certain settings, such as connective selection task (see Escouflaire et al., 2024).

Finally, FreCaDiS annotations on subjectivity as a binary feature crossed with discourse relations (such as explanations and justifications) make it highly relevant for sentiment and opinion mining, offering a richer basis for modeling nuanced evaluative language beyond simple polarity distinctions. Another methodological approach would be to use a Likert scale for subjectivity in order to provide more granular data, which could better support LLMs in nuanced stance detection tasks. In our study, however, we opted for a coarse-grained annotation scheme (*rather objective* vs. *rather subjective*) for two main reasons. First, our goal was to capture annotators' intuitive and holistic perception of subjectivity rather than to model subtle degrees of stance. We assume that finer distinctions are difficult to apply consistently without extensive guidelines and calibration, which could increase inter-annotator variability and reduce reliability. Second, agreement rates for a multi-point scale would likely have been even lower than for a binary feature, making the additional annotation burden disproportionate to the expected benefit.

7. Limitations

The corpus also presents a number of limitations. Since the task involves inherently subjective judgments, inter-annotator agreement is modest—an expected outcome in pragmatic annotation settings, but one that limits FreCaDiS's immediate suitability for high-stakes NLP benchmarking or fully automated, unsupervised model training. We acknowledge this constraint and view the present study as an initial step toward more robust modeling of subjectivity and discourse-based argumentation. In future work, we plan to strengthen the dataset by incorporating expert annotations, refining annotation guidelines, and introducing a supervising adjudicator to handle ambiguous or contested cases. These measures aim to improve annotation reliability while preserving the richness of intuitive human interpretation that characterizes subjectivity and discourse phenomena.

Furthermore, FreCaDiS is at this stage a research-oriented corpus that lacks a fully validated pipeline for how these findings can be integrated into broader SSH research infrastructures.

8. Acknowledgments

The work was supported by grant no. 34821046 from the Belgian FNRS, awarded to the first author.

9. Bibliographical References

Antici, F., Galassi, A., Ruggeri, F., Korre, K., Muti, A., Bardi, A., et al. (2024, May 24). A Corpus for Sentence-level Subjectivity Detection on English News Articles. arXiv. https://doi.org/10.48550/arXiv.2305.18034

Benveniste, E. (1966). *Problèmes de linguistique générale*. Paris: Gallimard.

Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation, 7(1)*, 55–68. https://doi.org/10.3233/AAC-160001

Blochowiak J., Grisot C., & Degand L. (2020). What type of subjectivity lies behind French causal connectives? A comparative study of French causal connectives car and parce que from a corpus-based perspective. *Glossa: A Journal of General Linguistics*, 5(1), 50. DOI: http://doi.org/10.5334/gjgl.1077

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems,* 33, 1877-1901.

Buckingham Shum, S., Uren, V., Li, G., Domingue, J., & Motta, E. (2003). Visualizing internetworked argumentation. In *Visualizing argumentation: Software tools for collaborative and educational sense-making* (pp. 185-204). London: Springer London.

Cabrio, E., & Villata, S. (2018). Five years of argument mining: A data-driven analysis. In IJCAI (Vol. 18, pp. 5427–5433). https://www.ijcai.org/proceedings/2018/0766.pdf.

Cougnon, L.-A. (2015). *Langage et sms: Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.

Das, N., & Sagnika, S. (2020). A Subjectivity Detection-Based Approach to Sentiment Analysis. In D. Swain, P. K. Pattnaik, & P. K. Gupta (Eds.), *Machine Learning and Information Processing* (Vol. 1101, pp. 149–160). Singapore: Springer Singapore.

https://doi.org/10.1007/978-981-15-1884-3_14
Ding, Y., Kashefi, O., Somasundaran, S. & Horbach, A. (2024) When Argumentation Meets Cohesion: Enhancing Automatic Feedback in Student Writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17513-17524, Turin, Italy. ELRA and ICCL.
Eemeren, F. H. van, & Grootendorst, R. (2016). *Argumentation, Communication, and Fallacies: A Pragma-dialectical Perspective.* New York: Routledge.
Escouflaire L., Blochowiak J., Degand L., & de Marneffe, M-C. (2024). “Which connective fits best: ‘car’ or ‘parce que’? A challenge for both humans and LLMs.” In: *JADT 2024 Proceedings: 17th International Conference on Statistical Analysis of Textual Data*, JADT: Brussels 2024 http://hdl.handle.net/2078.1/288770
Fairon, C., Klein, J., Paumier, S. (2006). Le corpus SMS pour la science. *Base de données de 30.000 SMS et logiciels de consultation.* Louvain-la-Neuve, Presses universitaires de Louvain.
Green, A. (2018). *Creativity in the distance: The neurocognition of semantically distant relational thinking and reasoning.* The Cambridge handbook of the neuroscience of creativity, 25, 363–381.
Grisot, C. (2017). A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement. *Journal of Pragmatics*, 117, 245–263.
Lambda-l, G. (1975). Car, parce que, puisque. *Revue romane,* 10(2), 248–280.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.
Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational linguistics*, 45(4), 765–818.
Lippi, M., & Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. ACM *Transactions on Internet Technology*, 16(2), 1–25. https://doi.org/10.1145/2850417
Liu, H., Xu, W., Lin, X., Huo, J., Chen, H., & Zhao, Z. (2024, May 30). AntCritic: Argument Mining for Free-Form and Visually-Rich Financial Comments.
arXiv. https://doi.org/10.48550/arXiv.2208.09612
Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. & Sagot, B. (2019). CamemBERT: a tasty French language model. *arXiv:1911.03894*.
Mochales, R., & Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1), 1–22.
https://doi.org/10.1007/s10506-010-9104-x
Pang, B., & Lee, L. (2004, September 29). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.
arXiv. https://doi.org/10.48550/arXiv.cs/0409058
Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 105–112).
Román, J. V., Cámara, E. M., Morera, J. G., & Zafra, S. M. J. (2015). Tass 2014-the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54, 61–68.
Ruiz-Dolz, R., Chen, C. C., Kando, N., & Chen, H. H. (2024). Learning strategies for robust argument mining: An analysis of variations in language and domain. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024* (pp. 10286–10292). European Language Resources Association (ELRA).
Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1), 1–35.
https://doi.org/10.1080/01638539209544800
Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Vol. 54). Cambridge University Press.
Van Eemeren, F. H., Grootendorst, R., Jackson, S., & Jacobs, S. (1997). Argumentation. In *Discourse as structure and process: Discourse studies: A multidisciplinary introduction* (pp. 208–229). SAGE Publications Ltd.
Wachsmuth, H., Lapesa, G., Cabrio, E., Lauscher, A., Park, J., Vecchi, E. M., et al. (2024, March 24). Argument Quality Assessment in the Age of Instruction-Following Large Language Models. arXiv.
https://doi.org/10.48550/arXiv.2403.16084
Walton, D. (2004). A new dialectical theory of explanation. *Philosophical Explorations*, 7(1), 71–89.
https://doi.org/10.1080/138697903200018686 3
Wei, S., Li, M., Sun, J., Davis J. & Moens, M.-F. (2024). DMON: A Simple Yet Effective Approach for Argument Structure Learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* pp. 5109-5118, Turin, Italy. ELRA and ICCL.
Wiebe, J., & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 3406, pp. 486–497). Berlin, Heidelberg: Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-540-30586-6_53
Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 129–136).

Text-only Domain Adaptation for Low-Resource ASR Using Large Language Models

William Lamb, Dongge Han, Ondřej Klejch, Peter Bell
University of Edinburgh, Microsoft Corporation UK, University of Edinburgh, University of Edinburgh
{w.lamb, o.klejch, p.bell}@ed.ac.uk, dongge.han@microsoft.com

Abstract

Automatic Speech Recognition (ASR) increasingly mediates access to broadcast media, public discourse and cultural archives. For minoritised languages, however, the development of robust ASR systems is constrained by limited and domain-restricted text data. This paper investigates cross-lingual text expansion (XLTE), a method that uses a Large Language Model (LLM) to generate in-domain text in a low-resource language from high-resource language summaries. We further examine whether supervised fine-tuning on a small set of human-authored texts enhances generation quality. Using Scottish Gaelic as a case study, we show that synthetic text generated via fine-tuned XLTE can be used to train an external language model that reduces Word Error Rate (WER) by 24.48% in a previously unseen broadcast domain. Our findings demonstrate that text-only domain adaptation through cross-lingual generation can strengthen speech technology in sparse data settings. Beyond engineering gains, the approach offers a scalable pathway for improving the digital representation, accessibility and sustainability of minoritised-language media and cultural heritage.

Keywords: speech recognition, domain adaptation, low-resource languages, Scottish Gaelic, large language models

1. Introduction

A central challenge in improving linguistic diversity in modern automatic speech recognition (ASR) systems is data sparsity. Although contemporary end-to-end ASR architectures jointly model acoustic and linguistic information, a strong external language model (LM) remains important in low-resource settings, where it can partially compensate for weaker acoustic representations (Wallington et al., 2021). For many minoritised languages, however, the difficulty lies not only in the limited quantity of available text but also in its narrow domain coverage.

Scottish Gaelic illustrates this problem clearly. The language is spoken by $69{,}700$ individuals in Scotland (1.3% of the population) (National Records of Scotland, 2022). Publicly available Gaelic text is dominated by BBC regional radio news scripts and is heavily skewed towards local reporting. When such data are used to train an LM for related but distinct domains, for example televised national or international news, performance deteriorates because of domain mismatch. This is not merely a technical inconvenience; it constrains the development of speech technologies that enable access to contemporary media, public discourse and cultural archives in minoritised languages.

Addressing domain mismatch is therefore central to broader efforts within the Social Sciences and Humanities (SSH) to build sustainable and equitable language technology infrastructures. ASR systems increasingly mediate access to minority-language media and oral heritage, yet creating sufficiently broad and representative training corpora remains prohibitively expensive for most low-resource languages.[1]

Ideally, an external LM for ASR would be trained on large volumes of in-domain transcribed text. In practice, such data rarely exist and using multilingual large language models (LLMs) presents one alternative to finding or creating genuine supervised speech data. If a target language appears in an LLM's pre-training data, it may be prompted to synthesise new text in that language using zero-shot or few-shot methods. Previous work has shown that fine-tuning can further improve generation quality when in-domain examples are available (Bell et al., 2021). However, in genuinely low-resource settings, even small quantities of domain-specific data may be unavailable.

Recent studies have explored synthetic data augmentation for low-resource languages (Lucas et al., 2024; Samuel et al., 2024; Alcoba Inciarte et al., 2024). Yet LLM outputs for such languages often exhibit structural interference from dominant high-resource languages, typically English (Robinson et al., 2023; Lai et al., 2023). This may reflect English-centric representational biases in multilingual models (Papadimitriou et al., 2023; Wendler et al., 2024). For ASR domain adaptation, however, perfectly fluent output is not required. Theoretically, improvements could come from synthetic text simply increasing coverage of n-grams present in evaluation data.

[1] Crowdsourcing is one option. See 'Opening the Well', a community-driven and ASR-assisted Gaelic folklore transcription initiative: `https://fosgladh.tobaranducalchais.co.uk/en`.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 95–102
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

In this study we simulate a scenario where limited LM training data exist for a low-resource language and propose a solution based on **cross-lingual text expansion (XLTE)**. XLTE generates in-domain text in a low-resource target language by prompting an LLM with summaries in a high-resource source language. The method combines cross-lingual transfer with controlled expansion: a short source-language summary is transformed into a longer target-language text aligned with the desired domain.

Beyond zero-shot XLTE, we introduce a supervised fine-tuning approach in which the LLM is trained to reconstruct original low-resource texts from summaries written in a high-resource language. Unlike prior work that improves n-gram LMs through, for example, transformer-based rescoring (Wang et al., 2019), our approach synthesises entirely new in-domain text in the target low-resource language.

For practical reasons, we fine-tune OpenAI's GPT-4o model,[2] adapting it to map English summaries of Gaelic regional news texts to their original Gaelic counterparts. We then prompt the fine-tuned model with wide-domain English summaries from the CNN news dataset (Hermann et al., 2015) to generate international news texts in Gaelic. These synthetic texts are then used to train an n-gram-based external LM, which is evaluated in a downstream ASR task involving the Gaelic television news programme *An Là* ('The Day').

Our central objective is to inject world knowledge encoded in a high-resource language, English, into a low-resource Gaelic training corpus through controlled cross-lingual generation. We evaluate zero-shot XLTE versus supervised fine-tuning, XLTE versus machine translation – as a data augmentation strategy – and the impact of synthetic data on downstream ASR performance. Building on our earlier introduction of XLTE (Lamb et al., 2025), which focused on intrinsic evaluation of synthetic narrative data, this paper shifts the emphasis to domain adaptation and extrinsic validation, demonstrating measurable reductions in WER in a low-resource ASR task.

Summary of contributions:

1. We evaluate an LLM-based cross-lingual data augmentation strategy for domain adaptation in low-resource ASR.
2. We demonstrate that supervised fine-tuning improves synthetic text quality over a baseline LLM across intrinsic evaluation metrics.
3. We achieve substantial reductions in Word Error Rate (WER) for a previously unseen domain in a real-world Gaelic ASR task.

[2]Model version: gpt-4o-2024-05-13.

2. Methodology

2.1. Text generation

XLTE requires aligned pairs of English summaries and Gaelic texts. To construct these, we prompted the baseline GPT-4o model to produce English summaries of the original Gaelic news articles (see Table 3) using the instruction: *Your role is to summarise the given news story in 3 to 4 sentences in English.* These summaries serve as the source-language stimuli for subsequent expansion. The remaining stages of the pipeline are illustrated in Figure 1 and detailed below.

We then tested whether supervised fine-tuning improves performance relative to the baseline GPT-4o model. For fine-tuning, the model was trained with the instruction: *You will receive a summary of a news story in English. Expand the summary into a much longer news story in Scottish Gaelic.* Prompting in Gaelic led to reduced performance and, therefore, was not pursued further.

At generation time, we expanded English in-domain summaries into corresponding Gaelic texts. Depending on the experimental condition, we generated between 100 and 2,701 synthetic news texts.

2.2. Evaluation metrics

We employ intrinsic metrics as proxies for downstream ASR performance: Mean Word Count (MWC), English-to-Gaelic ratio (en:gd), Neologism Ratio (Neo), Perplexity (PPL) and Self-BLEU (SB). Lower values are preferred for all metrics except mean word count. Each metric captures a different property of the generated text.

Mean Word Count (MWC) measures average volubility per generated sample. Higher MWC implies that fewer API calls are required to reach a target corpus size, thereby reducing generation cost.

The English-to-Gaelic ratio (en:gd) estimates language uniformity. It is computed by dividing the number of tokens matching a large English dictionary by those matching a large Gaelic dictionary. Tokens absent from both dictionaries are classified as neologisms, which include hallucinated forms and other out-of-vocabulary items. The Neologism Ratio (Neo) reports the proportion of such tokens.

Perplexity (PPL) measures the predictive fit between a language model and a text sample. Although widely used as a proxy for language model quality, it correlates imperfectly with human judgements (Stureborg et al., 2024) and is sensitive to punctuation (Wang et al., 2023) and length effects (Meister and Cotterell, 2021). For consistency, we lowercase and normalise all texts, to-

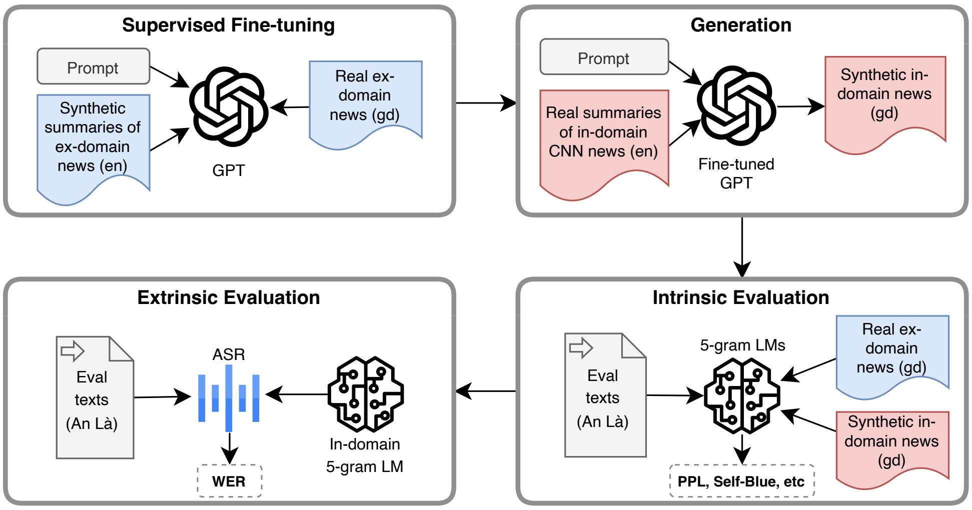


Figure 1: Training and evaluation pipeline (en 'English'; gd 'Gaelic'; PPL 'perplexity')

kenise using a byte-pair encoding model (Sennrich et al., 2016) trained on a dataset of Gaelic news scripts, and compute perplexity with 5-gram language models employing modified Kneser-Ney smoothing (KenLM (Heafield et al., 2013)).

Human-authored texts typically show greater lexical diversity than synthetic texts (Yu et al., 2024). To assess our texts' lexical diversity, we compute Self-BLEU (SB), which measures the average n-gram overlap between each sentence in a text and all others (Zhu et al., 2018). Lower SB indicates greater diversity. For a set of sentences $\{s_1, s_2, \ldots, s_m\}$, Self-BLEU is defined as:

$$\mathrm{SB} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{BLEU}\big(s_i, \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m\}\big) \quad (1)$$

For extrinsic evaluation, we use the best-performing fine-tuned GPT-4o model to generate a larger synthetic corpus. An n-gram LM is trained on this material and integrated into a low-resource ASR system. We then compare Word Error Rate (WER) against a baseline LM trained solely on the available training data (Tr). The key question is whether incorporating synthetic text yields measurable WER reduction.

3. Datasets

With the permission of BBC ALBA, we scraped our out-of-domain data from texts posted on the BBC's *Naidheachdan* 'News' website.[3] These originated as scripts for news programmes on the *Radio nan Gàidheal* radio station. Although the original radio programmes cover regional, national and international stories, the *Naidheachdan* web texts are heavily skewed towards regional reporting. We scraped approximately 37,000 bulletins (c. 6.5M words) spanning January 2012 to October 2023. After de-duplication, we split the corpus into three disjoint RnG subsets: a training set (Tr) and validation set (Val) for fine-tuning, and a generation set (Gen) used to create English summary stimuli using GPT-4o (see Section 2). We treat the BBC RnG scripts as the source domain and BBC ALBA's *An Là* as the target domain.

Our evaluation set (Eval) was drawn from BBC ALBA's *An Là* ('The Day') news programme[4] and comprises verbatim transcripts of six 30-minute episodes with aligned audio. We use Eval both for intrinsic evaluation (perplexity) and for extrinsic evaluation, where we report Word Error Rate (WER) on the corresponding audio.

The raw CNN data were from the training subset of the `cnn_dailymail` dataset (Hermann et al., 2015), available at `HuggingFace`.[5] The CNN news articles (all English-medium) originally appeared between April 2007 and April 2015. We deployed two separate generation subsets from these data: Gen1, comprising 1000 GPT-4o-generated summaries of CNN articles used during intrinsic evaluation; and Gen2, a larger, disjoint 2701 article dataset, which we used during extrinsic evaluation. Because Gen2 is used only as a stimulus source for generation, leveraging the dataset highlights rather than model-generated summaries does not affect the downstream comparison. Table 3 summarises these datasets.

[3] https://www.bbc.co.uk/naidheachdan

[4] https://www.bbc.co.uk/programmes/b00drynf

[5] https://huggingface.co/datasets/abisee/cnn_dailymail

Source	Lang	Subset	N	Words (gd)
RnG	gd	Tr	1000	240,866
RnG	gd	Val	200	40,729
RnG	gd	Gen	2000	444,953
An Là	gd	Eval	7	24,725
CNN	en	Gen1	1000	n/a
CNN	en	Gen2	2701	n/a

Table 1: Datasets (Key: RnG 'Radio nan Gàidheal'; Lang 'original language'; gd 'Gaelic'; en 'English'; Tr = fine-tuning training set; Val = validation set; Gen = summary stimuli; Eval = ASR evaluation set; N = number of documents.)

Input	PPL	SB	MWC	en:gd	Neo
Real (RnG)	**317.3**	0.41	226.1	0.02	0.03
Gen (RnG)	422.6	0.47	381.0	0.02	0.02
Gen (CNN)	494.5	**0.40**	377.7	0.03	0.04

Table 2: GPT-4o baseline: Intrinsic evaluation metrics for generated texts derived from RnG-Gen and CNN-Gen1 summaries, and ground-truth texts from the RnG-Gen set ($N = 400$). Key: PPL = perplexity; SB = Self-BLEU; MWC = mean word count; en:gd = English-to-Gaelic token ratio; Neo = proportion of out-of-dictionary tokens.

4. Experimental Results

4.1. Establishing Baseline Performance

We began by summarising the texts in English using the baseline GPT-4o model to produce the required summary-text pairs.[6] Then, we generated 400 texts using the RnG-Gen and CNN-Gen1 summaries and compared the synthesised texts to real ones from the RnG-Gen set. As seen in Table 2, when using the GPT-4o-base model, LMs built from the synthetic data obtained higher PPL values against the Eval set than an LM built from the real data. So, the real data produce an LM with a better fit to the target domain. Although past research has found real text to be more diverse than synthetic text (Yu et al., 2024), texts generated from CNN-Gen1 show a slightly lower Self-BLEU score than the real data and those generated from RnG-Gen. This textual diversity could be advantageous to the downstream ASR task. Yet, the English to Gaelic and Neologism ratios are higher than the real and generated RnG texts. The international focus of the CNN summaries may dispose the LLM to output more English and out-of-dictionary tokens.

[6]Summarisation hyperparameter settings: n=1, τ=1, top-p=0.85, frequency penalty=0, presence penalty=0.2 and max tokens=250. The maximum token count translates to roughly 3 to 4 sentences in English.

Model	PPL	SB	MWC	en:gd	Neo
FT100	**283.8**	**0.25**	457.5	0.05	0.09
FT200	313.4	0.25	454.9	0.06	0.09
GPT-4o	494.5	0.40	379.2	0.03	0.05

Table 3: Fine-tuning experiments: Intrinsic evaluation metrics for fine-tuned models (FT100, FT200) and the GPT-4o baseline.

4.2. Effect of Supervised Fine-tuning

To investigate whether supervised fine-tuning improves text synthesis quality, we fine-tuned GPT-4o on 100, 200 and 400 summary-text pairs from the RnG training set. We refer to these models as FT100, FT200 and FT400. After a short hyper-parameter study, we settled on fine-tuning for 3 epochs with a batch size of 1 and a learning rate multiplier of 2. Loss trends indicated effective early-stage learning with no signs of over-fitting.

We generated 400 texts from the CNN-Gen1 set using the GPT-4o-base, FT100 and FT200 models.[7] Table 3 compares them using the intrinsic evaluation metrics. The LM associated with FT100 had the lowest PPL value against the Eval set. This suggests that supervised fine-tuning boosts performance for this task over generating from the baseline GPT model. The FT100-based LM is also a better fit to the Eval set and more lexically diverse than an LM built from out-of-domain real data (cf. 'Real: RnG' in Table 2). Notably, given that the FT100 model's LM produces a lower PPL value than the FT200 model's LM, the ideal number of fine-tuning examples appears to be below 200 for these data and this use-case.

In general, Table 3 shows that, compared to the baseline model, the fine-tuned models produce more diverse text, a higher MWC, a higher proportion of English text and a higher proportion of neologisms. While the greater diversity and higher MWC are likely favourable for our task, the inflated English and neologism could be detrimental. Impressionistically, the baseline model produces text that is more coherent, but also more generic. The FT models, in comparison, are stylistically closer to the real news-scripts.

4.3. Translation versus Generation

To consider whether machine translation (MT) yields superior results to GPT-4o-based genera-

[7]We began using a top-p of 0.85, but reduced it to 0.7 after further testing. The other generation hyperparameters were: n=1, τ=1, frequency penalty=0.2; presence penalty=0.5 and max tokens=1000. Experimentation with the baseline GPT-4o model suggested that these settings provided good diversity, less repetition and the longer outputs required for our task.

Source	PPL	SB	MWC	en:gd	Neo
Gen	**381.9**	**0.43**	419.2	0.06	0.06
MT	505.1	0.45	640.7	0.05	0.05

Table 4: Generated vs. machine-translated text: Intrinsic evaluation metrics.

tion, we expanded 800 summaries from the CNN-Gen1 set using our FT100 model. In tandem, we translated English texts from the same dataset to Gaelic using Google Translate's API. After building n-gram language models from the generated and machine-translated text, we calculated the intrinsic evaluation metrics. To ensure fair comparisons between the MT and Generated (Gen) datasets, we controlled for word count.

Table 4 shows that an LM built from FT100-generated text (Gen) achieves lower perplexity on the Eval set than one built from the MT text. The lower Self-BLEU score indicates that the Gen text is also slightly more diverse than the MT text, despite its lower MWC. The stronger supervision signal associated with MT may explain the marginally lower en:gd and neologism ratios for the MT text. In sum – for this task, these models and these languages – XLTE produces higher quality and more diverse text than a state-of-the-art MT system, but shows marginally inflated English and neologism ratios.

Ultimately, improvements in intrinsic metrics are only meaningful if they translate into downstream ASR gains. We therefore evaluate whether synthetic data reduce WER under realistic decoding conditions.

4.4. Extrinsic Evaluation

We evaluated our approach in a real-world setting by generating a large volume of text with FT100, incorporating it into an external LM and using it for an ASR task within the target domain. The goal was to determine whether XLTE-derived text enhances WER, which would suggest effective domain adaptation.

The ASR system's acoustic model was our top-line Gaelic model at the time the research was carried out (details of architecture in Klejch et al., 2025[8]). Acoustic features were extracted from the 18$^{\text{th}}$ layer of an XLS-R 300M model (Babu et al., 2022), which underwent continual pre-training on Gaelic and English. The English text came from the MGB-1 corpus (Bell et al., 2015). The acoustic model employed a TDNN-F architecture with 1,000 BPE units and shared a tokeniser with the language model. Evaluation was conducted on the Eval set (175 mins of aligned acoustic and textual data) described in §3.

Dataset	Train	Gen	Train+Gen	Top-line
gd	29.37	22.86	22.18	–
gd+en	21.01	19.30	**18.86**	14.81

Table 5: WER (%) for the ASR task using LMs built from real data (Train), synthetic data (Gen), and their combination (Train+Gen). Interpolation with English data is shown in the gd+en row.

Table 5 reports WER results for external LMs trained on three datasets: real data from Tr ('Train': 60,994 words), synthetic data generated from 2,701 Gen2 summaries using FT100 ('Gen': 1,038,793 words) and the concatenation of these two datasets ('Train+Gen': 1,099,787 words).

To simulate a low-resource setting, the real data were restricted to the 100 texts used to fine-tune FT100. To better model code-switching, which is common in spoken Gaelic (Smith-Christmas, 2012), we also trained variants that incorporated English data from the MGB-1 corpus (Bell et al., 2015).

For reference, Table 5 includes the WER of our current production-grade external LM ('Top-line') evaluated on the same test set. This model is trained on the full available Gaelic corpus and therefore represents an upper-bound benchmark rather than a system operating under the same low-resource constraints.

Using an LM built from 100 real Gaelic texts yields a WER of 29.37%, which drops to 21.01% when interpolating with English data. Using LMs built from the synthetic data further lowers the WER (22.86% for Gaelic only, 19.3% for Gaelic+English), validating our approach. Fine-tuning GPT-4o on just 100 summary-text pairs synthesises an effective, in-domain corpus over an order of magnitude larger than the original, achieving a 22.17% relative WER reduction. Combining the real and generated data brings the reduction to 24.48% ($29.37\%\to22.18\%$) and interpolating the English data increases the relative WER reduction to 35.78% ($29.37\%\to18.86\%$). The WER of 18.86% is only 4 percentage points away from the top-line on this task.

Typical of modern Gaelic speech, our Eval set contains many English words: roughly 17% of the total. Notably, our synthetic data have a higher en:gd ratio than the training data (0.8 vs 0.34). This, along with the fact that deletions account for the biggest reduction in WER between the training and generated data, led us to investigate whether our approach mainly enhances English speech recognition.

Table 6 shows that 66% of the total WER reduction (514 of 784 errors) can be attributed to improve-

[8] Between the time that the research was conducted and the publication of Klejch et al., 2025, the WER on the *An Là* testset dropped from 14.81% to 10.4%.

Language	Subs	Ins	Dels	Sum	% Total
English	-282	49	-281	-514	66%
Gaelic	-214	-11	-45	-270	34%

Table 6: Change in error counts after replacing the external LM trained on real data with one incorporating synthetic data. Negative values indicate reductions in errors. The final column shows each language's proportion of the total reduction in errors ($n = 784$).

ments with English tokens, with 55% of these gains arising from fewer deletions. Although English insertions increase slightly ($n = 49$), the net error reduction remains strongly positive. Indeed, 34% of the total error reduction (270 errors) relates to Gaelic tokens, confirming that the synthetic data improve recognition of both languages. While gains for English are more pronounced, the reduction in Gaelic errors provides evidence that XLTE enhances in-domain language modelling; it does not merely boost English coverage. This is encouraging for future efforts applying XLTE to low-resource ASR settings (cf. Joshi and Singh, 2022).

5. Conclusions

This study demonstrates that fine-tuning an LLM on a modest set of human-authored texts can generate a substantially larger in-domain corpus for a low-resource language. Intrinsic evaluation shows that cross-lingual text expansion (XLTE) produces material better aligned with the target domain than machine translation, yielding lower perplexity and greater internal diversity. When used to train an external language model, the synthetic corpus delivers substantial downstream gains, reducing WER by up to 24.48% under low-resource conditions and 35.78% when interpolated with English data.

These results establish synthetic text generation as an effective mechanism for domain adaptation in low-resource ASR. Improvements are observed for both Gaelic and English tokens in a bilingual setting, indicating enhanced in-domain coverage rather than inflated majority-language recognition. The method is computationally lightweight, scalable and does not require additional parallel data, making it readily transferable to other minoritised languages and under-represented domains.

The study also highlights the continuing value of human-generated texts (Bird and Yibarbuk, 2024). Even relatively small, carefully curated human-produced datasets provide the structural and stylistic signal necessary for effective fine-tuning. In this sense, community-authored material functions as critical digital infrastructure for low-resource language technology.

Several limitations remain. Fine-tuning was conducted via a proprietary API, which constrains transparency. The growing availability of open-weight models makes replication with them feasible, although recent work suggests that open-weight models are generally less performative for Gaelic than leading proprietary ones (Devine et al., 2026). The extrinsic evaluation assumes a strong acoustic model; future work should examine stricter low-resource acoustic conditions. Finally, validation was limited to a single broadcast domain, and broader genre coverage will be required to assess generalisability.

Beyond engineering gains, the findings carry implications for the Social Sciences and Humanities. Speech recognition increasingly provides a conduit for searching and accessing broadcast media, folklore and oral history archives. For minoritised languages, domain mismatch in language modelling affects not only accuracy but cultural visibility. Strengthening domain-specific language models is therefore a prerequisite for equitable digital representation and sustainable access to linguistic heritage. In the Gaelic context, initiatives such as *Opening the Well* illustrate how advances in language technology can enhance the digital presence of cultural materials, while community engagement in turn strengthens the technological ecosystem, creating a mutually reinforcing cycle.

Statement on Ethics

Institutional ethical review was initiated on 2 March 2023 and approved on 13 March 2023 by the Ethics Officer of the host institution. The study involved no human participants and was assessed as presenting minimal risk. The authors nonetheless recognise the environmental implications associated with training and deploying large language models. While the primary computational costs arise during large-scale pre-training, fine-tuning and text generation also entail energy consumption. Because this study relied on a proprietary model accessed via API, precise estimates of compute usage and associated carbon emissions are not publicly available.

Code and Data Availability

All code and synthetic data generated by the fine-tuned GPT-4o model is available at `https://github.com/razorfish17/XLTE`. The BBC *Naidheachdan* corpus cannot be redistributed due to third-party licensing restrictions, but the scraping methodology, including the date range, URL structure and data selection criteria, is described in §3 to facilitate replication.

Acknowledgements

This work has benefited from the support of the Scottish Government (Grant name: 'Ecosystem for Interactive Speech Technologies'). Thanks to the anonymous peer reviewer for their helpful comments. Thanks also to Mr Reamonn Lenkas at BBC ALBA for his assistance gathering the *naidheachdan* data and Mr Cailean Gordon for transcribing the *An Là* episodes.

References

Alcides Alcoba Inciarte, Sang Yun Kwon, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2024. On the utility of pretraining language models on synthetic data. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 265–282, Bangkok, Thailand. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech*.

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2021. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693. IEEE.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Peter Devine, William Lamb, Beatrice Alex, Ignatius Ezeani, Dawn Knight, Mícheál J. Ó Meachair, Paul Rayson, and Martin Wynne. 2026. GaelEval: Benchmarking LLM performance for Scottish Gaelic. In *Proceedings of LLMs4SSH*, Palma de Mallorca.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Raviraj Joshi and Anupam Singh. 2022. A simple baseline for domain adaptation in end to end ASR systems using synthetic data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 244–249.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic. In *Interspeech 2025*, pages 728–732.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

William Lamb, Dongge Han, Ondrej Klejch, Beatrice Alex, and Peter Bell. 2025. Synthesising a corpus of Gaelic traditional narrative with cross-lingual text expansion. In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 12–26, Abu Dhabi [Virtual Workshop]. International Committee on Computational Linguistics.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation. 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 6385–6397.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. 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 5328–5339, Online. Association for Computational Linguistics.

National Records of Scotland. 2022. Scotland's census 2022. `https://www.`

scotlandscensus.gov.uk. Accessed: 2024-08-14.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs Augment Low-Resource Reading Comprehension Datasets? Opportunities and Challenges.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Cassie Smith-Christmas. 2012. *I've lost it here dè a bh'agam: Language shift, maintenance, and code-switching in a bilingual family*. Ph.D. thesis, University of Glasgow.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Electra Wallington, Benji Kershenbaum, Peter Bell, and Ondřej Klejch. 2021. On the learning dynamics of semi-supervised training for ASR. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, pages 716–720. International Speech Communication Association.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2023. Perplexity from PLM is unreliable for evaluating text quality.

Yiren Wang, Hongzhao Huang, Zhe Liu, Yutong Pang, Yongqiang Wang, ChengXiang Zhai, and Fuchun Peng. 2019. Improving n-gram language models with pre-trained deep transformer.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Benchmarking LLMs for Aspect-Based Sentiment Classification in Slovene Historical Periodicals

Tina Munda[2], **Filip Dobranić**[1], **Uroš Šmajdek**[3], **Oliver Pejić**[1], **Ciril Bohak**[3], **Vojko Gorjanc**[1,2], **Darja Fišer**[1]

[1] Institute of Contemporary History
Privoz 11, SI-1000 Ljubljana
{filip.dobranic, oliver.pejic, darja.fiser}@inz.si

[2] Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
{tina.munda, vojko.gorjanc}@ff-uni-lj.si

[3] Faculty of Computer Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
{uros.smajdek, ciril.bohak}@fri-uni-lj.si

Abstract

Historical newspapers present substantial challenges for computational sentiment analysis due to OCR errors, archaic linguistic features, and the absence of domain-specific labelled training data. This paper investigates whether instruction-following LLMs can facilitate aspect-level sentiment inference under such conditions. We benchmark four instruction-following LLMs on a manually annotated sample of collective-identity mentions drawn from Slovene historical newspapers. The results provide a benchmark for targeted sentiment classification in OCR-degraded historical Slovene and offer an empirically grounded assessment of the capabilities and limitations of an instruction-tuned LLM in digital humanities research.

Keywords: historical newspapers, digital humanities, aspect-level sentiment analysis, ALSA, ABSA, LLMs, benchmark, GaMS, sPeriodika

1. Introduction

Computational approaches have long been central to Digital Humanities (DH), enabling large-scale analyses of textual corpora through rule-based methods (e.g., lexicon-driven sentiment analysis), unsupervised modelling techniques such as topic modelling, and supervised machine-learning approaches. While these approaches have expanded the analytical scope of DH research, many rely on curated linguistic resources or supervised training data, while others require extensive pre-processing and normalisation. Such requirements pose substantial challenges for analysing historically noisy corpora, where annotated datasets and normalisation tools are limited.

The emergence of instruction-following large language models (LLMs) has fundamentally altered this landscape. Unlike traditional supervised approaches that require substantial amounts of task-specific annotated data, instruction-tuned LLMs can perform complex linguistic tasks in zero-shot settings, including sentiment inference, potentially lowering the barrier for computational analysis in DH contexts. This shift is particularly relevant for historical Slovene, where labelled sentiment datasets are scarce and OCR degradation further complicates large-scale annotation and model training. In such contexts, building and maintaining domain-specific training resources is costly and often infeasible.

At the same time, the robustness of instruction-following LLMs under historically degraded conditions remains insufficiently examined. It is unclear how instruction-following models behave in OCR-extracted text material, in morphologically rich languages, and in tasks that require fine-grained attribution of sentiment to specific lexical targets rather than to entire sentences or documents. Moreover, the extent to which language-adapted models outperform widely used general-purpose instruction-tuned LLMs in such settings has not been systematically evaluated for Slovene historical data.

This study addresses these questions by benchmarking instruction-following LLMs on targeted sentiment classification in late 19th–early 20th century Slovene newspapers from the *sPeriodika* corpus (Dobranić et al., 2023). The task involves classifying sentiment toward explicitly marked collective identity mentions—both nominal and adjectival realisations—in OCR-extracted historical text.

We evaluate GaMS3-12B-Instruct, a publicly available instruction-tuned LLM based on the Gemma 3 family and continually pretrained with a focus on Slovene-language data (Vreš et al., 2024), and compare it with other publicly available general-purpose instruction-tuned models of comparable scale: Gemma-3-12B-IT (Gemma Team, 2025),

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 103–113
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

LLaMA 3.1 (Grattafiori and Dubey et al., 2024), and DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025). To probe the effect of model size within two model families, we additionally include the smaller Gemma-3-4B-IT (Gemma Team, 2025) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025) variants. We restrict the comparison to publicly available downloadable models that can be run and versioned in a controlled, reproducible evaluation setup. Each model is assessed on a manually annotated sample of 371 mentions using standard classification metrics. Beyond overall performance, we examine variation across grammatical realisation and referential type, identifying systematic asymmetries that affect dataset-scale inference.

Finally, we investigate whether the best-performing model can be applied at dataset scale by aggregating mention-level predictions across newspapers. This aggregation functions as a methodological demonstration of scalability rather than as a substantive historical interpretation.

Taken together, this study contributes (i) a controlled benchmark for aspect-level sentiment classification in OCR-degraded historical Slovene text; (ii) a systematic comparison of aspect-level sentiment classification in a few-shot setting between a Slovene-adapted instruction-tuned LLM and general-purpose instruction-tuned LLMs of comparable scale; and (iii) a diagnostic analysis of performance variation across grammatical realisation and referential type, highlighting systematic asymmetries that must be considered in dataset-scale sentiment aggregation. By situating LLM evaluation within a historically degraded and morphologically complex dataset, the paper provides empirical evidence on the reliability and limitations of instruction-tuned models as analytical instruments in digital humanities research.

2. Related Work

Sentiment analysis has been a longstanding research area with early approaches being lexicon-based, utilising tools like VADER (Hutto and Gilbert, 2014) and the Liu-Hu lexicon (Hu and Liu, 2004) to assign sentiment scores based on predefined sentiment lexica. While simple and interpretable, these methods were limited by lexicon coverage and struggled to capture linguistic nuances such as sarcasm and idiomatic expressions (Cambria et al., 2017). Machine learning methods, such as Support Vector Machines (Cortes and Vapnik, 1995) and Naive Bayes classifiers (Pang et al., 2002), improved sentiment classification by leveraging features like n-grams (Cavnar et al., 1994) and part-of-speech tags (Marcus et al., 1993). However, these approaches relied heavily on manual feature engineering and lacked the ability to model contextual information, which limited their performance on more complex sentiment tasks.

In recent years, language models have set new benchmarks in sentiment analysis. The introduction of transformer-based models, such as BERT (Devlin et al., 2019), has reshaped the field. BERT's bidirectional context modelling enables it to consider both preceding and succeeding words in a sentence, significantly boosting performance on sentiment classification tasks. When fine-tuned for ABSA, BERT has demonstrated remarkable improvements in identifying and classifying aspect-specific sentiment (Sun et al., 2019).

In a comprehensive experiment testing capabilities of LLMs in performing various sentiment analysis tasks, Zhang et al. (2024) highlight the strengths and limitations of LLMs. LLMs excel in simpler tasks, such as binary or trinary sentiment classification, even in zero-shot and few-shot settings, often matching or surpassing fine-tuned smaller language models. This makes them particularly effective when training resources are limited.

For Slovene, sentiment analysis research has developed largely around contemporary, non-historical text types and has produced several widely used datasets. Early work includes sentiment annotation of user-generated content in the Janes corpus, where Fišer et al. (2016) applied an SVM-based three-class classifier (POS/NEG/NEU) trained on a large manually labelled tweet collection and used it to automatically assign sentiment metadata across heterogeneous Slovene genres (tweets, forums, blogs, news comments, and Wikipedia talk pages). A major step toward target- and entity-oriented sentiment is SentiNews 1.0, a manually annotated news corpus (10,427 articles; five-point Likert scale; document/paragraph/sentence levels) and its derivative SentiCoref 1.0, which enriches 837 selected news articles with named entities, coreference chains, and target-level sentiment labels for entities in context. Žitnik et al. (2022) operationalise this setup as target-level sentiment analysis over entity-based document representations and show strong gains from BERT-based approaches over traditional feature-based methods. Recent work continues to explore ABSA pipelines for Slovene on SentiCoref, comparing lexicon-based and tree-based feature extraction with neural embedding approaches (Adhikari et al., 2024). Finally, Slovene sentiment research has also expanded into specialized domains such as finance, where recent benchmarking work evaluates LLMs for target-based financial sentiment in news (Muhammad et al., 2025). In contrast to these largely contemporary, clean-text settings, the present study targets OCR-degraded historical newspapers and frames the task as aspect-level targeted sentiment classification without supervised fine-tuning.

3. Data

The dataset on which we evaluate instruction-following LLMs for targeted sentiment classification in historical Slovene comprises three Slovene-language newspapers: *Slovenec* [The Slovene] (1873–1945), *Slovenski narod* [The Slovene Nation] (1868–1943), and *Slovenka* [The *Slovene* Woman] (1897–1902), sourced from the *sPeriodika* collection Dobranić et al. (2023). For copyright reasons, the collection only includes issues of newspapers published until 31 December 1914.

This dataset spans distinct political orientations and readership profiles, providing a heterogeneous discursive environment for evaluating targeted sentiment classification.

The newspapers differ substantially in size (Table 1). *Slovenski narod* contains approximately 183 million tokens, *Slovenec* 137 million tokens, and *Slovenka* 1.6 million tokens, amounting to over 320 million tokens in total. This volume enables dataset-level aggregation of model predictions while preserving variation across newspapers.

All texts are derived from historical scans using OCR of heterogeneous quality. Spelling inconsistencies, segmentation errors, and recognition noise introduce ambiguity at the token and sentence level. In addition, the data reflect morphologically rich and historically evolving language use, including orthographic variation and derivational complexity, affecting collective-identity expressions. The data were thus further refined through cleaning and preprocessing before undergoing automatic linguistic annotation (Dobranić et al., 2024).

Newspaper	Number of tokens
Slovenski narod	183,294,799
Slovenec	137,506,802
Slovenka	1,632,695

Table 1: Size of newspaper data

4. Methodology

4.1. Collective-Identity Extraction

In this study, collective identity refers to nouns and adjectives derived from ethno-national, regional/provincial, or other geography-based designations. Identity mentions therefore include: (i) ethno-national denominations (e.g., Španec [Spaniard], nemški [German], Jud [Jew]), (ii) regional or provincial identities (e.g., Istrijan [Istrian], Moravka [Moravian]), and (iii) other geography-based identities (e.g., evropski [European], južnoameriški [South American]).

Collective identity mentions were extracted using a manually constructed lexicon of nationality- and identity-denoting lemmas.

For nominal references, we manually inspected *Slovenka* and compiled a list of lemma types referring to national, regional, continental, or ethnic groups. This noun-lemma inventory was then applied unchanged to *Slovenec* and *Slovenski narod* to retrieve all matching nominal mentions across newspapers.

For adjectival references, lemma candidates were first extracted automatically based on Slovene derivational suffixes (e.g., -ski, -ški, including historical orthographic variants such as -zki and -žki). In *Slovenka*, all candidate adjective lemmas were manually reviewed. In the larger corpora (*Slovenec* and *Slovenski narod*), candidates were first filtered by frequency (minimum 90 occurrences) and subsequently manually inspected. The validated adjective lists from all three newspapers were merged, deduplicated, and consolidated into a single lexicon used for dataset-wide extraction.

To enable unified analysis, adjectival lemmas were mapped to their corresponding nominal identity categories (e.g., nemški [German] → Nemci [Germans], italijanski [Italian] → Italijani [Italians]), allowing nominal and adjectival realisations to be grouped under shared identity labels during aggregation.

No distinction was made at the extraction stage between adjectival mentions modifying collective actors (e.g., German army) and those modifying non-agentive entities (e.g., Slovene bread). This referential distinction is introduced later during manual evaluation (cf. Section 4.5).

4.2. Task Definition

The task is formulated as aspect-level sentiment classification. For each extracted collective identity mention (cf. Section 4.1), the model predicts the sentiment expressed toward that specific mention within its local context.

For adjectival realisations, sentiment is annotated with respect to the collective-identity adjective, even though the nominal head it modifies remains visible in context. The model must therefore attribute sentiment specifically to the marked identity expression rather than to the broader sentence or topic.

Each instance is represented as a structured entry containing: (i) a unique mention identifier, (ii) the target identity mention explicitly marked using XML-style tags, and (iii) a context window consisting of the sentence containing the mention together with the two preceding and two following sentences.

Given this input, the model assigns one of three labels to the marked mention: positive (POS), neutral (NEU), or negative (NEG).

4.3. Evaluation Dataset Construction

To evaluate model performance, we initially sampled 400 collective identity mentions from the three newspapers. The dataset was stratified at sampling time to ensure equal representation across newspapers and grammatical realisations: each newspaper contributed an equal number of instances, with 50% nominal and 50% adjectival mentions per outlet.

Mentions were randomly sampled within these constraints. To prevent over-representation of highly frequent identities, a frequency cap was applied during sampling: if a single identity accounted for more than 15% of a newspaper-specific subset (i.e., more than 20 instances within approximately 133 samples), excess instances were replaced via random resampling within the same newspaper and mention-type category. This threshold was set heuristically, as several identities were extremely prevalent in the dataset.

During manual annotation, annotators could assign the label *unknown* in cases where sentiment could not be determined due to pre-processing errors. A total of 29 instances received this label and were excluded from evaluation. The final evaluation dataset therefore contains 371 annotated mentions, which form the basis for all reported performance metrics.

The resulting dataset[1] spans a broad range of identity categories across newspapers and grammatical forms, providing a balanced and linguistically varied benchmark for mention-level sentiment classification under historical OCR noise.

4.4. Annotation Procedure

The evaluation set was divided evenly among three annotators, all trained linguists, with each mention annotated independently by a single annotator. Annotators consulted the sentence containing the target mention and, when necessary, up to two preceding and two following sentences, using the smallest sufficient context required to determine sentiment.

In addition to sentiment labelling, annotators recorded the referential type (group vs. non-group) for adjectival mentions. This variable was later used to analyse differences in model performance across referential contexts.

Inter-annotator agreement was not assessed, as each mention was annotated by only one annotator.

[1]Available for download through the GitLab repository at `https://dihur.si/muki/llm4dh/`.

4.5. Annotation Guidelines

Sentiment was defined as the evaluative stance expressed toward the referenced collective identity within its immediate context. Annotators were instructed to rely only on linguistic cues and other pragmatic signals available in the local context, and not on broader historical background knowledge. Mentions were labelled as positive or negative only when a linguistically explicit or clearly implied evaluative judgment was present, including cases of irony, patronizing stance, and related pragmatic effects where these could be inferred from the available context. In the absence of such evaluation, mentions were labelled as neutral. All interpretable mentions were assigned a sentiment label; the rest were marked as *unknown*.

Referential type was annotated independently of sentiment. All nominal identity mentions were treated as group references, as they inherently denote collective actors (e.g., Nemci [Germans], Slovenci [Slovenes]). For adjectival identity expressions, referential type was determined based on syntactic context. Adjectives modifying collective actors (e.g., German army) were classified as group references, whereas adjectives modifying inanimate or abstract entities (e.g., Slovene bread, German politics) were classified as non-group references.

Sentiment labelling was performed strictly with respect to the marked identity expression and not the broader topic or event described in the passage.

4.6. LLM Sentiment Inference Setup

Sentiment classification was performed in a few-shot setting using four openly available instruction-following LLMs representing a range of architectures and training regimes: GaMS3-12B-Instruct (Slovene-adapted), Gemma-3-12B-IT, Llama-3.1-8B-Instruct, and DeepSeek-R1-Distill-Qwen-14B. These models were selected to compare a language-adapted model against widely used general-purpose instruction-following models of comparable scale. We additionally included two smaller model variants, Gemma-3-4B-IT and DeepSeek-R1-Distill-Qwen-7B, to assess the impact of model size on performance and whether this impact generalises across model families.

Each collective-identity mention was processed independently using the structured input described in Section 4.2.

Models were prompted in Slovene with explicit instructions to perform targeted sentiment classification of the marked mention only. The prompt required that: (1) sentiment be evaluated exclusively for the tagged identity expression; (2) output be returned as valid JSON; and (3) the assigned label be one of the predefined categories: POS, NEU, or NEG.

Short illustrative examples of positive, negative, and neutral classifications were included in the prompt to clarify labelling criteria.

Model responses were required to conform to a predefined JSON schema containing the mention text and predicted sentiment label. No recalibration, filtering, or manual correction of sentiment labels was applied; all reported results reflect the models' raw predictions.

4.7. Dataset-Level Aggregation of Sentiment Predictions

Sentiment predictions were generated at the level of individual collective-identity mentions. To examine whether the selected model can be applied on the entire dataset, we aggregated these mention-level predictions by collective identity and newspaper. For each identity within each newspaper, we counted the number of POS, NEU, and NEG predictions and computed their relative proportions.

The resulting class distributions are presented in 5.2. Given the class-specific performance differences observed in Section 5.1—particularly lower recall for positive sentiment—the aggregated proportions should be interpreted as reflecting the model's relative classification tendencies rather than exact estimates of historical evaluative stance.

5. Results

This section presents a comparison of multiple instruction-following LLMs, then analyses the performance profile of the top-scoring model in greater detail. Finally, we demonstrate dataset-level sentiment distributions to assess the scalability and interpretative plausibility of large-scale inference.

5.1. Model Evaluation

5.1.1. Overall Performance Across Models

We evaluated four cutting-edge instruction-following LLMs in a few-shot setting: GaMS3-12B-Instruct (GaMS3), Gemma-3-12B-IT (Gemma-3-12B), Llama-3.1-8B-Instruct (Llama 3.1), DeepSeek-R1-Distill-Qwen-14B (DeepSeek-R1-14B). To assess the effect of model size within two model families, we additionally included two smaller variants: Gemma-3-4B-IT (Gemma-3-4B) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-R1-7B). All models were evaluated using the same prompt template, decoding parameters, and manually annotated evaluation sample (N=371 mentions). Table 2 reports overall accuracy, macro-averaged F1, weighted F1, and per-class F1 scores for each model.

Model	Acc	M-F1	W-F1	$F1_{POS}$	$F1_{NEU}$	$F1_{NEG}$
GaMS3	**0.693**	0.538	**0.708**	0.343	**0.794**	**0.478**
Gemma-3-12B	0.679	**0.555**	0.701	0.411	0.776	0.477
LLaMA 3.1	0.485	0.459	0.515	0.427	0.545	0.406
DeepSeek-R1-14B	0.580	0.519	0.616	**0.442**	0.665	0.450
Gemma-3-4B	0.620	0.428	0.637	0.154	0.742	0.388
DeepSeek-R1-7B	0.253	0.244	0.260	0.197	0.267	0.269

Table 2: Cross-model performance on mention-level sentiment classification (N=371). M-F1 denotes macro-averaged F1; W-F1 denotes support-weighted F1. Class-specific scores are reported as $F1_{POS}$, $F1_{NEU}$, and $F1_{NEG}$, corresponding to POS, NEU, and NEG sentiment labels.

GaMS3 achieves the highest overall accuracy (0.693) and the highest weighted F1 (0.708), indicating the strongest performance when class frequency is taken into account. Gemma-3-12B attains the highest macro-averaged F1 (0.555), suggesting slightly more balanced performance across sentiment classes. The difference between GaMS3 and Gemma-3-12B is small in overall accuracy ($\Delta \approx 0.014$), and in macro F1 ($\Delta \approx 0.017$), but the models differ slightly in their performance profile: GaMS3 performs better under class imbalance, whereas Gemma-3-12B shows slightly more balanced behaviour across sentiment classes. In relation to our research questions, these results suggest that Slovene adaptation yields a measurable but limited advantage over a strong general-purpose instruction-tuned model from the same family (Gemma-3-12B).

Comparing the smaller models, DeepSeek-R1-7B performs notably worse than its larger counterpart (0.253 vs. 0.580 accuracy). By contrast, Gemma-3-4B shows a more limited performance drop relative to Gemma-3-12B (0.620 vs. 0.679 accuracy), although its performance declines substantially on positive sentiment detection (0.154 vs. 0.411 $F1_{POS}$). This asymmetry may be related to the distillation process used to train DeepSeek-R1-7B, which can reduce sensitivity to low-frequency linguistic features important for handling Slovene. As a morphologically rich and relatively low-resource language, Slovene may be particularly sensitive to such degradation.

At the class level, all larger model variants perform best on neutral sentiment, though the magnitude differs substantially. GaMS3 achieves the highest F1 for neutral instances (0.794), followed closely by Gemma-3-12B (0.776). In contrast, both DeepSeek-R1-14B (0.665) and LLaMA 3.1 (0.545) have greater difficulties distinguishing descriptive from evaluative contexts in this domain.

Performance on negative sentiment is low for GaMS3 (0.478), Gemma-3-12B (0.477) and DeepSeek-R1-14B (0.450), and even lower for LLaMA 3.1 (0.406). Notably LLaMA 3.1 exhibits extremely high negative recall (0.950), paired with substantially lower precision, suggesting a tendency to overpredict negative sentiment. This pattern indicates reduced calibration rather than stronger discrimination.

Positive sentiment is consistently the most difficult class across models, especially smaller variants. GaMS3 yields an F1 of 0.343, Gemma-3-12B improves to 0.411, LLaMA 3.1 reaches 0.427, while DeepSeek-R1-14B achieves the best result (0.442). On the other hand, smaller models (Gemma-3-4B and DeepSeek-R1-7B) under-perform considerably. These results confirm that explicit positive evaluation in historical newspaper discourse is both less frequent and more challenging for models to detect reliably.

Taken together, these results indicate that no single model uniformly dominates across all evaluation criteria. GaMS3 provides the strongest overall performance under realistic class imbalance, while Gemma-3-12B demonstrates competitive class-balanced behaviour. The divergence between models—despite identical prompts and evaluation data—highlights the importance of empirical validation when deploying LLMs for sentiment inference in historically noisy corpora. Model performance in such DH settings appears sensitive not only to general instruction tuning, but also to language adaptation and domain robustness. These findings underscore the need for cross-model comparison when LLMs are used as analytical instruments in humanities research.

Given these results, subsequent dataset-level analyses are conducted using GaMS3. Although Gemma-3-12B yields slightly higher macro-averaged F1, GaMS3 provides the highest overall accuracy and weighted F1, as well as the strongest performance on neutral detection, which dominates the historical dataset distribution. This combination of robustness under class imbalance and stable neutral classification makes GaMS3 suitable for large-scale aggregation. At the same time, the cross-model variation observed above underscores that polarity estimates should be interpreted comparatively rather than as absolute measures of evaluative intensity.

5.1.2. Performance Profile of GaMS

Table 3 summarizes the performance of GaMS3-12B-Instruct on the manually annotated evaluation set. GaMS's overall accuracy reaches 0.693, with a weighted F1 of 0.708 and a macro-averaged F1 of 0.538. The gap between weighted and macro F1 reflects class imbalance and uneven performance across sentiment categories.

Neutral sentiment is detected most reliably (F1=0.794, P=0.858, R=0.739; support=287). Both precision and recall are comparatively high, indicating stable behaviour with relatively few false positives and false negatives in neutral contexts.

Sentiment	M	Ska	Snec	SN	**Total**
POS	P	0.400	0.500	0.200	**0.400**
	R	0.261	0.625	0.111	**0.300**
	F1	0.316	0.556	0.143	**0.343**
NEU	P	0.812	0.865	0.909	**0.858**
	R	0.812	0.703	0.700	**0.739**
	F1	0.812	0.776	0.791	**0.794**
NEG	P	0.154	0.511	0.222	**0.351**
	R	0.400	0.767	0.889	**0.750**
	F1	0.222	0.613	0.356	**0.478**
Macro	P	0.455	0.625	0.444	**0.536**
	R	0.491	0.698	0.567	**0.596**
	F1	0.450	0.648	0.430	**0.538**
Weighted	P	0.709	0.760	0.803	**0.749**
	R	0.694	0.713	0.669	**0.693**
	F1	0.697	0.724	0.708	**0.708**

Table 3: Performance of GaMS3 on the manually annotated sample for mention-level sentiment classification for *Slovenka* (Ska), *Slovenec* (Snec), *Slovenski narod* (SN), and the full dataset (Total). M denotes evaluation metric (P precision, R recall, F1-score). Macro and weighted averages are computed across classes.

Negative sentiment achieves an F1 of 0.478 (P=0.351, R=0.750; support=44). The substantially higher recall than precision indicates that the model captures most annotated negative instances but over-assigns negative labels in some neutral or positive contexts. In other words, negative sentiment is detected readily, though not always selectively.

Positive sentiment proves more challenging (F1=0.343, P=0.400, R=0.300; support=40). The relatively low recall suggests that a considerable proportion of positive instances are not recognized and are instead classified as neutral or negative. Compared to the negative class, the model appears more conservative in assigning positive sentiment.

Taken together, the results show a clear asymmetry in class behaviour. Neutral sentiment is most stable. Negative sentiment is detected with high sensitivity but reduced precision. Positive sentiment is under-detected relative to its annotated frequency. For downstream aggregation, this implies that dataset-level summaries are more likely to under-represent positive evaluations than negative ones, while neutral proportions remain comparatively robust.

5.1.3. GaMS Performance by Grammatical Category

We analyse GaMS's performance separately for nominal identity mentions (e.g., Nemci [Germans], Slovenci [Slovenes]) and adjectival modifiers (e.g., nemški [German], slovenski [Slovene]) to assess whether grammatical form affects classification behaviour (see Table 4).

Sentiment	M	Ska	Snec	SN	**Total**
Nouns					
POS	P	0.333	0.000	0.000	**0.222**
	R	0.143	0.000	0.000	**0.100**
	F1	0.200	0.000	0.000	**0.138**
NEU	P	0.702	0.714	0.919	**0.773**
	R	0.767	0.641	0.723	**0.713**
	F1	0.733	0.676	0.810	**0.742**
NEG	P	0.000	0.519	0.333	**0.385**
	R	0.000	0.667	0.857	**0.645**
	F1	0.000	0.583	0.480	**0.482**
Macro	P	0.345	0.411	0.417	**0.460**
	R	0.303	0.436	0.527	**0.486**
	F1	0.311	0.420	0.430	**0.454**
Weighted	P	0.581	0.615	0.799	**0.645**
	R	0.583	0.619	0.702	**0.633**
	F1	0.572	0.613	0.726	**0.630**
Adjectives					
POS	P	0.444	0.556	0.333	**0.476**
	R	0.444	1.000	0.167	**0.500**
	F1	0.444	0.714	0.222	**0.488**
NEU	P	0.918	1.000	0.900	**0.938**
	R	0.849	0.750	0.679	**0.759**
	F1	0.882	0.857	0.774	**0.839**
NEG	P	0.333	0.500	0.111	**0.310**
	R	1.000	1.000	1.000	**1.000**
	F1	0.500	0.667	0.200	**0.473**
Macro	P	0.565	0.685	0.448	**0.574**
	R	0.765	0.917	0.615	**0.753**
	F1	0.609	0.746	0.399	**0.600**
Weighted	P	0.833	0.898	0.818	**0.846**
	R	0.797	0.803	0.639	**0.749**
	F1	0.809	0.820	0.701	**0.777**

Table 4: Performance of GaMS3 on manually annotated nominal (N=180) and adjectival (N=191) mentions. M denotes evaluation metric (P precision, R recall, F1 score). Macro and weighted averages are computed across classes.

For noun mentions (N=180), GaMS achieves an accuracy of 0.633 and a weighted F1 of 0.630, indicating moderate overall performance. Neutral sentiment is identified most reliably (F1=0.742, P=0.773, R=0.713), showing relatively balanced behaviour. Negative sentiment yields an F1 of 0.482, with higher recall (0.645) than precision (0.385), suggesting a tendency to over-assign negative labels in ambiguous contexts.

Positive sentiment proves particularly difficult in the nominal subset. Although 20 positive instances are present, recall drops to 0.1, resulting in a very low F1 of 0.138. This indicates that the model rarely detects positive sentiment when it is expressed through nominal identity references, frequently defaulting instead to neutral or negative predictions.

For adjectival mentions (N=191), performance improves substantially. Accuracy increases to 0.749 and weighted F1 to 0.777, indicating greater overall stability. Neutral sentiment again achieves the highest F1 (0.839, P=0.938, R=0.759), reflecting strong precision and fewer false positives.

Positive sentiment shows marked improvement compared to nouns (F1=0.488, R=0.500), suggesting that evaluative meaning is more readily captured when embedded in adjectival modification rather than nominal reference.

Negative sentiment for adjectives displays a different error profile: recall reaches 1.00, but precision drops to 0.310, indicating systematic overprediction of negative labels in this subset. In other words, the model successfully captures all annotated negative adjectival instances but at the cost of labelling a substantial number of neutral contexts as negative.

Taken together, the results demonstrate that grammatical form influences classification behaviour not only in overall performance but in the precision–recall balance of individual classes. Nominal mentions are associated with missed positive evaluations, whereas adjectival mentions are more prone to over-attributing negative sentiment. This asymmetry is important for dataset-scale analysis, where mention-level predictions are aggregated into identity-level sentiment distributions, combining both grammatical realisations and therefore inheriting their respective error tendencies.

5.1.4. GaMS Performance by Referential Type

We next assess whether GaMS's performance varies according to referential type, distinguishing between group-referential mentions (direct references to collective actors, e.g., Nemci [Germans], or adjectival expressions modifying group-denoting heads, e.g., nemška vojska [German army]) and non-group mentions (typically adjectival nationality markers modifying inanimate or abstract heads, e.g., nemška politika [German politics]; see Table 5).

For group-referential mentions (N=245), GaMS achieves an accuracy of 0.645 and a weighted F1 of 0.660. Neutral sentiment is detected most reliably (F1=0.749, P=0.827, R=0.685), indicating reasonably balanced behaviour. Negative sentiment reaches an F1 of 0.491, with higher recall (0.711)

Sentiment	M	Ska	Snec	SN	**Total**
Group Modifiers					
POS	P	0.333	0.286	0.250	**0.304**
	R	0.235	0.500	0.200	**0.269**
	F1	0.276	0.364	0.222	**0.286**
NEU	P	0.741	0.816	0.936	**0.827**
	R	0.727	0.635	0.698	**0.685**
	F1	0.734	0.714	0.800	**0.749**
NEG	P	0.000	0.526	0.280	**0.375**
	R	0.000	0.741	0.875	**0.711**
	F1	0.000	0.615	0.424	**0.491**
Macro	P	0.358	0.543	0.489	**0.502**
	R	0.321	0.625	0.591	**0.555**
	F1	0.337	0.564	0.482	**0.509**
Weighted	P	0.619	0.710	0.822	**0.701**
	R	0.587	0.660	0.684	**0.645**
	F1	0.601	0.671	0.722	**0.660**
Non-Group Modifiers					
POS	P	0.667	1.000	0.000	**0.714**
	R	0.333	0.750	0.000	**0.357**
	F1	0.444	0.857	0.000	**0.476**
NEU	P	0.905	0.960	0.867	**0.907**
	R	0.927	0.857	0.703	**0.830**
	F1	0.916	0.906	0.776	**0.867**
NEG	P	0.500	0.429	0.091	**0.273**
	R	1.000	1.000	1.000	**1.000**
	F1	0.667	0.600	0.167	**0.429**
Macro	P	0.690	0.796	0.319	**0.631**
	R	0.753	0.869	0.568	**0.729**
	F1	0.676	0.788	0.314	**0.591**
Weighted	P	0.859	0.919	0.766	**0.856**
	R	0.857	0.857	0.643	**0.786**
	F1	0.848	0.874	0.688	**0.803**

Table 5: Performance of GaMS3 on manually annotated group-referential (N=245) and non-group mentions (N=126). M denotes evaluation metric (P precision, R recall, F1 score). Macro and weighted averages are computed across classes.

than precision (0.375), suggesting that the model captures most negative group evaluations but also assigns negative labels to a notable number of non-negative instances. Positive sentiment performs more weakly (F1=0.286, P=0.304, R=0.269), indicating that positive evaluations directed toward collective actors are frequently missed.

For non-group mentions (N=126), overall performance improves. Accuracy increases to 0.786 and weighted F1 to 0.803. Neutral sentiment again shows strong performance (F1=0.867, P=0.907, R=0.830), reflecting both low false-positive and low false-negative rates in descriptive contexts. Positive sentiment also improves compared to the group subset (F1=0.476, P=0.714, R=0.357). While precision is relatively high, recall remains moderate, indicating that some positive cases are still not detected. Negative sentiment achieves perfect recall (1.00); however, this result must be interpreted cautiously due to the very small number of negative instances in this subset (support=6), which limits stability and inflates recall.

Overall, the comparison indicates that direct group-referential mentions are more difficult for the model than non-group contexts. In group contexts, negative sentiment is more readily identified than positive sentiment, and positive evaluations are disproportionately missed. In non-group contexts, classification is more stable across classes, particularly for neutral and positive sentiment. This distinction is relevant for dataset-level aggregation, as summaries of sentiment toward collective actors may under-estimate positive polarity more than negative polarity.

5.2. Sentiment Distribution by Identity and Newspaper

Following the evaluation and diagnostic analysis above, we applied GaMS3-12B-Instruct to all identity mentions in the dataset (2.65 million instances) and aggregated predicted sentiment labels by identity and newspaper. Figure 1 shows the proportional distribution of POS, NEU, and NEG predictions for the five most frequently mentioned collective identities in the dataset. For each identity, three stacked bars correspond to the historical newspapers in Slovene: *Slovenka*, *Slovenec*, and *Slovenski narod*.

The distributions differ across identity categories. Nemci [Germans] show a consistently higher proportion of negative predictions across all three newspapers, with relatively smaller positive shares. In contrast, references to Slovenci [Slovenians] display a more balanced or mixed distribution, including a visibly larger proportion of positive predictions, particularly in *Slovenka*. Identities such as Avstrijci [Austrians] and Rusi [Russians] are dominated by neutral predictions across newspapers, with only limited positive or negative shares. The distribution for Čehi [Czechs] appears more mixed, with moderate levels of both positive and negative labels depending on the newspaper.

These patterns suggest that the model differentiates between identity categories rather than assigning sentiment uniformly. On the one hand, some of our preliminary findings largely align with mainstream historiographical narratives surrounding turn-of-the-century Slovene history. For example, the overwhelmingly negative sentiment that all three journals show towards Germans is indicative of contemporary Slovene-German nationalist political conflict (Čuček, 2016). Likewise, the pre-

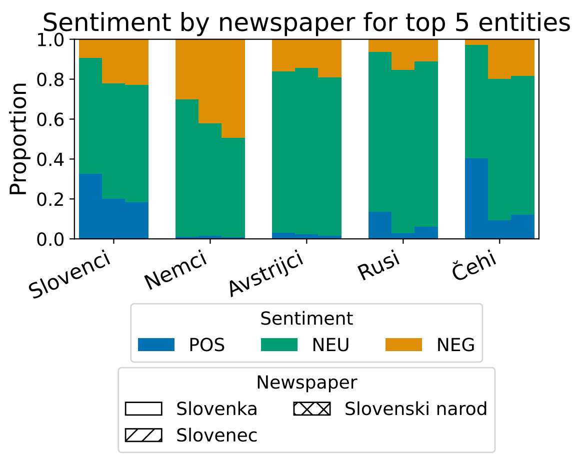


Figure 1: Sentiment class composition (POS/NEU/NEG) for the five most frequent collective identities in the dataset. For each identity, three stacked bars show the predicted class proportions in *Slovenka* (no hatch), *Slovenec* (//), and *Slovenski narod* (xx).

dominantly neutral sentiment expressed towards Austrians is expected given that Slovene political discourse of the time was mostly supportive of the Austrian state and Austrian political identity (Luthar et al., 2008).

Conversely, the results that we have gained by analysing the lemma Čehi [Czechs] demonstrate some of the interpretative complexities faced when compiling data using such models. While the overwhelmingly positive sentiment shown towards this group in *Slovenka* is not surprising, it is curious to see an overwhelmingly negative assessment of Czechs in the other two journals knowing that Slovene society was overwhelmingly sympathetic to Czechs during this period (Keršič-Svetel, 1996). While our interpretation remains speculative, we assume that the negative sentiment in the two journals was connected to the intensity of German-Czech nationalist conflict in the multiethnic crownland of Bohemia [Češka] — a province that is otherwise referred to using the same adjective [češki] as the Czech nation.

This analysis is intended as a plausibility check rather than a substantive historical argument. As shown in Section 5.1, model performance varies by class, with the highest reliability for neutral sentiment and lower performance for positive instances.

Figure 1 demonstrates that mention-level predictions can be aggregated at dataset scale while preserving identity-specific variation, provided that model performance characteristics are taken into account.

6. Conclusion

This study evaluated whether instruction-following LLMs can reliably perform targeted, mention-level sentiment classification in OCR-extracted text from historical Slovene newspapers and whether these predictions can support large-scale historical analyses when aggregated across the dataset. Using a manually annotated sample of 371 collective-identity mentions, we benchmarked four instruction-tuned LLMs and selected the Slovene-adapted GaMS3-12B-Instruct model for large-scale application. The comparison further shows that Slovene adaptation is beneficial, but that its advantage over a strong general-purpose model from the same family remains modest in a few-shot environment.

We show that the best-performing model is usable for this task, but its performance is class-dependent and varies across grammatical realisation and referential type. Neutral sentiment is detected most reliably. Negative sentiment is captured with relatively high recall but lower precision, indicating a tendency toward over-attribution, while positive sentiment is systematically under-detected, especially in nominal and group-referential contexts. These asymmetries directly affect dataset-level interpretation and require that aggregated polarity patterns be read as directional tendencies rather than exact measurements of historical evaluative stance.

This study also highlights an often overlooked disconnect between technical benchmarks and scholarly needs. For DH workflows, a model's F1 score is ultimately less important than its interpretative validity: how it behaves when used to map complex and often ambiguous human expression.

More broadly, the study provides a benchmark for targeted sentiment classification in OCR-degraded historical Slovene, highlights the necessity of cross-model comparison in DH settings, and offers an empirically grounded assessment of both the reliability and limitations of instruction-following LLMs as analytical instruments in DH research.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the Slovenian Research and Innovation Agency research programme “Digital Humanities: resources, tools and methods” (2022–2027) [grant number P6-0436], by the DARIAH-SI research infrastructure, by the Slovene Common Language Resources and Technology Infrastructure (CLARIN.SI), and by the project “Large Language Models for Digital Humanities” (2024–2027) [grant number GC-0002].

7. Bibliographical References

Swoichha Adhikari, Manan Gangwani, and Adithi Varadarajan. 2024. Aspect-based sentiment analysis for slovene texts: Models, lexicons, and embeddings. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 2, pages 1–6.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2017. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

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*, pages 4171–4186.

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2024. A lightweight approach to a giga-corpus of historical periodicals: The story of a slovenian historical newspaper collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 695–703.

Darja Fišer, Jasmina Smailovic, Tomaž Erjavec, Igor Mozetic, and Miha Grcar. 2016. Sentiment annotation of slovene user-generated content. In *Proceedings of the 2016 conference language technologies and digital humanities (JTDH 2016)*, pages 65–70.

Gemma Team. 2025. Gemma 3.

Aaron Grattafiori and Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225.

Marjeta Keršič-Svetel. 1996. *Češko-slovenski stiki med svetovnima vojnama*. Zbirka Zgodovinskega časopisa. Zveza zgodovinskih društev Slovenije, Ljubljana.

Oto Luthar, Igor Grdina, Marjeta Šašel Kos, Petra Svoljšak, Peter Kos, Dušan Kos, Peter Štih, Alja Brglez, and Martin Pogačar. 2008. The land between: a history of slovenia.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Iftikhar Muhammad, Marco Rospocher, Timotej Knez, and Slavko Žitnik. 2025. Benchmarking large language models for target-based financial sentiment analysis. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 785–795, Cagliari, Italy. CEUR Workshop Proceedings.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–385.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. *arXiv preprint arXiv:2410.06898*.

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. Association for Computational Linguistics.

Filip Čuček. 2016. Svoji k svojim: na poti k dokončni nacionalni razmejitvi na spodnjem Štajerskem v 19. stoletju.

Slavko Žitnik, Neli Blagus, and Marko Bajec. 2022. Target-level sentiment analysis for news articles. *Knowledge-Based Systems*, 249:108939.

8. Language Resource References

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. Corpus of slovenian periodicals (1771-1914) sPeriodika 1.0. Slovenian language resource repository CLARIN.SI.

Automatic Metrical Scansion of Poetry in a Low-Resource Setting

Pablo Ruiz Fabo[1,2], Anxo Alonso[1], Pablo Rodríguez Fernández[1], Paulo Gamallo[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
[2]Université de Strasbourg – LiLPa UR 1339
{pablo.ruiz.fabo,anxo.alonso.perez,pablorodriguez.fernandez,pablo.gamallo}@usc.gal

Abstract

We present the first neural systems for automatic metrical scansion of poetry in Galician, a Romance language close to Portuguese and Spanish. The task is threefold: First, identifying metrical syllables based on lexical ones; both syllable series may differ given metrical licenses modifying a line's syllable structure to enable stress-related rhythms. Second, identifying stress patterns, and third identifying the metrical syllable count, based on stressed positions. We manually annotated a corpus of 4,287 examples, a first in Galician, and fine-tuned an 8B-parameter LLM specialized in Galician and Portuguese, and two encoder–decoder models: ByT5, a token-free byte-to-byte model, and the multilingual mT5, which includes Galician. We also tested our recent symbolic scansion system. Several fine-tuning setups reached exact per-line accuracy above 90% on our test-set at all three scansion subtasks, using orthographic syllables with explicit stress marks as input. Encoder–decoders performed better than the LLM. The token-free ByT5 was best, particularly when adding the two surrounding lines to the input. The symbolic system (89.9% acc.) managed rare metaplasms infrequent in training data better than the neural ones, and the approaches can be seen as complementary.

Keywords: automatic metrical scansion, Galician poetry, transformers and LLM

1. Introduction

The automatic metrical analysis of poetry, or scansion, poses some challenges for NLP. The task requires identifying patterns such as the alternation between stressed and unstressed syllables. This is not trivial, as syllable boundaries and stress placement can be modified in poetry for rhythmic effects. Scansion automation can assist the development of large metrically annotated corpora, useful for comparing versification traditions and to help understand the distribution of metrical patterns across them. The value of such corpora was recently illustrated in Nagy et al. (2025), who performed a computational modeling of verse evolution in classical Latin, in the European Renaissance and in 19th-century Europe.

Both symbolic and neural approaches have been used for automatic scansion. Recently, large language models (LLMs) have started being used, particularly via fine-tuning and prompting proprietary models accessed via cloud platforms (Valença and Calegario, 2025; Kranti and Vajjala, 2025). Such work showed the feasibility of the task for LLMs. However, the reproducibility of cloud-based models is limited. Their long-term availability is not guaranteed and API changes may impact results. Besides, for large scale annotation, the cost may become unaffordable. A problem of the most capable LLMs, even those that can be deployed locally, is the computing demands posed by their large size.

Taking such limitations into account, we compare automatic scansion with models that can be run with the limited resources commonly available in Humanities teams: An 8B-parameter LLM (decoder), several smaller encoder–decoder models (0.3M parameters), and a symbolic baseline.

Some of the earlier studies on scansion-model training had access to generous preexisting training corpora. Here we focus rather on the case where no prior training data exists, and we need to create a first neural system from scratch. We work with Galician, a Romance language close to Portuguese and Spanish, co-official in the Spanish region of Galicia, and recognized in the European Charter for Regional or Minority Languages (Council of Europe, 1992). In the last decade, NLP projects have developed ample resources for Galician, covering the basic NLP pipeline (e.g. Gamallo et al., 2018), but also LLMs and specialized benchmarks to evaluate these (Gamallo et al., 2024; Rodríguez et al., 2025b). However, most NLP resources focus on contemporary Galician, whose orthography was normalized in the late 20th century. In the context of a Computational Literary Studies (CLS) project studying modern Galician poetry diachronically, we need to analyze 19th-century text, where orthographic practices vary given lack of a standard, decreasing the performance of available NLP models. In this sense, together with the prior unavailability of annotated data for scansion, the task corresponds to a low-resource scenario.

The paper's contributions are the following:

- First transformers-based systems for scansion in Galician evaluated on unnormalized 19th-century data, publicly available under an open license.[1] As a baseline for evaluation

[1]See Appendix A for model URLs.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 114–125
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

we use our recent symbolic scansion system (Ruiz Fabo et al., 2026).

- A manually annotated corpus for Galician scansion (lexical and metrical syllables, metrical licenses, syllable count) with 4,287 examples.[2]
- A comparison between several transformers-based models, encoder–decoder and decoder-only, with each other and with the symbolic baseline, pointing out the strengths of each. This can be informative for research on developing a first neural scansion system in other languages: Smaller encoder–decoder models, particularly the token-free ByT5, obtained best results overall, and the symbolic system outperformed neural ones with rare metrical phenomena, harder to learn statistically.

Although very large decoders excel at challenging semantic tasks, scansion relies less on semantic information, making comparisons with smaller and alternative architectures particularly relevant.

The paper is structured as follows: Section 2 outlines the state of the art. Sections 3 and 4 define the task and describe the corpora. The models developed are presented in Section 5, and the results are discussed in Section 6; Section 7 concludes.

2. Related Work

Symbolic approaches, statistical ones based on classical machine learning, and neural approaches are all present in the state of the art.

Systems we consider **symbolic** may include, besides rule-based methods, a statistical component that is not driven by machine learning (e.g. to disambiguate among scansion alternatives). For Portuguese, one of the traditions closest to Galician, an early rule-based system was created by Araújo and Mamede (2002), and Mittmann (2016) created *Aoidos*, a rule-based system with a thorough set of 159 rules, which achieves 97.5 exact line accuracy (higher with some test corpora), as tested on canonical authors, mostly Brazilian from the 18th-19th cts. In Spanish, Gervás (2000) created an early symbolic system. Navarro-Colorado's (2018) system specializes on hendecasyllable verse, reaching 95% perfect stress-pattern match per line, as tested on the ADSO 100 corpus, with 1,404 classical sonnets. The system performs automatic syllabification, using parts of speech (PoS) to determine syllable tonicity; rules resolve metrical ambiguities. De la Rosa et al.'s (2020) system (*Rantanplan*), also uses PoS, syllabification and metrical disambiguation heuristics. It achieves 95% exact stress-pattern match per line in ADSO 100, reaching, 65.02% on the more challenging *Carvajal* corpus (Pérez Venegas, 2015), with mixed-meter poems and a much larger metrical variety. *LibEscansión* by Sanz-Lázaro (2024) uses PoS and syllabification based on a phonological transcription, reaching 97.01% exact per-line stress match on ADSO 100. A final symbolic tool for Spanish is *Jumper* (Marco Remón and Gonzalo, 2021). It identifies stress patterns without prior syllabification. It reaches 95% perfect stress-match per line on ADSO 100 and 82% on the harder, mixed-meter Carvajal corpus. It is thus the best available Spanish mixed-meter scansion tool according to this benchmark. Our symbolic baseline for Galician is derived from this tool.

Symbolic systems have also been implemented for languages further removed from our task: For French, Delente and Renault (2015), and Bobenhausen and Hammerich (2015) for German. Several exist for English, to name just two, *Prosodic* (Anttila and Heuser, 2016), a metrical phonology parser (which also has some trainable components), and *ZeuScansion* (Agirrezabal et al., 2016b), based on Finite State Technology.

Statistical approaches, deploying classical machine learning (ML), were used by Estes and Hench (2016), using Conditional Random Fields (CRF) for Middle High German. Agirrezabal et al. (2016a) experimented with sequence labeling (CRF and Hidden Markov Models) and other classical ML models, for English scansion. Statistical methods appear as well in Plechac (2016), for Czech. Barbosa and Barbosa (2025) developed statistical scansion methods targeting Brazilian Northeastern phonology, not addressed in earlier Portuguese systems.

Several **neural systems** have been developed. Among those based on recurrent networks, Agirrezabal et al. (2017) and Agirrezabal (2017) used bidirectional LSTM-CRF for Basque, English and Spanish, in the latter case reaching 90.84% exact per-line accuracy on ADSO 100. Haider (2021) used a Bi-LSTM-CRF to predict syllable stress and other prosodic features, with above 83.1% per-line accuracy in English and 87.7% in German, using ca. 3,500 annotated examples in each language (Haider, 2023, 217). The same implementation was applied to Czech by Klesnilová et al. (2024), training with ca. 59,000 poems from the Corpus of Czech Verse (CCV) (Plecháč and Kolár, 2015), which totals ca. 2.3 million lines (ca. 66,500 poems). Several configurations were tested, e.g. giving the entire poem or single lines as the unit to tag. With the best configuration, results were above 99% exact-line accuracy, improving upon Plechac's (2016) statistical model, which achieved 81.9% exact-line accuracy. Koziev (2025) combined neural and other paradigms for Russian scansion.

[2]See project repository at `https://github.com/compellit/gama-trf`

Transformers have also been used. De la Rosa et al. (2021) fine-tuned encoders for Spanish scansion, with 8,748 examples. In the best setup (RoBERTa large and 100 epochs), perfect stress match per line was 93.43% on the ADSO 100 corpus. Glaser (2025) trained BERT (encoder) and T5 (encoder–decoder) for the scansion of 18th-century English iambic pentameter using >100K training examples, achieving 96% per-line accuracy.

Still within transformers, **LLMs** were used by Valença and Calegario (2025), who fine-tuned GPT 3.5 for Portuguese scansion, obtaining 88.6% per-line accuracy using 7,200 examples (87.19% if using 3,520). Kranti and Vajjala (2025) performed scansion via prompting, in Telugu. The results suggest that, without fine-tuning, quality is low: 60% accuracy for GPT-5 and 20% for Gemini-2.5-Pro (syllable classification task, Table 2 in their work).

Discussing transformer models for poetry-related tasks, Rosa et al. (2025) point out that, for tasks where manipulating syllables is important, the models' pre-trained tokenizers, which learn a subword vocabulary efficient for modeling, rather than targeting units like syllables, can underperform compared to a syllable-based or character-based tokenization. This informed our choice of a token-free option among our compared models (Section 5).

Some of the systems above showed remarkable accuracy, above 95%. However, in some cases, this involved massive training data (like the Czech neural tagger), or was achieved for a specific period or meter, like the T5 examples or the symbolic hendecasyllable taggers. Observing that extraordinary accuracy was only possible under specific conditions suggests that the task can have difficulties, regardless of the technological paradigm chosen.

3. Task Definition

We defined scansion as articulated into three subtasks, that are interrelated but can be evaluated individually, described below.

3.1. Metrical syllabification

In the context of our fine-tuning experiments, we defined metrical syllabification as obtaining metrical syllables based on lexical ones. Lexical syllables depend on general and language-specific phonological constraints. In Romance metrics, metrical syllables need not match lexical ones, which can be merged or split to allow stressed syllables to fall into specific positions, which helps create rhythmic patterns. For Galician, the main metaplasms are the following (see Table 2 for distribution):

- **Synalepha:** The final syllable of a word ending in a vowel merges with the initial syllable of the following vowel-initial word, forming a single metrical syllable across the word boundary. In Galician it is very frequent and is the default realization for vowel sequences across the word boundary. E.g. *a* and *o* in *Sin mi-rar, fi-x**a_o**s o-llos*.[3]
- **Syneresis:** Within a word, two vowels belonging to separate syllables and not constituting a diphthong are merged into a single syllable. It is relatively rare in Galician metrics. E.g. *e* and *o* in *Dé-ches-me fi-d**eo**s con gre-los*.[3]
- **Dialepha:** Takes place when the last syllable of a word and the first one of the following word could be pronounced as a single syllable, but are pronounced as separate ones. It can be seen as an exceptional absence of synalepha: *í* and *a* in *a-quí **a**-que-las vei-gas*.[3]
- **Dieresis:** Within a word, a diphthong is split into two syllables, adding a metrical syllable. It is rare in Galician, e.g. *ia* pronounced as two syllables in *do sil-ves-tre_ar-bo-re-do su-b**i-a**n-do*.[3]

3.2. Stress-pattern detection

Stress-pattern detection consists in identifying which of the metrical syllables are stressed. The output can be formalized in several ways, like a syllable tonicity boolean vector, with as many dimensions as syllables in a line, or as a list with the positions of the line's stressed metrical syllables.

In current Galician (using the official ILG/RAG norm), orthographic cues for syllable tonicity are more ambiguous than in Spanish: Stressed interrogative pronouns do not bear an accent mark and are homographic with unstressed relative pronouns and conjunctions. Word-final stressed syllables containing a falling diphthong do not bear an accent. In our 19th-century corpus, the challenge increases because there was no written norm, practices to represent stress vary, and an accent mark can represent vowel aperture or stress.

3.3. Syllable count

Following Spanish-style counting practices, which apply to Galician (cf. Carballo Calero, 1981; Fer, 1991), syllable count is affected by the position of the last stressed metrical syllable in the line. If stressed, one syllable is added to the count. If the last metrical syllable in the line is the antepenultimate, a syllable is deducted. When the line is divided into hemistichs, the same rules apply at the end of the first hemistich. For instance, an alexandrine (14 metrical syllables) is divided into hemistichs of 7 metrical syllables. These can

[3]English glosses: 1. Without looking, she fixes her eyes (R. de Castro). 2. You gave me noodles with rapini (J. M. Posada). 3. Here those plains (X. M. Cabada). 4. Whistling in the wild grove (F. Vaamonde).

have only 6 lexical syllables, if their 6th syllable is stressed.[4]

3.4. Challenges and applications

As a first challenge, lines can have metrical ambiguities. It may be possible to apply different sets of metaplasms, which would result in different stress patterns and syllable counts. For humans, lines in the context, particularly metrically unambiguous ones, can help decide how to scan a given line: what syllable count to target, which metaplasms to choose and for which syllables. This all poses challenges for an automatic scansion system, which must resolve ambiguities identifying possible solutions and excluding unlikely or impossible ones.

In some ambiguous cases, human experts accept more than one scansion, or even disagree as to the correct one. This poses a challenge for automatic evaluation of scansion, as several scholars have commented (recently Cuéllar, 2025, p. 10, Martin, 2025, p. 128). We discuss how this manifested in our corpora in Section 4.

As regards the subtasks' relative importance, the correct detection of stress-patterns is arguably more important than exact syllabification. It encodes metrical prosody more directly and, unlike exact syllable match, it is largely unaffected by mismatches at the character level unrelated to prosody. Besides, syllable count as defined can be derived deterministically from the stress pattern.

In terms of downstream applications, syllable count can provide a coarse overview of the form of a large versified corpus. Stress patterns have richer applications, and have been used to cluster corpora across languages and traditions (Nagy et al., 2025), or as a stylometric signal for authorship studies (e.g. Plecháč, 2021; Cuéllar, 2025).

Concerning alternative task definitions, it would be possible to define metrical syllabification as taking orthographic words as input instead of lexical syllables, as in Valença and Calegario (2025). Similarly to their experiment, GPT-4.1 fine-tuned on our corpus succeeded, with ca. 75% accuracy. Still, a smaller locally deployed decoder (Sec. 5) did not manage the task so defined. We thus use lexical syllables as input, as do most studies reviewed (Sec. 2). It would also be possible to use phoneme-based syllabification instead of orthography-based (as Klesnilová et al., 2024; Sanz-Lázaro, 2024 among others), a future work possibility.

[4] For the names of meters (in the sense of line types based on their syllable count) we follow the Spanish/Italian convention (termed *contagem grave ou espanhola* in Chociay, 1974) rather than French convention (*contagem aguda ou francesa*). The latter is currently more common in Portuguese versification studies but not in Galician based on our sources. Accordingly, in our descriptions, a hendecasyllable is a line with the last stress on the 10th metrical syllable, called *decassílabo* in Portuguese convention. Likewise, an alexandrine has 14 metrical syllables in our descriptions rather than the 12 it has under French/Portuguese-style syllable counts.

4. Corpora

We manually annotated a corpus of 19th-century poetry in Galician, totalling 4,287 lines from 98 poems by 29 authors. We used 3,487 lines for training (among which 697 lines for validation), and 800 lines as a held-out test-set. An example of the corpus format is given in Table 1.

The corpus contains a variety of meters representative of metrical poetry in modern Galician (see Fig. 1). About 34% of lines belong to mixed-meter poems, where scansion is harder because at least two meters (in the sense of syllable counts) appear (rarely more than three). The original orthography was largely preserved, but we performed a lightweight typographical and orthographic normalization which did not alter any metrically relevant features (see 5.1). The corpus covers authors from the mid 19th century and the Galician Renaissance (*Rexurdimento*) in the second half, when sustained literary production in the language reemerged, after centuries of decreased activity.

As a departure point for manual annotation, we carried out an automatic pre-annotation of syllabification, stress and metaplasms, thanks to heuristics that combine two sources: First, the output of our recent symbolic scansion system, which identifies stress patterns, syllable count and metaplasms (see 5.1). Second, the output of an automatic lexical syllabification tool we developed.

All these automatic pre-annotations were corrected manually, yielding a human-validated corpus annotated with lexical and metrical syllabification, stress patterns, metaplasms and metrical syllable counts for each line. To promote data quality, some error patterns that can be common in manual annotation were identified algorithmically and corrected manually: misalignments between lexical and metrical syllables given differences at the segment-level (rather than in stress), or impossible stress patterns, where the series of stressed positions is not compatible with the syllable count.

The entire corpus was annotated by the first author, and the test corpus was annotated manually by two of the authors. We computed inter-annotator agreement (IAA) between both for all subtasks defined in Section 3. IAA was 97.63% for exact metrical syllabification match, 98.63% for stress-pattern match, and 99.63% for syllable counts. We consider agreement substantial and in line with values reported in the literature. Navarro-Colorado (2018) report 96% IAA with three annotators in their 100-sonnet test corpus (fixed meter).

Line Text	Lexical Syllables	Metrical Syllables	Stress Pattern	Syllable Count
co eco das harpas	co / *e- / co / das / *har- / pas	c**o / *e**- / co / das / *har- / pas	2 5	6
renóvese a vida	re- / *nó- / ve- / se / a / *vi- / da	re- / *nó- / ve- / s**e a** / *vi- / da	2 5	6
Hoxe o meu eido	*Ho- / xe / o / *meu / *ei- / do	*Ho- / x**e o** / *meu / *ei- / do	1 3 4	5
que onte blanqueaba	que / *on- / te / blan- / que- / *a- / ba	qu**e *o**n- / te / blan- / qu**e- *a**- / ba	1 4	5

Table 1: Two groups of two manually annotated lines. A slash delimits syllables, stars indicate stress. Metaplasms are bolded: Dialepha applies in first line and synalepha in the second.
Note: First group: Lines by F. M. de la Iglesia (1880). Gloss: *with the echo of the harps / may life be renewed*. Second group: Lines by E. Martelo Paumán (1893). Gloss: *today my field / which yesterday was whitening.*

Metaplasm	train		test	
	N	%	N	%
Synalepha	1661	47.63	397	49.62
Complex (>2 syllables)	50	1.43	12	0.75
Syneresis	121	3.47	36	4.50
Dialepha	107	3.07	32	4.00
Dieresis	25	0.72	6	0.75

Table 2: Number and percentage of lines with metaplasms in the corpus splits.

As said in 3.4, there can be difficulties in establishing the reference scansion for some lines. In our test corpus, of the 11 lines where both annotators' stress patterns did not match, none of the cases was due to conceptual disagreement. Two cases would match if we consider alternative patterns proposed by annotators, one case was due to a missing criterion in the annotation guidelines, and 8 cases were due to errors by one of the two annotators (which were corrected after computing agreement). Regarding alternative patterns, these were provided for 18 lines in total, suggesting that a clear solution existed for humans in most cases.

Concerning corpus metadata, poems' titles, authors, and year of publication were recorded.

5. Experiments

This section presents the fine-tuning experiments, including data preprocessing workflow, baselines and experimental conditions tested.

5.1. Preprocessing and Lexical Syllabification

We use lexical syllabification with stress marks as the input for fine-tuning (Table 1). Syllable segmentation is largely deterministic in Galician, with some exceptions like cases of full-vowel vs. glide variation (Freixeiro Mato, 1998; Regueira Fernández, 2010), that can be managed with lexical resources. Syllable tonicity detection, however, presents some challenges because there is ambiguity in orthographic cues, more so in the unnormalized 19th-century variants from our corpus (see 3.2).

To tackle the task, we created a rule-based syllabification tool. This also implements a lightweight preprocessing, aimed at resolving syllable tonicity ambiguity in Galician (historical) orthography, helping detect stressed syllables by restoring stress marks where they are absent in 19th-century text, or otherwise assigning syllable stress to ambiguous syllables based on parts-of-speech (PoS) and context information. The workflow is fully described in Ruiz Fabo et al. (2026). It relies on candidates generated via regex and weighted edit distances, ranked in context with an n-gram language model. The in-vocabulary (IV) lexicon is based on resources from the LinguaKit and Apertium libraries (Gamallo et al., 2018; Forcada et al., 2011) and the 5-gram language model was trained on 126 million tokens in Galician from *CorpusNÓS* (De-Dios-Flores et al., 2024), with KenLM (Heafield, 2011). Although the approach is based on classical techniques, it allowed good results without the need to develop any training data and with few computational resources.

Preprocessing errors in lexical stress detection affected 24 of 800 lines in the test-set (3%). The remaining 2 errors were irrelevant for metrics, not affecting stress placement or syllable count.

The goal of the experiments was to assess the models' performance at learning metrical syllabification based on lexical syllables, rather than evaluating lexical syllabification based on orthographic words. The latter was implemented as a preprocessing step. To isolate lexical-to-metrical syllabification, we corrected preprocessing errors in the training and test data prior to fine-tuning.

5.2. Baselines

The literature (Section 2) suggests that prompting alone is not sufficient for scansion. To assess this on our language and corpus, we used prompting with GPT-5.2 and GPT-5 mini as a first baseline, in zero-shot and few-shot modes (20 examples).

The literature shows that symbolic systems can achieve high scansion quality, in some cases competitive with statistical and neural ones. As a symbolic baseline, we use our recent system (Ruiz Fabo et al., 2026), an adaptation to Galician of Jumper (Marco Remón and Gonzalo, 2021), which performs stress pattern detection without syllabification in Spanish poetry. We adapted its lexical resources to work with Galician. It generates scansion candidates (syllable tonicity vectors) based on vowel sequences that could be merged or split. The candidates are ranked based on their similarity to well-attested patterns in a stress pattern inventory, also taking into account the candidates selected for lines in a context window. The algorithm requires orthographic input with unambiguous tonicity, for which we used the preprocessing in 5.1. We corrected preprocessing errors before evaluation to assess scansion independently of preprocessing accuracy. Sec. 6 reports results before and after corrections. The symbolic system does not perform exactly the same task as the fine-tuned systems: it operates on orthographic words to infer stress patterns and syllable count directly, without syllabified input. Nevertheless, it provides a useful baseline, allowing us to gauge to what extent neural models implicitly acquire representations relevant to solve a task that the symbolic system encodes through explicit expert knowledge.

5.3. Conditions: Varying Input Context

We structured data for fine-tuning according to two different conditions. In ***single-line***, the model input and output consist in lexical and metrical syllables for a single verse-line respectively. In ***context-lines***, the input contains lexical syllables for the previous, current, and following lines (within the same poem), with delimiters to mark structure clearly. The output contains the metrical syllables for the current line only.

The two conditions test the influence of added context in fine-tuning. Humans use context lines to decide on the parse for a metrically ambiguous line; we thus tested whether context was also beneficial for the models.

5.4. Neural Model Fine-Tuning

Fine-tuned LLMs (decoder only) can succeed at scansion, as shown by Valença and Calegario (2025) with Portuguese and GPT-3.5. We wanted an LLM that can be deployed locally with limited resources (e.g. a Colab session, a usual tool in humanities teams in our experience). We chose 8B-parameter Nos-PT/Llama-Carvalho-PT-GL, specialized in Galician and Portuguese, created via continual pretraining of Llama-3.1-8B with 232M Galician and 250M Portuguese tokens, along the lines of methods in Rodríguez et al. (2025a), also using English and Spanish text to prevent catastrophic forgetting.

Scansion can be seen as transforming the input sequence (lexical syllables in our case) into an output one (metrical syllables). The task can involve removing or adding tokens, to apply metaplasms which erase or insert syllable boundaries. This is a natural fit for an encoder–decoder architecture, fine-tuned for sequence-to-sequence generation. Glaser (2025) recently fine-tuned T5 for scansion with success, so we chose T5-variants for our experiments.

Our first encoder–decoder base model was mT5-small (Xue et al., 2021). This is a multilingual T5 variant which includes Galician among its pretraining languages. Our training data is about 4,000 examples, the small version is appropriate for such data volume.

Our second encoder decoder model was ByT5-small (Xue et al., 2022), a token-free T5 variant. This model does not rely on a sub-word tokenizer, learning to perform byte-to-byte generation. The literature has shown that pre-trained tokenizers can perform worse than syllable- or character-based tokenization at manipulating poetic form (Rosa et al., 2025). By testing a token-free model we wanted to see if our setup also shows benefits from character-to-character learning.

We applied supervised fine-tuning, 5 runs per model with the same set of seeds. T5 variants were trained for 30 epochs, selecting the best checkpoint based on exact metrical syllabification match per-line, which is the most demanding one of our evaluation subtasks and also improves results at the others. Effective batch-size was 16 and learning rate 5×10^{-5}. The decoder was fine-tuned with LoRA (Hu et al., 2022) loaded in 4-bit precision, for 3 epochs, with an effective batch size of 8 and a learning rate 10^{-5}. Other hyperparameters are in the project repository.[5] We used `transformers` (Wolf et al., 2020) and, for the decoder, Unsloth (Han-Chen et al., 2025).

6. Results and Discussion

We evaluated all subtasks defined in Section 3. For each, our metrics are based on exact match per-line. We use the following abbreviations to discuss results: *sym*, *spm* and *scm* refer to exact-match per line in metrical syllabification, stress patterns and syllable counts respectively. We report results for the baselines and neural models. For the latter, we performed inference using greedy decoding and we report results averaged over 5 seeds.

The prompting baselines do not show exploitable results, as Valença and Calegario (2025) reported.

[5] https://github.com/compellit/gama-trf

	cd	sym	spm	scm
Baselines				
gpt5m-zs	sg	47.5	52.38	58.13
gpt52-fs	cx	61.2	68.12	82.5
symbolic		dna	89.88	**97.38**
Fine-tuning				
carvalho	sg	86.43 (.41)	87.05 (.53)	87.46 (.37)
	cx	87.75 (.78)	88.60 (.77)	88.80 (.87)
mt5-sm	sg	89.48 (.14)	90.46 (.14)	90.38 (.13)
	cx	90.37 (.51)	91.12 (.46)	91.02 (.53)
byt5-sm	sg	*92.02* (.44)	*92.42* (.53)	92.42 (.53)
	cx	**92.95** (.19)	**93.50** (.20)	*93.48* (.18)

Table 3: Exact-match per-line accuracy (%) and its *std* (5 runs) in metrical syllabification (*sym*), stress patterns (*spm*) and syllable counts (*scm*), in single-line (*sg*) and context (*cx*) conditions (*cd*).

Table 3 shows the worst (GPT-5 mini, zero shot *single-line*) and best results (GPT-5.2, few shot *context-lines*), which only reached 61.2 *sym*.

The symbolic baseline was very strong. In *spm*, both encoder–decoder models improved upon the symbolic baseline, with ByT5 in the *context-lines* condition achieving the largest margin (3.6 percentage points). The symbolic system, however, scored 0.72 points higher than the best decoder in *spm*. In *scm*, the symbolic system achieved the highest score, 3.9 points above the best fine-tuned model. When evaluating the symbolic system without prior correction of preprocessing errors, *spm* decreased to 88.12%, and *scm* remained unaffected.

Regarding results per neural architecture, encoder–decoders were better than the decoder. ByT5 was best, agreeing with earlier literature on improved performance of token-free models at formal poetry-related tasks. Accuracy differences per model within the same experimental conditions were significant, based on (i) a non-parametric bootstrap test per-item and (ii) a sign-flip permutation test treating poems as the unit of analysis. Statistical significance was assessed at $\alpha = 0.05$. We controlled the false discovery rate using the Benjamini–Hochberg step-up procedure; adjustments were applied over the full set of pairwise model and condition comparisons within each metric.

In terms of experimental conditions, for all models, *context* outperformed *single*, suggesting that the previous and following lines are helpful for learning and inference. However, evaluating the differences between conditions with the statistical tests above showed a significant difference only for ByT5. Note that we had tested other ways of adding context, like grouping 2 or 4 lines as input-output pairs, but they did not consistently improve performance across models and could degrade results, likely due to the increased output complexity. Adding context lines to the input while predicting a single-line output proved more effective.

Results in lines with specific metrical difficulties (Table 4) show a contrast between the symbolic and neural paradigm. As the distributions in Table 2 showed, the test corpus contains almost 50% of lines with the frequent synalepha license (*slp* in Table 4), and 74 lines with less frequent metaplasms: syneresis (*srs*), dialepha (*dlp*), dieresis (*die*). The best fine-tuned model outperforms the symbolic system by 3.6 points in stress-pattern match, but examining results for lines containing specific metaplasm types shows that improvement takes place mainly in lines with synalepha, whether it involves 2 syllables (*slp*) or more (*slpc*). For the infrequent metaplasms, the symbolic model outperforms the neural ones, more clearly with dialepha and dieresis, likely because they are the least represented in training data (Table 2), posing challenges for statistical learning. Results for syneresis (erasing a word-internal syllable boundary between vowels) were similar in the symbolic and the encoder–decoder, suggesting the latter's capacity to model task-specific transformations, or perhaps reflecting that erasing the boundary is easier with byte-to-byte learning than with subword tokens.

Per-meter results (Figure 1) show that the average per-line accuracy decreases as syllable count increases, likely due to the higher complexity of predicting an exact per-line pattern when the number of positions to consider grows. Results for the symbolic system also illustrate the tension between symbolic modeling and data-driven learning: The symbolic system outperformed neural models with alexandrines (14 syllables, 2 hemistichs) by 2 lines. Error analysis shows that the symbolic system correctly updated syllable counts if the first hemistich's last stressed position required it (see 3.3) also avoiding synalepha across the hemistich boundary. The rarity of these configurations makes them harder for learning-based systems to capture.

In summary, the fine-tuned systems, particularly ByT5, managed frequent metrical licenses robustly. Rare ones (particularly dieresis) were managed more adeptly by the symbolic model. Both approaches have practical value to assist in large-scale metrical annotation, followed by human validation. It would be possible to devise heuristics to select cases likely to require revision. The symbolic system marks metaplasms in its output; this signal could be used to flag lines that potentially have rare metaplasms for human verification. Although LLMs have become dominant for challenging tasks, from our results it is unclear if they are the most

Model	cd	**slp** · n=397		**slpc** · n=12		**srs** · n=36		**dlp** · n=32		**die** · n=6	
		spm	N	spm	N	spm	N	spm	N	spm	N
symbolic		86.9	345	83.3	10	**75**	**27**	**78.1**	**25**	**50**	**3**
byt5	sg	92.1	365.8	**100**	**12**	67.8	24.4	28.1	9	0	0
	cx	**93.1**	**369.6**			61.7	22.2	36.2	11.6		
mt5	sg	88.8	352.6	**100**	**12**	60	21.6	35	11.2	0	0
	cx	90.5	359.2			49.4	17.8	34.4	11		
carvalho	sg	80.7	320.2	71.7	8.6	38.3	13.8	58.8	18.8	0	0
	cx	85	337.4	**100**	**12**	40.6	14.6	42.5	13.6		

Table 4: Exact-match per-line accuracy in stress-patterns (*spm*) in lines with metaplasms: Synalepha (*slp*; *slpc* if $>$ 2 syllables), syneresis (*srs*), dialepha (*dlp*), dieresis (*die*) per model, in context (*cx*) or single (*sg*) conditions. The total number of metaplasms of each type follows *n*= in the header. For fine-tuned models, results are averaged over 5 runs, *N* is the average of correct lines.

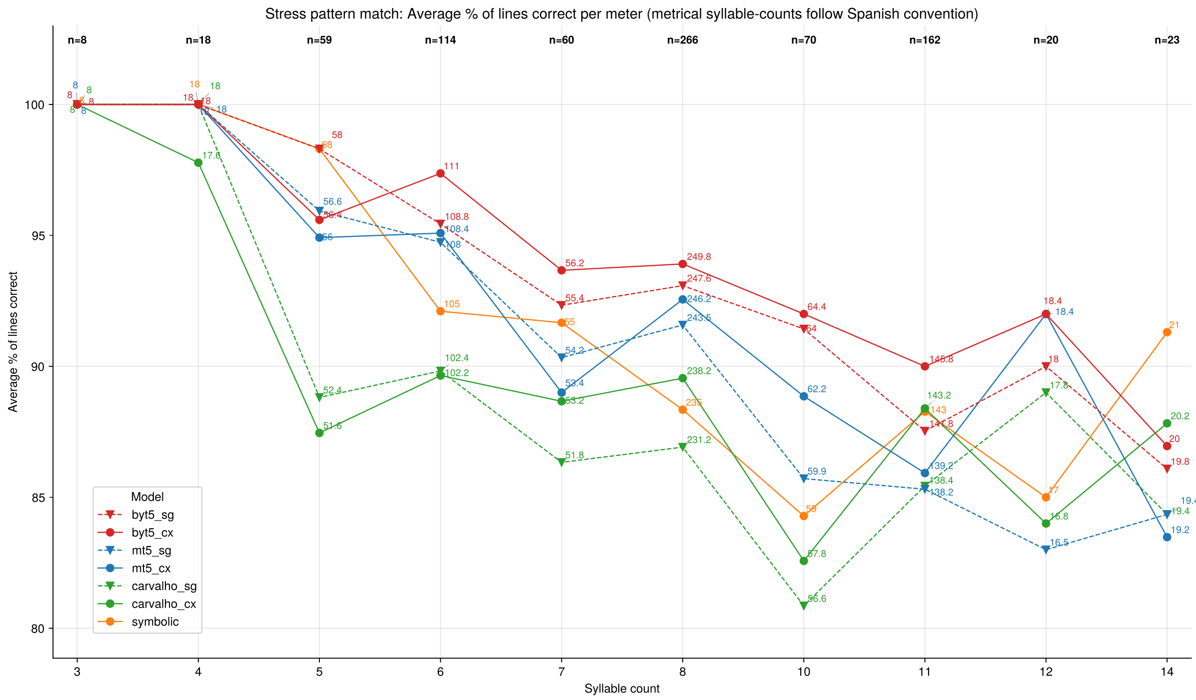


Figure 1: Stress pattern match per-meter (i.e. according to line syllable counts), using Spanish-style conventions (e.g. hendecasyllables are considered to have 11 metrical syllables and alexandrines 14).

efficient neural architecture for scansion, as pre-trained encoder–decoder models were consistently more successful at a fraction of the parameter size.

7. Conclusion and Outlook

We presented experiments on fine-tuning encoder–decoder models (ByT5, mT5) and an LLM (8B-parameter decoder-only Llama-Carvalho-PT-GL) for metrical scansion of poetry in Galician, including three subtasks: metrical syllabification, stress pattern detection and syllable count, evaluated with exact per-line accuracy. We also tested the influence of additional input context vs. single-line input. We created the first manually annotated training corpus for scansion in Galician for public release under an open license, including metaplasms or metrical license annotations. Using 3,487 examples for training and the remaining 800 for testing, several systems attained >90% on all three subtasks. The best model was the token-free ByT5, suggesting the usefulness of byte-to-byte learning for tasks involving units like syllables, which need not correspond to pre-trained tokenizers' subword tokens. Encoder–decoder models were better than the decoder-only. This has practical implications, since the former are smaller and proved faster in inference, for which they also demanded less VRAM than the decoder. The influence of additional input context was positive, but only when asking to predict metrical syllables for a single line given its two surrounding lines of lexical syllables in the input. Asking to predict multiple lines at a time underperformed compared to single-line prediction with our data; this behaviour may differ with a larger training set. There was a contrast between the symbolic and the fine-tuned systems. The latter obtained better results overall. However, on lines that have rare metaplasms (particularly dialepha and dieresis), most infrequent in training data, the symbolic system did better. Both approaches can thus be seen as complementary. In a scansion workflow, the robust analysis of more general patterns by the fine-tuned models could be combined with symbolic input to identify lines likely to involve exceptional cases and require human inspection.

Future work may explore several directions. Regarding the results with rare metaplasms, oversampling or other data augmentation methods may be attempted. In this work we operated on orthographic input. It would be possible to compare this with phonetic-transcription-based training, as was done in some related works, and might contribute to generalization as it would neutralize orthographic differences. Another future topic might be assessing cross-language transfer in closely related traditions, such as Galician, Portuguese and Spanish.

Acknowledgements

This work was supported by the European Union, under the Marie Skłodowska-Curie Actions, HORIZON MSCA-2023-PF, Grant ID 101149659, COMPEL – Computational Analysis of Peripheral Literatures.

The work was also supported by Xunta de Galicia – Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024–2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022–2026, ED431C 2022/19) and by the European Union's European Regional Development Fund – ERDF.

We are grateful to Elisa Fernández Rei (Instituto da Lingua Galega, ILG), for full-text access to 19th-century electronic sources in *Tesouro informatizado da lingua galega* (Version 4.1), directed by Antón Santamarina, Ernesto González Seoane, and María Álvarez de la Granja (http://ilg.usc.gal/TILG/).

Limitations

One constraint of the study is the size of the test set (800 items), which reflects the need to create annotations from scratch while maintaining a sufficiently large training set. This may limit the robustness of the findings and a larger test set would strengthen the conclusions.

In terms of evaluation, as noted in Section 3.4, certain verse lines admit more than one plausible analysis. Although alternative annotations were recorded for the small subset of such cases, only the first of the recorded annotations was used for system evaluation. This may slightly underestimate system performance, as outputs corresponding to valid alternative analyses are counted as incorrect.

Ethics Statement

We acknowledge a gender imbalance in the corpus, which predominantly reflects male authorship. While one of the central figures of 19th-century Galician literature and the *Rexurdimento*, Rosalía de Castro, is well represented, other women writers of the period are less visible in available digital resources and, consequently, in our dataset. Besides de Castro, the only additional woman author identified in our sources was Filomena Dato Muruais. Expanding the representation of women writers would require further archival research and the development of additional digital materials.

The study required the use of GPUs, which are associated with higher energy consumption than more modest computational setups. We compared resource-efficient models (e.g., 0.3M-parameter encoder–decoders) with an 8B-parameter LLM, as well as a symbolic system with minimal computational requirements. Since the encoder–decoder and symbolic systems achieved better results than the LLM, additional experimentation with larger models did not appear warranted given their higher computational cost.

Code and Data Availability

The project repository is at `https://github.com/compellit/gama-trf`.

8. Bibliographical References

Manex Agirrezabal. 2017. *Automatic Scansion of Poetry*. Ph.D. thesis, University of the Basque Country.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016a. Machine Learning for Metrical Analysis of English Poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages

772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A Comparison of Feature-Based and Neural Scansion of Poetry. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 18–23. Incoma Ltd. Shoumen, Bulgaria.

Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016b. ZeuScansion: A Tool for Scansion of English Poetry. *Journal of Language Modelling*, 4(1):3–28.

Arto Anttila and Ryan Heuser. 2016. Phonological and Metrical Variation across Genres. *Proceedings of the Annual Meetings on Phonology*, 3.

Paulo Alexandre Araújo and Nuno J Mamede. 2002. Classificador de Poemas. In *CCTE conference*, Lisbon.

Bryan K S Barbosa and Marcela Y A Barbosa. 2025. CordelSextilha.BR: A Benchmark for Poetic Form in Brazilian Cordel Verse Generation. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 736–747, Fortaleza/CE, Brasil.

Klemens Bobenhausen and Benjamin Hammerich. 2015. Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme *Metricalizer2*. *Langages*, 199(3):67.

Ricardo Carballo Calero. 1981. *Historia da Literatura Galega Contemporánea*. Galaxia, Vigo. Facsimile reprint, 2019.

Rogério Chociay. 1974. *Teoria do Verso*. McGraw-Hill do Brasil, São Paulo.

Council of Europe. 1992. European charter for regional or minority languages. European Treaty Series No. 148.

Álvaro Cuéllar. 2025. From Atoms to Waves: Rhythmic Stylometry for Authorship Studies of Early Modern Spanish Theatre. *Janus. Estudios sobre el Siglo de Oro*, (14).

Javier De la Rosa, Álvaro Pérez, Mirella de Sisto, Laura Hernández, Aitor Díaz, Salvador Ros, and Elena González-Blanco. 2021. Transformers analyzing poetry: multilingual metrical pattern prediction with transfomer-based language models. *Neural Computing and Applications*.

Javier De la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, and Elena González-Blanco. 2020. Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. *Procesamiento del Lenguaje Natural*, pages 83–90.

Eliane Delente and Richard Renault. 2015. Traitement automatique des formes métriques des textes versifiés. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 116–122, Caen, France. ATALA.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Workshop on Computational Linguistics for Literature (CLfL)*, pages 1–8.

Claudio Rodríguez Fer. 1991. *Arte literaria*. Xerais, Vigo.

Xosé Ramón Freixeiro Mato. 1998. *Gramática da lingua galega I: Fonética e fonoloxía*. A Nosa Terra, Vigo.

Pablo Gamallo, Marcos Garcia, César Piñeiro, Rodrigo Martinez-Castaño, and Juan C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Pablo Gamallo, Pablo Rodríguez, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open Generative Large Language Models for Galician. *Procesamiento del Lenguaje Natural*, 73:259–270.

Pablo Gervás. 2000. A logic programming application for the analysis of Spanish verse. In *Computational Logic—CL 2000*, pages 1330–1344. Springer.

Ben Glaser. 2025. TrochAIc: Metrical Tools for AI Interpretability. *Anthology of Computers and the Humanities*, 3:1438–1453.

Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725, Online. Association for Computational Linguistics.

Thomas Haider. 2023. *Computational Stylistics of Poetry*. PhD Thesis, Universität Stuttgart.

Daniel Han-Chen, Michael Han-Chen, and Unsloth AI. 2025. Unsloth. `https://github.com/unslothai/unsloth`.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. 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*.

Kristýna Klesnilová, Karel Klouda, Magda Friedjungová, and Petr Plecháč. 2024. Automatic Poetic Metre Detection for Czech Verse. *Studia Metrica et Poetica*, 11(1):44–61.

Ilya Koziev. 2025. Automated Evaluation of Meter and Rhyme in Russian Generative and Human-Authored Poetry. ArXiv:2502.20931 [cs].

Chalamalasetti Kranti and Sowmya Vajjala. 2025. MetricalARGS: A Taxonomy for Studying Metrical Poetry with LLMs. ArXiv:2510.08188 [cs].

Guillermo Marco Remón and Julio Gonzalo. 2021. Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural*, 66(0):77–87.

Meredith Martin. 2025. *Poetry's data: Digital humanities and the history of prosody*. Princeton University Press.

Adiel Mittmann. 2016. *Escansão automática de versos em português*. PhD Thesis, Universidade Federal de Santa Catarina.

Ben Nagy, Artjoms Šeļa, Mirella De Sisto, and Petr Plecháč. 2025. Metronome: tracing variation in poetic meters via local sequence alignment. *Computational Humanities Research*, 1.

Borja Navarro-Colorado. 2018. A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Petr Plechac. 2016. Czech Verse Processing System KVĚTA – Phonetic and Metrical Components. *Glottotheory*, 7.

Petr Plecháč. 2021. *Versification and Authorship Attribution*. Karolinum Press.

Xosé Luís Regueira Fernández. 2010. *Dicionario de pronuncia da lingua galega*. Real Academia Galega ; Instituto da Lingua Galega, A Coruña, [Santiago de Compostela].

Pablo Rodríguez, Pablo Gamallo, Daniel Santos, Susana Sotelo, Silvia Paniagua, José Ramom Pichel, Pedro Salgueiro, Vítor Nogueira, Paulo Quaresma, Marcos Garcia, and Senén Barro. 2025a. Enhancing large language models for underrepresented varieties: Pretraining strategies in the galician-portuguese diasystem. *Journal of the Brazilian Computer Society*, 31(1):1049–1062.

Pablo Rodríguez, Silvia Paniagua Suárez, Pablo Gamallo, and Susana Sotelo Docio. 2025b. Continued Pretraining and Interpretability-Based Evaluation for Low-Resource Languages: A Galician Case Study. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4622–4637, Vienna, Austria. Association for Computational Linguistics.

Rudolf Rosa, David Mareček, Tomáš Musil, Michal Chudoba, and Jakub Landsperský. 2025. EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 524–542, Albuquerque, USA. Association for Computational Linguistics.

Pablo Ruiz Fabo, Pauline Moreau, and Anxo Alonso Pérez. 2026. Automatic metrical scansion of galician poetry: First results. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese*. (Accepted, to appear).

Fernando Sanz-Lázaro. 2024. libEscansión: A Recursive Precedence Approach to Metrical Scansion. *Digital Humanities Quarterly*, 18(3).

Andre Valença and Filipe Calegario. 2025. Experimenting with Large Language Models for Poetic Scansion in Portuguese: A Case Study on Metric and Rhythmic Structuring. In *Proceedings of ICCC, the 16th international conference on computational creativity*, Campinas, Brasil. Association for Computational Creativity.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.

9. Language Resource References

Iria De-Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia, and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 593–599, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Pablo Gamallo, Marcos Garcia, César Piñeiro, Rodrigo Martinez-Castaño, and Juan C Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Petr Plecháč and Robert Kolár. 2015. The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1):107–118.

Dionisio Pérez Venegas. 2015. *El endecasílabo y su combinatoria en "Extravagante jerarquía"*. PhD Thesis, Universidad de Granada. Appendix (metrical annotations) to the PhD thesis.

A. Fine-tuned Model URLs

Model IDs and URLs are provided in Table 5 below. Accuracy (*Acc*) may differ slightly from results reported in Table 3 for the same base-model and fine-tuning condition, because the earlier table reports averages across 5 seeds, and the table below reports accuracy for the specific checkpoint uploaded to Hugging Face.

Cd	ID	Acc
sg	compellit/llama-carvalho-scansion-gl-sg	86.75
cx	compellit/llama-carvalho-scansion-gl-cx	88.75
sg	compellit/mt5-scansion-gl-sg	89.5
cx	compellit/mt5-scansion-gl-cx	91
sg	compellit/byt5-scansion-gl-sg	92.5
cx	compellit/byt5-scansion-gl-cx	93.25

Table 5: Model IDs on Hugging Face. Input-type refers to the (*sg*) and context (*cx*) conditions (*Cd*) from Section 5.3). *Acc* refers to exact match per-line in metrical syllabification.

SACRED: A Faithful Annotated Multimedia Multimodal Multilingual Dataset for Classifying Connectedness Types in Online Spirituality

Qinghao Guan[1], **Yuchen Pan**[2], **Donghao Li**[2], **Zishi Zhang**[2], **Yiyang Chen**[2], **Lu Li**[3,4], **Flaminia Canu**[1], **Emilia Volkart**[1], **Gerold Schneider**[2]

[1]Department of Communication and Media Research, University of Zurich, Zurich, Switzerland
[2]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3]Department of Informatics, University of Zurich, Zurich, Switzerland
[4] Ping An Technology, Shanghai, China

Abstract

In religion and theology studies, spirituality has garnered significant research attention for the reason that it not only transcends culture but offers unique experience to each individual. However, social scientists often rely on limited datasets, which are basically unavailable online. In this study, we collaborated with social scientists to develop a high-quality multimedia multi-modal datasets, **SACRED**, in which the faithfulness of classification is guaranteed. Using **SACRED**, we evaluated the performance of 13 popular LLMs as well as traditional rule-based and fine-tuned approaches. The result suggests DeepSeek-V3 model performs well in classifying such abstract concepts (i.e., 79.19% accuracy in the Quora test set), and the GPT-4o-mini model surpassed the other models in the vision tasks (63.99% F1 score). Purportedly, this is the first annotated multi-modal dataset from online spirituality communication. Our study also found a new type of connectedness which is valuable for communication science studies.

Keywords: LLMs, Connectedness, Spirituality, Dataset, Health communication

1. Introduction

From the primal reverence of natural forces and totems in ancient times to the complex theological systems today, spirituality has been an inextricable part of human society. The concept of totems, for instance, exemplifies the enduring spirituality pursuit throughout history. Rooted in the earliest human societies, totemism reflects a profound connection between people and the natural world, symbolizing a relationship that is both sacred and essential for the cultural identity of indigenous communities. These totems, be they animals, plants, or celestial bodies (Bolatova et al., 2019) were not mere symbols but represented a deep, spiritual kinship and understanding of the interconnectedness of all life. Therefore, human spirituality or religion are considered indispensable to human functioning and survival (Lee and Kanazawa, 2015).

Spirituality holds that all life is interconnected (Spaniol, 2002). At its core, relationality constitutes the essence of spiritual belief (de Souza, 2003; de Souza et al., 2004; de Souza, 2012). It can be argued that the impulse to connect represents one of humanity's most fundamental and inherent desires. The ability to forge spiritual connections – whether with the divine, within oneself, with others, or with the natural world – plays a crucial role in shaping individual identities and in defining our roles within the world.

Although very popular with audiences (Ramasubramanian, 2014, p.47), spirituality in the context of social media is still considered to be an understudied area of research (Janicke and Raney,

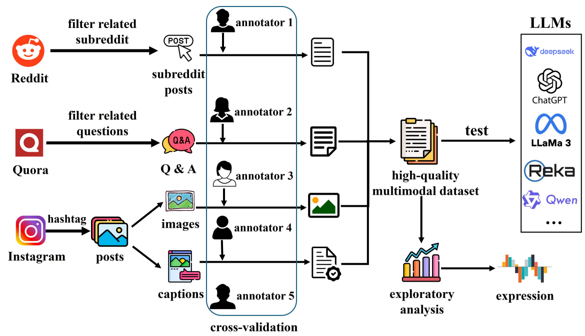


Figure 1: Workflow Diagram of SACRED Dataset Annotation and LLMs Evaluation

2016). Our paper makes the following contributions:

- This paper introduces **SACRED**, the first multimodal annotated dataset specifically designed for online spirituality research.
- The paper evaluates the visual and textual classification ability of 11 popular LLMs, including LLaMA3.1, Qwen2.5-VL, GPT-4o-mini, GPT-4o, DeepSeek-V3, GPT-4.1-mini, Claude Sonnet-4.5, Gemini-2.5-Flash, Reka, Gemma3, and LLaVA-1.5.
- This study, with the assistance of LLMs, uncovers a new type of connectedness within online spirituality and LLMs' understanding of spirituality. Our codes are available on GitHub[1].

[1]https://github.com/Qinghao-Guan/Spirituality_LLMs4SSH-2026

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 126–138
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Table 1: Concepts of Connectedness and Examples

Concept	Description	Example
Connectedness to Self	authenticity, inner harmony and inner peace, consciousness, self-knowledge, and experiencing and searching for meaning in life	Just spent the morning in deep meditation and self-reflection. Amazing how connecting with my inner self brings clarity and peace to my spirit. #InnerPeace #SelfAwareness
Connectedness to Others	compassion, caring, gratitude, and wonder	Just returned from volunteering at the local shelter, and my spirit feels so enriched! Connecting with others through acts of kindness truly uplifts the soul. #CommunityLove #MentalWellbeing
Connectedness to Nature	the deep sense of relationship that individuals feel with the natural world and understanding of humanity's place within the broader ecological system	Hiking through the forest today, I felt an incredible connection to the earth. The serenity of nature rejuvenates my spirit and calms my mind. #NatureHealing #SoulConnection #EarthSpirit
Connectedness to Transcendence	something or someone beyond the human level, such as the universe, transcendent reality, a higher power, or God	Attended a beautiful church service this morning. As we sang together and lifted our voices in prayer, I felt an incredible connection to God that filled my spirit with joy and hope. Truly a transcendent experience. #Faith #DivineConnection #SpiritualUpliftment

2. Theoretical Background

2.1. Online Spirituality

Since the 1980s, the practice of religion in online settings have consistently grown. Initially marked by the creation of religious sub-groups on platforms like Usenet and through email, a variety of religious activities started to surface, drawing attention from both the media and academic circles (Campbell, 2011). Examples of this include the establishment of virtual places of worship and online sites for spiritual pilgrimages (Campbell, 2006). By the mid-1990s, academics began to seriously study these distinctive online socio-spiritual practices and contemplate the implications of transferring traditional religious beliefs and rituals to the digital realm. Over the last decade, even more creative and unique forms of online religious expression have emerged, ranging from religious "podcasting" (Campbell and Teusner, 2015) to religion-online (Cheong et al., 2009) where served doctrinal interpretative and communal integrative functions are accomplished (Helland, 2002), and virtual worship spaces in Second Life for religious groups including Christians, Muslims, and Jews.

2.2. Dimensions of Connectedness

Motivated by conceptual considerations (Hill et al., 2000) and assessments of the significance of spirituality in the general population (Skrzypińska, 2014), much research focused on the multidimensional concepts of spirituality (Demmrich and Huber, 2019).

This research adopts the conceptualization of spirituality proposed by de Jager Meezenbroek et al. (2012), who defined it as "one's striving for and experience of connection with oneself, others, nature and the transcendent". Their definition considered the spirituality as a universal human experience and emphasized the multidimensional nature of spirituality. Most of the current research, thus, is grounded in this comprehensive conceptualization.

Four dimensions of connectedness (Ellison, 1983; Emmons, 2006) are often defined as follows: connectedness to self (involving authenticity, inner harmony/inner peace, consciousness, self-knowledge and experiencing and searching for meaning in life), connectedness to others (involving "compassion, caring, gratitude and wonder"), connectedness to Transcendence (referring to "something or someone beyond the human level"), and connectedness to nature (referring to the deep sense of relationship that individuals feel with the natural world and understanding of humanity's place within the broader ecological system). Examples are shown in Table 1.

2.3. Definition of Spirituality

A major challenge in researching internet-based religion is staying up-to-date with its swift evolution and transformations. This has posed a considerable problem in establishing theoretical models to study religious involvement on the World Wide Web (Helland, 2005). The accompanying issue is the definition of spirituality, which has been constantly changing from the early 20th century to the present (Peng-Keller, 2019). Initially, spirituality was viewed primarily in terms of the sacred (Otto, 1926; Eliade, 1959). By 2014, the definition of spirituality had evolved to include both theistic and non-theistic dimensions (Ramasubramanian, 2014). This study considers both the theistic and non-theistic dimensions and use a strict definition of spirituality:

> Spirituality is the pursuit and practice of experiences and beliefs that influence and nurture the spirit, fostering personal growth, meaning, and a sense of connection to something greater than oneself.

3. Related Work

3.1. Textual Analysis of Online Spirituality

Traditionally, textual data is analyzed by human coders who discern various characteristics and elements inherent in the text (i.e., content analysis in social science methodology). Manual annotation undoubtedly yields high-quality results. However, online platforms generate vast amounts of textual data on spirituality. Manually analyzing this data is impractical due to the overwhelming volume of data (Hilbert et al., 2019), thus NLP techniques are essential tools for efficient analysis (cf. Guan, 2025 and Guan and Lawi, 2024). The prevalent techniques employed in analyzing posts related to spirituality are mainly term analysis, classification and topic modeling. Demmrich and Huber (2019) test six dimensions of Huber's model (Huber, 2013) via frequency analysis and found that the "public practices" aspect, accounting for 11.69% of the total coding, includes rites of passage, religious service, holiday celebrations, and public meditation. Additionally, machine learning models were used in this research field. Holmberg et al. (2016), for instance, classified the tweets into 8 categories and concluded that these posts contain prayers that express praise, thanksgiving, devotion, care for other people, concern for one's own life and actions, questioning, despair, and even anger towards God. Sánchez-Garcés et al. (2021) employed a two-phase methodology (coding and sentiment analysis) to summarize key factors of patient interviews on their experiences during the disease. Their research revealed that patients often adopted a positive attitude (spiritual resilience) towards the symptoms and complications, significantly influenced by their religious faith. Kim et al. (2020) utilized structural topic modeling (STM) to identify the latent dimensions of Religiosity/Spirituality, and they extracted three distinct topics, namely Experience of, Engagement in, and Essence of Transcendence. Similarly, Winiger et al. (2025) presented a bottom-up approach to exploring the definitions of "religion" and "spirituality" by combining qualitative analysis with methods of distributional semantics. Their study demonstrates the potential of computational textual analysis for advancing research in religious studies. In the current LLM era, computational linguists (e.g., Gao et al., 2025), from an engineering perspective, developed a chat-based web interface called SpiritRAG, which is built on retrieval-augmented generation techniques. This tool, based on a database of 7,500 United Nations resolution documents addressing Religiosity/Spirituality issues, enables researchers and policymakers to efficiently search for information related to Religiosity/Spirituality.

3.2. Visual Analysis of Online Spirituality

The visual culture prevalent on the web naturally connects with holistic spirituality, which emphasizes intuitive understanding and spiritual experience (Noomen et al., 2011). Therefore, images posted online serve as reflective mediums, offering insights into the mental health of individuals. They not only depict visual narratives but also convey underlying psychological states and circumstances, making them valuable for understanding and assessing mental well-being. Tanhan and Strack (2020) studied the biopsychosocial spirituality of Muslims using an online photovoice methodology. They advocated for the shaping of public and mental health professional training, especially addressing issues related to the public and mental health services for Muslim communities. Xue et al. (2024) investigated the therapeutic potential of religion-related films by analyzing their distinctive emotional trajectories, through facial recognition techniques with YOLOv5, observing that the subtler emotional arcs foster introspection and personal growth. Also, images can be a tool serving as a sense of personal connection. Psychic practitioners draw on visual representations to convey their belonging to an exclusive cultural group using symbolic imagery to communicate shared knowledge and enhance their credibility within the psychic-spiritual commu-

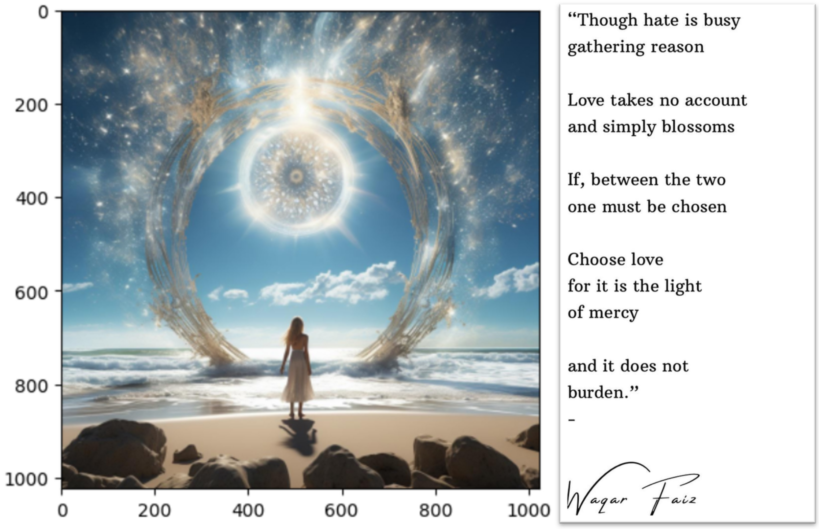


Figure 2: Instagram Image Example: The Left One Only Contains Visuality and The Right One Only Contains Textual Information

nity (Ryan, 2012). Sebek (2019) studied the storytelling on Instagram and analyzed the transformative effects of spirituality when people share visual posts, such as selfies. Trillò et al. (2021) also explored the visual contents on Instagram. They used a bottom-up taxonomy to classify images into categories and explored which visual repertoires are associated with users' value.

Generally, few studies integrate visual analysis techniques within the field of spirituality research (Goldenfein, 2019). This is due to two main challenges: firstly, automatic computer vision analysis techniques lack complete accuracy; secondly, hiring annotators is costly (cf. Gilardi et al., 2023), and annotating large datasets is both time-consuming and expensive. Our study intends to alleviate this.

4. Data Creation

4.1. Dataset Sampling

To collect data for this study, we developed a Python script utilizing the PRAW library to extract a substantial volume of posts from Reddit. The script adhered to Reddit's API usage guidelines, ensuring ethical and efficient data acquisition. We conducted a thorough examination of spirituality-related subreddits and selected ten for analysis: "Buddhism", "Christianity", "Enlightenment", "Meditation", "Mindfulness", "OpenChristian", "Soulnexus", "TrueChristian", "Spirituality", and "Taoism". The data was acquired on September 1st, 2024.

As to Instagram, people would like to share pictures and videos. Thus, inspired by the existing Python script, *instagram-scraper*[2], we extracted both images and their associated textual data including captions, hashtags, and mentions. In total, we collected data from 17 tags, including #Buddhism, #audiomeditation, #awakening, #christian, #christainity, #enlightenment, #meditation, #mindfulness, #openchristian, #soulnexus, #spiritualawakening, #spirituality, #taoism, #hindu, #confucianism, #islam, and #truechristian. Considering the data analysis capability, we only collected 500 pictures for each tag. Due to limited data, some tags only contain less than 500 pictures, such as only 69 images with the tag #soulnexus. Comparatively, the token numbers of the Instagram caption are much more than the subreddit posts.

For Quora data scraping, we drew inspiration from the existing Python script *quora-plus-bypass*[3]. We got access to the Quora URLs specified in a text file, extracting both the accepted and suggested answers. There were 12 topics (Spirituality, Meditation, Enlightenment (spiritual), Buddhism, Christians, Taoism, Tao (Chinese philosophy), Yoga, Beliefs, Philosophy of Religion, Theology, and Atheism) in total involving four specific cultures: Christianity, Taoism, Buddhism, Atheism.

Eventually, we collected 6769 Instagram images as well as their captions, 3819 answers from Quora platform, and 4922 subreddit posts.

4.2. Data Annotation

Annotation Guidelines

We invited five annotators [4] whose major or minor is Computational Linguistics. Each annotator

[2]https://github.com/meetmangukiya/instagram-scraper

[3]https://github.com/NitinN77/quora-plus-bypass

[4]the fifth one is an external annotator for evaluation

possesses a rich interdisciplinary background, having taken courses in social sciences and humanities in previous semesters. All annotators were trained to familiarize themselves with the annotation requirements and challenges of the dataset. We provided them with a detailed codebook (See Appendix 1. Annotation codebook), which included key term definitions and procedures, as well as illustrative examples. The annotation was conducted over one month on a part-time basis. This allowed annotators to thoroughly engage with the dataset, thereby enhancing the reliability of the annotations. All personal information has been deleted.

Annotating Textual Data

For the present research, we use three distinct textual datasets: subreddit posts, Quora answers, and Instagram captions. Our annotators were tasked with a two-step annotation process. Initially, they determined whether a text was related to spirituality. If a text was spirituality-related, the next step involved classifying it into one of four types of connectedness. For Instagram captions, the annotators were also required to identify and record the language of the caption, which is useful for multilingual research of spirituality.

Annotating Images

As part of our multimodal dataset, the annotation process extended beyond textual data to include images on Instagram. Annotators first judged whether an image could be classified as spiritual. For those images identified as conveying spiritual content, annotators were asked to determine the specific type of connectedness it represented. This classification drew upon the same categories used for textual data, as detailed in our codebook, ensuring a consistent approach across modalities. In addition to identifying spiritual content and its type, annotators were also instructed to note the language used in the images (e.g., memes or screenshots). Not all spiritual imagery includes textual language (see the left example in Figure 2) and some images only contain textual information (see the right example in Figure 2).

5. Overview of the SACRED Dataset

5.1. Data Exploratory Statistics

Among these, Quora exhibits the highest proportion of spirituality-related content at 53.94%, significantly surpassing the other platforms. Instagram images show a relatively low engagement with spirituality-related content at 28.87%, followed by Instagram captions at 21.27%. Reddit features the lowest percentage of such content, standing at 17.76%.

Given that our dataset is multi-lingual, we explored the language distribution of spiritual posts on Instagram, focusing specifically on posts excluding English. This sub-section provides a comprehensive overview of the linguistic diversity in spiritual communications on this platform. The language distribution of Instagram captions, represented by green bars in Figure 3, indicates a significant prevalence of German, Spanish, and Chinese.

Table 2: Percentage of Spirituality-related Content by Platform

Source	**Spirituality-related (%)**
subreddit	17.76%
Instagram captions	21.27%
Instagram images	28.87%
Quora topics	53.94%

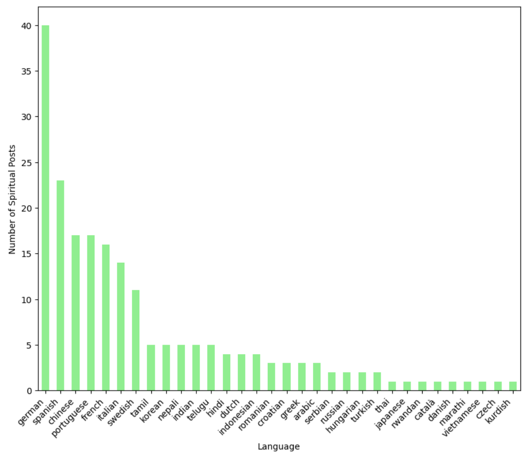


Figure 3: Language Distribution of Instagram Caption Excluding English

Conversely, the language distribution for text within Instagram images, shown by purple bars in Figure 4, presents a different pattern. German remains dominant, followed by Chinese and Arabic.

5.2. A New Type of Connectedness

A notable finding of data analysis is that 12 posts were identified as relevant to spirituality but did not align with any of the four predefined types of connectedness. We employed ChatGPT-o3 to analyze the content of these unclassifiable posts (the prompt is given in Appendix 6). This analysis led to the identification of a new type of connectedness.

Connectedness to Art There are some posts about creating or engaging with art that reflects spiritual themes, such as character drawings with spiritual significance or abstract art that evokes spiritual contemplation. They highlight a connec-

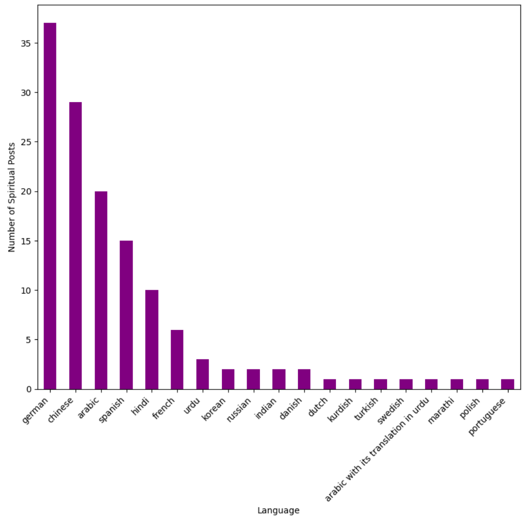


Figure 4: Language Distribution of Instagram Images Excluding English

tion to spirituality through creativity and artistic expression. Also, some posts described authors' life changes by reading spiritual books like "The Untethered Soul" or "The Bible Tells Me So", which shows a connection to spirituality through the lens of literary influence and personal interpretation. In discussions with several experts in spirituality studies, one scholar suggested that connectedness to art could be understood as either connectedness to the self or to transcendence. Works of art, such as paintings or classical music, are created by artists (e.g., Vincent van Gogh) and resonates deeply with the audiences. Philosophically, this form of connection can be interpreted as an engagement with the artist's spirit or creative essence, which differs fundamentally from the inner spiritual purification typically achieved through meditation.

6. Evaluation

6.1. Tasks

To rigorously assess the efficacy of different computational approaches in classifying spirituality as well as connectedness types within our annotated datasets, we implemented a comparative evaluation framework. This framework included a variety of methods: a lexicon-based rule method, a fine-tuned BERT model, and LLMs for classifying the connectedness types of 193 images and 447 posts [5]. The selection of these images and posts was proportionally sampled according to the distribution of each hashtag to ensure representative coverage of different thematic categories.

6.2. Experimental Setting

We tested the performance of 13 types of models (11 LLMs and 2 traditional models). LLM prompts were attached in Appendix (see Appendix 7), and an introduction to each model is shown below.

Rule-based method We employed rule-based methods to develop four dictionaries, each comprising keywords representative of a specific connectedness type, and subsequently performed classification based on these dictionaries.

Fine-Tuning BERT We utilized the uncased version of the BERT model (Devlin et al., 2019) and fine-tuned it on a selection (20%: 875 rows in total, proportionally sampled from each platform) of the whole dataset.

LLaMA3.1-8B We incorporated the LLaMA3.1 model with 8B parameters (text-only), which is a small, open-source, but prominent model of the LLaMA series (Touvron et al., 2023).

Reka Flash-21B It is a powerful 21B-parameter language models trained on 5T text tokens (Reka Team, 2024). We selected this model because 15% of the pretraining data of Reka is multilingual involving 32 diverse languages.

GPT We integrated three close-source GPT models (OpenAI, 2023), including GPT-4o, GPT-4o-mini, and GPT-4.1-mini. The comparison is able to reveal the performance difference between smaller, distilled version and full-featured version.

Qwen2.5VL-7B To test the visual classification ability, we utilized Qwen2.5VL model, which is a multimodal model with 7B parameters by Alibaba Group (Qwen Team, 2025).

DeepSeek-V3-671B DeepSeek-chat is built on Multi-Head Attention (MHA) block of a transformer and Mixture-of-Experts architecture (DeepSeek-AI, 2024). The MHA accelerate inference requiring far less KV cache. Also, Deepseek proposed fine-grained expert segmentation and shared expert isolation.

[5] We limited our evaluation to a smaller dataset because the small LLMs were tested locally on Ollama, and processing the entire dataset would have required several months of computation time. In addition, the outputs of small LLMs often failed to follow the given instructions, requiring extensive human post-processing of the annotation results.

LLaVA-v1.6-13B LLavA is "a multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding" (Liu et al., 2023). This model performed well in computational social science, such as image captioning (Sun et al., 2025).

Claude-Sonnet-4.5 Released on 29 September 2025, Claude-Sonnet-4.5 demonstrated significant advancements over its predecessors, particularly in handling complex agents and coding tasks. It represents Anthropic's most advanced multimodal model available at the time of writing.

Gemini-2.5-Flash Gemini-2.5-Flash, which excels in multimodal understanding, such as the ability to process up to three hours of video content, is Google's most advanced multimodal model available at the time of writing (Google Gemini Team, 2025).

Gemma-3-12B Gemma3 is a lightweight, open-source large language model by Google. It is built from the same research and technology used to create the Gemini model (Team, 2025).

6.3. Main Results

Our experiment evaluated the textual classification performance of the rule-based method, fine-tuned BERT, and a series of LLMs (see Table 3). DeepSeek-V3 outperforms all other methods in classifying the connectedness types of posts. For instance, in the definition-injected setting, it achieves the highest scores on Quora Answers (79.19%). Across the board, models that utilize the definition-injected setting show possible improved performance compared to their zero-shot counterparts. For example, LLaMA3.1's score on Instagram caption increases from 45.27% to 54.05% when definitions are injected. The traditional methods (rule-based and fine-tuning BERT) lag behind the advanced LLMs, which shows the potentials of LLMs in settings where contextual understanding is crucial.

In the Instagram image classification task, GPT-4o-mini achieved the highest F1 score (63.99%) (see Appendix 8), while GPT-4.1-mini attained the highest accuracy. These results indicate that the ChatGPT model series continues to outperform other models in the task of classifying visual spirituality connectedness. What surprised us was the performance of the Qwen2.5-VL model, which achieved an accuracy of 68.31% despite being a relatively small model with 7 billion parameters, outperforming LLaVA-1.5 model with 13 billion parameters and Gemma3 model with 12 billion parameters. Besides, although Gemma-3 and Qwen-2.5-VL are smaller models, their outputs remain consistently well formatted as large models, such as GPT-4o and DeepSeek-V3.

Comparing the classification performance among platforms, models generally perform better on Reddit posts and Instagram captions than on Quora answers. Our error analysis revealed that Quora answers tend to be relatively long, which means that a single response may contain elements of multiple connectedness types, even if the overall focus is on one primary type. While human annotators can readily identify the dominant connectedness type, LLMs may be more influenced by local contextual cues (Zeng et al., 2025), leading to misclassifications.

7. Discussion

To control for variables, we did not enable reasoning in our experiments, even though we believe that doing so could have yielded higher performance scores (e.g., DeepSeek-R1[6] and GPT-5[7]). Our findings revealed that while advanced models like DeepSeek-V3 and GPT-4o-mini exhibit superior performance, they are not yet fully reliable in classifying abstract concepts inherent in spirituality. Notably, the inclusion of a strict definition of spirituality (see Section 2.3) can help the models understand the task in some unclear posts, highlighting the importance of contextual information in processing complex tasks. Also, we observed an interesting phenomenon that GPT models sometimes generated the label "connectedness to transcendence" even though the prompt explicitly instructed them to choose from five predefined categories. We know the reason is some scholars use "connectedness to transcendence" in their studies (e.g., Rahe and Jansen, 2024). This suggests a limitation in the model's ability to strictly adhere to prompt constraints (i.e., hallucination), even when the instruction was explicit.

Despite these contributions, our study has several limitations:

Our dataset, while multimodal, predominantly features English-language content. The language distribution analysis shows some diversity, but languages like German and Spanish are underrepresented. This limits the applicability of our findings to a global context where spirituality is expressed in myriad languages. In the further research, we will incorporate data from a wider array of platforms and cultural contexts to enhance the representatives of our dataset.

[6]https://api-docs.deepseek.com/news/news250528

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

Table 3: Accuracy Performance Comparison Across Platforms and Methods

Model	Reddit		Instagram Caption		Instagram Image	Quora	
	Zero-shot	Definition-injected	Zero-shot	Definition-injected	Definition-injected	Zero-shot	Definition-injected
Rule-based	22.67		20.27	-	-	10.74	-
Fine-tuning BERT	53.33	-	46.62	-	-	48.32	-
LLaMA3.1	58.00	66.00	45.27	54.05	-	53.02	52.35
Reka Core	52.00	55.33	43.00	45.33	-	31.00	28.33
DeepSeek-V3	**72.00**	**78.67**	**77.70**	**77.70**	-	**73.15**	**79.19**
GPT-4o	67.33	76.00	59.46	65.54	74.60	47.65	59.73
GPT-4o-mini	67.33	68.67	54.05	61.49	75.82	50.34	55.70
GPT-4.1-mini	-	-	-	-	**76.04**	-	-
Gemini-Flash	-	-	-	-	73.06	-	-
Sonnet-4.5	-	-	-	-	73.96	-	-
Qwen2.5VL-7b	-	-	-	-	68.31	-	-
LLaVA-13b	-	-	-	-	40.37	-	-
Gemma3-12b	-	-	-	-	45.03	-	-

8. Conclusion

In this paper, we introduced **SACRED**, the first multimodal multilingual annotated dataset specifically designed for online spirituality research. Our work underscores the complexity of classifying abstract and deeply personal concepts using computational methods for communication sciences. While advanced LLMs show promise, they are not infallible and require careful contextualization and guidance. The **SACRED** dataset provides a valuable foundation for future research, but it also highlights the need for ongoing refinement of both computational tools and theoretical models.

9. Acknowledgements

This paper was inspired by the seminar "Language Technology for Social Media Analysis" held in 2023 by Martin Volk (1961-2025†). This work is dedicated to his memory, honoring his lifelong contributions to the field of computational linguistics.

Asiat Dautovna Bolatova, Fatima Taulanovna Uzdenova, and Rauzat Abdullakhovna Kerimova. 2019. Totem beliefs in Karachay-Balkar folklore and literature. *Journal of History Culture and Art Research*, 8:223–227.

Heidi Campbell. 2006. Religion and the Internet. *Communication Research Trends*, 26:3–24.

Heidi Campbell. 2011. *Internet and religion*. Blackwell.

Heidi Campbell and Paul Emerson Teusner. 2015. *Internet and social networking*. Routledge.

Pauline Hope Cheong, Jessie P. Poon, Shirlena Huang, and Irene Casas. 2009. The Internet highway and religious communities: Mapping and contesting spaces in religion-online. *The Information Society*, 25:291–302.

Eltica de Jager Meezenbroek, Bert Garssen, Machteld van den Berg, Dirk Van Dierendonck, Adriaan Visser, and Wilmar B. Schaufeli. 2012. Measuring spirituality as a universal human experience: A review of spirituality questionnaires. *Journal of Religion and Health*, 51:336–354.

Marian de Souza. 2003. Contemporary influences on the spirituality of young people: Implications for education. *International Journal of Children's Spirituality*, 18:269–279.

Marian de Souza. 2012. Connectedness and connectedness: The dark side of spirituality–implications for education. *International Journal of Children's Spirituality*, 17:291–303.

Marian de Souza, Patricia Cartwright, and E. Jacqueline McGilp. 2004. The perceptions of young people who live in a regional city in Australia of their spiritual wellbeing: Implications for education. *Journal of Youth Studies*, 7:155–172.

DeepSeek-AI. 2024. Deepseek-V3 Technical Report. Technical report, DeepSeek-AI.

Sarah Demmrich and Stefan Huber. 2019. Multidimensionality of spirituality: A qualitative study among secular individuals. *Religions*, 10:613.

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, pages 4171–4186.

Mircea Eliade. 1959. *The sacred and the profane: The nature of religion (Vol. 81)*. Houghton Mifflin Harcourt Inc.

Craig W. Ellison. 1983. Spiritual well-being: Conceptualization and measurement. *Journal of Psychology and Theology*, 11:330–340.

Robert A. Emmons. 2006. *Spirituality: Recent progress*. Oxford University Press.

Yingqiang Gao, Fabian Winiger, Patrick Montjourides, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller, and Gerold Schneider. 2025. SpiritRAG: A Q&A system for religion and spirituality in the United Nations Archive. In *Empirical Methods in Natural Language Processing*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120:e2305016120.

Jake Goldenfein. 2019. The profiling potential of computer vision and the challenge of computational empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Gemini Team.

Qinghao Guan. 2025. Mental distress in English posts from r/AmITheAsshole subreddit community with language models. *Corpus-based Studies across Humanities*, 3:165–187.

Qinghao Guan and Melanie Nicole Lawi. 2024. An unsupervised learning study on international media responses bias to the war in Ukraine. *Corpus-based Studies across Humanities*, 1:79–97.

Christopher Helland. 2002. Surfing for salvation. *Religion*, 32:293–302.

Christopher Helland. 2005. Online religion as lived religion. Methodological issues in the study of religious participation on the Internet. *Online-Heidelberg Journal of Religions on the Internet*.

Martin Hilbert, George Barnett, Joshua Blumenstock, Noshir Contractor, Jana Diesner, Seth Frey, Sandra Gonzalez-Bailon, PJ Lamberso, Jennifer Pan, Tai-Quan Peng, Cindy Cuihua Shen, Paul E. Smaldino, Wouter Van Atteveldt, Annie Waldherr, Jingwen Zhang, and Jonathan J. H. Zhu. 2019. Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, 13:3912–3934.

Peter C. Hill, Kenneth II Pargament, Ralph W. Hood, Michael E. McCullough, James P. Swyers, David B. Larson, and Brian J. Zinnbauer. 2000. Conceptualizing religion and spirituality: Points of commonality, points of departure. *Journal for the Theory of Social Behaviour*, 30:51–77.

Kim Holmberg, Johan Bastubacka, and Mike Thelwall. 2016. "@God please open your fridge!" Twitter messages to @God in content analysis: Hopes, humor, spirituality, and profanities. *Journal of Religion, Media and Digital Culture*, 5:339–355.

Stefan Huber. 2013. *Zentralität und Inhalt: ein neues multidimensionales Messmodell der Religiosität (Vol. 9)*. Springer-Verlag.

Sophie H. Janicke and Arthur A. Raney. 2016. *Spirituality, Media, and Well-Being*. Routledge.

Seong-Hyeon Kim, Narae Lee, and Pamela Ebstyne King. 2020. Dimensions of religion and spirituality: A longitudinal topic modeling approach. *Journal for the Scientific Study of Religion*, 59:62–83.

Yueh-Ting Lee and Satoshi Kanazawa. 2015. An introduction to the special issue on the nature and evolution of Totemism, Shamanism, religions, and spirituality. *Psychology of Religion and Spirituality*, 7:265.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Ineke Noomen, Stef Aupers, and Dick Houtman. 2011. In their own image? Catholic, Protestant and holistic spiritual appropriations of the Internet. *Information, Communication & Society*, 14:1097–1117.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Rudolf Otto. 1926. *The Idea of the Holy: An Inquiry into the non-rational factor in the idea of the divine and its relation to the rational*. Oxford University Press.

Simon Peng-Keller. 2019. Genealogies of spirituality: An historical analysis of a travelling term. *Journal for the Study of Spirituality*, 9:86–98.

Qwen Team. 2025. Qwen2.5-VL Technical Report. Technical report, Qwen Team, Alibaba Group.

Martina Rahe and Jansen Jansen. 2024. What predicts well-being: connectedness to oneself, nature, others, or the transcendent? *Cogent Psychology*, 11:2371024.

Srividya Ramasubramanian. 2014. *Media and spirituality*. Routledge.

Reka Team. 2024. Reka Core, Flash, and Edge: A series of powerful multimodal language models. *arXiv*.

Tamlyn Ryan. 2012. Virtual spirituality: The negotiation and (re)-presentation of psychic-spiritual identity on the Internet. *Doctoral dissertation, University of York*.

Petra P. Sebek. 2019. *Spirituality in the selfie culture of Instagram*. Wipf and Stock Publishers.

Katarzyna Skrzypińska. 2014. The threefold nature of spirituality (TNS) in a psychological cognitive framework. *Archive for the Psychology of Religion*, 36:277–302.

LeRoy Spaniol. 2002. Spirituality and connectedness. *Psychiatric Rehabilitation Journal*, 25:321.

Yibing Sun, Varsha Pendyala, Ruixue Lian, Haohang Xin, Puja Patel, Erik P. Bucy, and Dhavan V Shah. 2025. From Internet meme to the mainstream: Using computer vision to track "Pepe the Frog" across news platforms. *Visual Communication Quarterly*, 32:33–55.

Jorge Sánchez-Garcés, Javier Linkolk López-Gonzales, Miguel Palacio-Farfán, Víctor Coronel-Sacón, Yonny Ferney-Teheran, Jahisber Peñuela-Pineda, and Himer Avila-George. 2021. Exploratory analysis of fundamental spiritual support factors to a positive attitude in patients with COVID-19 using Natural-Language Processing algorithms. *Applied Sciences*, 11:9524.

Ahmet Tanhan and Robert W. Strack. 2020. Online photovoice to explore and advocate for Muslim biopsychosocial spiritual wellbeing and issues: Ecological systems theory and ally development. *Current Psychology*, 39:2010–2025.

Gemma Team. 2025. Gemma 3 Technical Report. Technical report, Gemma Team, Google DeepMind.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv*.

Tommaso Trillò, Rebecca Scharlach, Blake Hallinan, Bumsoo Kim, Saki Mizoroki, Paul Frosh, and Limor Shifman. 2021. What does #freedom look like? Instagram and the visual imagination of values. *Journal of Communication*, 71.

Fabian Winiger, Gerold Schneider, Janis Goldzycher, David Neuhold, and Simon Peng-Keller. 2025. The 'Spiritual' and the 'Religious' in the Twittersphere: A topic model and semantic map. *Journal of Religion, Media and Digital Culture*, 14.

Bai Xue, Zhongrui Wang, Yuqing Liu, and Yao Song. 2024. Faith in frames: unveiling therapeutic narratives in religion-related cinema through computational analysis. *Frontiers in Public Health*, 12:1385379.

Jing Zeng, Qinghao Guan, Ariadna Matamoros-Fernández, and Xiran Liu. 2025. How do multimodal large language models understand non-english visual hate? insights from studying hate speech in Chinese-speaking communities on Instagram. *Platforms & Society*, 2:1–18.

Appendices

1. Annotation codebook

The topic of the document is spirituality. You are expected to annotate the connectedness type for all posts and images. In this research, we use a strict definition of spirituality:

> Spirituality is the pursuit and practice of experiences, beliefs, and values that influence and nurture the spirit, fostering personal growth, meaning, and a sense of connection to something greater than oneself.

Connectedness Type:

(1) *Connectedness to Nature*

Connectedness to nature refers to the deep sense of relationship that individuals feel with the natural world and understanding of humanity's place within the broader ecological system.

(2) *Connectedness to the Self*

Connectedness to the self includes authenticity, inner harmony/inner peace, consciousness, self-knowledge and experiencing and searching for meaning in life.

(3) *Connectedness to Others*

Connectedness to Others emphasizes empathy, compassion, and a sense of community, recognizing that interpersonal connections contribute to personal growth, a deeper understanding of oneself and others, and overall spiritual well-being.

(4) *Connectedness to Transcendence*

Connectedness to Transcendence pertains to "something or someone beyond the human level, such as the universe, transcendent reality, a higher power or God".

Note: Meditation is not necessarily connectedness to self. If something superhuman appears in it, it also belongs to connectedness to transcendence.

I. Instagram Annotation

There are two types of Instagram data: text data (Caption) and image data (image).

(1) *Text data annotation*

Text data annotation is to classify the content in the "Caption" column.

- First determine whether the caption is spirituality-related (whether it has an impact on the spirit);

- If it is spirituality-related, mark it as 1; otherwise, mark it as 0;

- If it is spirituality-related, check which connectedness type it is (if it does not belong to any of them, you can mark it as **None of them** or name it yourself, but please note that the labels you mark should be consistent throughout the entire annotation process).

- You need to additionally mark the language of the content (i.e. you can use ChatGPT or copying the caption to Google/DeepL Translate).

(2) *Image data annotation*

Image data annotators need to find the corresponding image folder based on the "Tag" column.

- Label the image as spirituality based on its content

- If it is spirituality-related content, check which connectedness type it is (if it does not belong to any of the above, you can mark it as **None of them** or name it yourself, but please make sure that the labels are consistent throughout the entire annotation process).

- You need to additionally mark the language on the image (i.e. you can use ChatGPT or copying the caption to Google/DeepL Translate).

II. Quora data annotation

The annotator needs to annotate the "Answer" column. - If it is spirituality-related content, annotate 1; otherwise, annotate 0;

- If it is spirituality-related content, check which connectedness type it is (if it does not belong to any of them, you can mark it as **None of them** or name it yourself, but please make sure that the labels you annotate are consistent).

III. Reddit data annotation

The Reddit contents that need annotating is post. The annotator needs to annotate the "post" column.

- If it is spirituality-related content, annotate 1; otherwise, annotate 0;

- If it is spirituality-related content, check which connectedness type it is (if it does not belong to any of them, you can mark it as **None of them** or name it yourself, but please make sure that the labels you annotate are consistent).

2. Agreement among annotators

We engaged five annotators to independently review and annotate a dataset ($\mathcal{N}$ = 100) previously curated by their peers. What they had to classify was the connectedness type. Then, we compiled the results and created a heatmap to visually represent the agreement and discrepancies among the annotations provided by each annotator. The figure 5 shows a great agreement among five annotators.

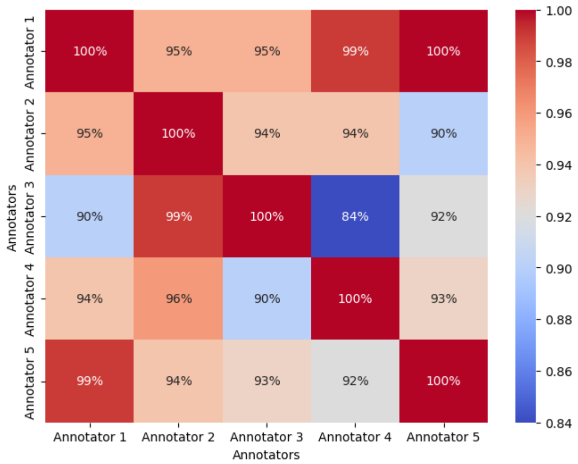


Figure 5: Heatmap of Annotation Agreement Among Annotators

3. Connectedness types across different platforms

This table displays the distribution of posts related to four types of connectedness across three social media platforms. The numbers represent the count of posts for each type of connectedness on each platform.

4. Unique representation of four types of connectedness

Given that each connectedness type has its unique terms, we did a quantitative analysis, which focuses on quantifying the data by counting the frequency of each unique term associated with different types of connectedness. Initially, the texts

Table 4: Connectedness Types Across Different Platforms

Type	Reddit Posts	Instagram Captions	Instagram Images	Quora Answers
Self	542	766	1166	436
Others	28	146	89	109
Nature	22	80	83	12
Transcendence	282	444	873	1503

were converted into its lemma form to standardize the vocabulary. Subsequently, terms specific to each connectedness category were identified by excluding words that appeared in more than one category. Each unique term was then counted to determine its frequency within its respective category. The results were displayed in horizontal bar charts.

The figure presents four separate horizontal bar charts, each representing the frequency of unique terms associated with four types of connectedness: Nature, Others, Self, and Transcendence.

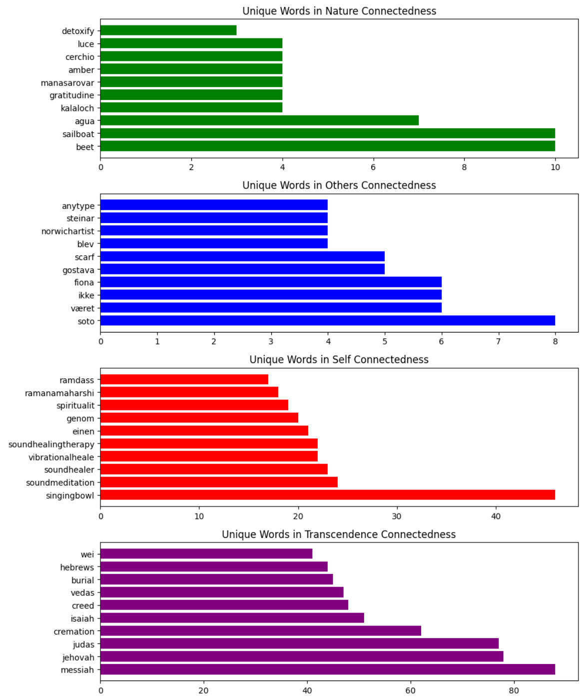


Figure 6: Unique Words in Four Types of Connectedness

The nature-related terms, such as "agua", "sailboat", "luce", and "manasarovar", suggest a focus on natural elements and experiences. As to connectedness to others, the feature words are "anytype", "steinar", and "norwichartist". Dominated by terms such as "ramdass", "spiritualiti", and "vibrationalheale", the chart of self connectedness suggests a focus on personal development, spirituality, and healing. The frequency of terms like "soundhealingtherapy" and "soundmeditation" highlights the sound-based therapeutic and meditative practices. The most populated words correlated to Transcendence like "messiah", "jehovah", and "vedas" reflect deep spiritual or religious beliefs and spiritual practices.

5. Classification results of Fine-tuning BERT

We employed the BERT model, specifically fine-tuned on our annotated dataset of Reddit posts. The fine-tuning was carried out 30 epochs, allowing the model sufficient iterations to adjust its parameters to the subtleties embedded in the dataset concerning different types of spiritual connectedness. The hyperparameter of learning rate is 2e-5, while we set the eps as 1e-8, which is a small constant added to prevent division by zero in the optimizer's calculations (AdamW optimizer), ensuring numerical stability. The figure below illustrates the trajectory of the model's accuracy over these epochs, showing the learning progress throughout the fine-tuning process. The model was uploaded on Huggingface (see `https://huggingface.co/QinghaoGuan/bert-spirituality-connectedness`).

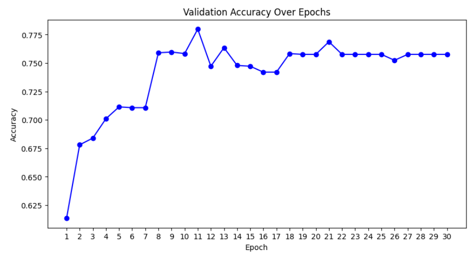


Figure 7: Validation Accuracy Curve over 30 Epochs

6. Prompt to guide LLMs to analyze topics of unclassifiable posts

Here's our prompt to guide ChatGPT-4 in assisting with topic analysis:

> I am conducting research on spirituality by categorizing posts based on four types of connectedness: to self, to others, to nature, and to the transcendence. However, some posts do not fit neatly into these predefined categories. As an assistant, you are expected to analyze the topics of these unclassifiable posts. Based on your analysis, please identify and suggest an alternative type of connectedness that these posts might represent, which is distinct from the four established categories.

7. Prompts of zero-shot and definition-injected zero-shot

Our prompts involve two settings, zero-shot and definition-injected zero-shot, and they are formulated as follows:

Zero-shot You are a human annotator and need to classify the sentence below into one of the following five labels: "connectedness to self", "connectedness to others", "connectedness to nature", "connectedness to transcendence", or "hard to classify". For each sentence, simply return self, others, nature, transcendence, or hard to classify. Sentence: "{post}"

Definition-injected zero-shot You are a human annotator and need to classify the sentence below into one of the following five labels: "connectedness to self", "connectedness to others", "connectedness to nature", "connectedness to transcendence", or "hard to classify". Now I give you the definition of four types of connectedness. {definitions}. Based on the definitions above, classify the sentence "post" into one of five labels. For each sentence, simply return "self", "others", "nature", "transcendence", or "hard to classify".

→ ***Definition given by LLMs***

As mentioned in section 2.3, the definition of spirituality is changing all the time. Given that social scientists are also interested in LLMs' understanding of spirituality, we asked Reka model to conclude the definition of spirituality based on human annotation results. Here is the answer by the Reka model.

> Based on the task of inferring definitions for the four types of "connectedness", here are my interpretations:
>
> 1. **Connectedness to Nature**: This refers to the sense of bond, harmony, and interdependence an individual feels with the natural world around them. It encompasses an emotional, spiritual, and sometimes physical relationship with ecosystems, flora, fauna, and the broader environmental context. This connectedness often involves a respect for and appreciation of natural processes and a feeling of belonging within the natural order.
>
> 2. **Connectedness to Self**: This type of connectedness involves a deep understanding, acceptance, and integration of one's own identity, values, emotions, and thoughts. It's about having a strong sense of self-awareness and self-compassion, and feeling a cohesive and authentic sense of who one is. Connectedness to self also includes being in tune with one's needs, desires, and boundaries, and living in alignment with these internal truths.
>
> 3. **Connectedness to Others**: This refers to the quality and depth of relationships and interactions an individual has with other people. It encompasses feelings of empathy, love, support, and mutual understanding with family, friends, colleagues, and even strangers. Connectedness to others involves a sense of community, belonging, and interdependence, and the ability to form and maintain meaningful bonds.
>
> 4. **Connectedness to Transcendence**: This type of connectedness involves a sense of connection to something greater than oneself, which may be understood through spiritual, philosophical, or existential lenses. It can include feelings of awe, wonder, and reverence for the universe, a higher power, or the inherent meaning and purpose of life. Connectedness to transcendence often involves a sense of unity with all things, a feeling of being part of a larger cosmic or spiritual plan, and a quest for deeper understanding and insight beyond the material realm.

8. Precision, recall, and precision of image annotation by LLMs

Model	Recall (%)	Precision (%)	F1 (%)
GPT-4o-mini	**63.42**	**65.66**	**63.99**
GPT-4o	60.46	62.87	61.34
GPT-4.1	48.42	59.68	49.83
Gemini-2.5	55.07	63.79	56.99
Claude-Sonnet	56.03	60.60	57.18
LLaVA-1.5	34.83	47.44	38.96
Qwen2.5-VL	61.94	55.76	54.74
Gemma3	43.35	52.50	34.87

This table presents the classification performance of seven models on the task of categorizing images into different types of connectedness. Overall, GPT-4o-mini achieved the highest overall performance, with a recall of 63.42%, a precision of 65.66%, and an F1-score of 63.99%. GPT-4o followed closely, with slightly lower scores across all metrics (recall = 60.46%, precision = 62.87%, F1 = 61.34%). This small performance gap suggests that while both GPT-4o and GPT-4o-mini are capable of handling complex visual-semantic classification tasks, the lightweight variant demonstrates slightly better generalization in this specific setting. Open-source small LLMs perform worse. Our findings show that scaling and fine-tuning strategies in recent GPT architectures have improved their visual-semantic reasoning capabilities, making them promising tools for computational spirituality and online religion research.

Integrating Knowledge Graphs and Multilingual Scholarly Corpora for Domain-Adaptive LLMs in SSH

Adam Faci[1], Alessio Miaschi[2], Anne Combe[3], Pascal Cuxac[4], Francesca Frontini[2], Nicolas Larrousse[1], Stéphane Pouyllau[1]
[1] Huma-Num, CNRS, Paris, France
[2]CNR - Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa, Italy
[3]Inria, France
[4]Inist, CNRS, 2 rue Jean Zay, 54500 Vandoeuvre-lès-Nancy, France
{name.surname}@huma-num.fr, {name.surname}@ilc.cnr.it
anne.combe@inria.fr, pascal.cuxac@inist.fr

Abstract

The integration of Large Language Models (LLMs) into scientific research workflows, particularly for bibliographic discovery and literature synthesis, raises significant methodological, epistemic and regulatory challenges for the Social Sciences and Humanities (SSH), especially with regard to disciplinary diversity, multilingual access to sources and the evaluation of results. This paper presents an on-going use case developed within the European project LLMs4EU and the ALT-EDIC infrastructure, aimed at adapting foundation models to SSH research practices and supporting tasks such as question answering, comparative document analysis and literature review. The evaluation framework follows the LLMs4EU protocol and encompasses both independent quantitative benchmarking (retrieval, summarisation, traceability and hallucination detection) and a qualitative assessment involving a panel of Digital Humanities experts. By embedding model adaptation within research infrastructures and a structured legal and ethical compliance framework, the use case explores how domain-sensitive and regulation-aware generative AI can support SSH scholarship while preserving reliability and epistemic responsibility.

Keywords: Scientific research workflows, Digital Humanities, Retrieval Augmented Generation

1. Introduction

In recent years, Large Language Models (LLMs) have profoundly transformed the landscape of language technologies and research practices in Social Sciences and Humanities (SSH - see, among possible examples, Arachchige et al. 2025). At the same time, the development of both LLMs and AI-enhanced scientific discovery platforms remains strongly concentrated on dominant languages and cultural contexts. It also tends to privilege epistemological models and validation practices that are typical of the so-called "hard" sciences.

This imbalance is reflected in many established and emerging AI-powered platforms for exploring scientific literature. Semantic search engines (e.g. Semantic Scholar), AI-assisted discovery tools (e.g. Elicit), large-scale citation indexes (e.g. Scopus and Web of Science), and conversational systems integrated into publishing ecosystems (e.g. Scite Assistant) are predominantly trained on large corpora of English-language journal articles. In most cases, they allow querying primarily or exclusively in English, prioritise peer-reviewed journal publications over other forms of scholarly output (such as monographs, critical editions, or research data), and rely heavily on citation-based metrics—such as impact factor or h-index—to rank and filter results.

Such design choices are not neutral. They reflect and reinforce specific models of scientific production, visibility, and authority that may not align with the diversity of practices in SSH, where multilingualism, book-based scholarship, local research traditions, and qualitative evaluation criteria remain central. As a consequence, SSH scholars working in less-resourced languages or within epistemic traditions that do not conform to citation-driven models risk being further marginalised in the emerging AI-mediated research ecosystem.

Moreover, research in the SSH often relies on interpretation and contextualisation. Texts, data, and metadata are not treated as fixed or neutral objects. Their meaning depends on research questions, theoretical perspectives, and historical context. The same material can therefore be read and organised in different ways. A well-known author such as Dante offers a simple example. A literary scholar may focus on narrative structure, poetic language, and intertextual references in the *Commedia*. A philologist may study manuscript traditions, textual variants, and editorial history. A historian may examine Dante's political thought in relation to the conflicts of medieval Italy. A Digital Humanities scholar, in turn, might build a digital scholarly edition, encode the text in TEI-XML, compare manuscript witnesses computationally, or analyse the circulation of Dante's works through network analysis and corpus-based methods. In such contexts, an algorithm cannot rely on topic matching alone, but should also recognise method-

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 139–146
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ological affinities, disciplinary perspectives, and also research methodologies.

To address these challenges, Europe has developed a rich ecosystem of data repositories and publication platforms designed to preserve the epistemic and methodological specificities of SSH. Cross-border disciplinary infrastructures such as CLARIN ERIC, DARIAH ERIC, and OPERAS have been conceived to support the shared use of digital resources while maintaining strong attention to disciplinary traditions and linguistic diversity (Branco et al., 2023; Dumouchel et al., 2020; König et al., 2023). These infrastructures have developed dedicated discovery services and meta-catalogues—such as the CLARIN Virtual Language Observatory, the SSH Open Marketplace, and the GoTriple platform—explicitly tailored to the needs of SSH communities. Such environments provide an important alternative to commercial discovery tools and could benefit from enhanced functionalities powered by LLMs. At the national level, infrastructures such as the French Huma-Num offer search facilities including Nakala and ISIDORE, which are widely used by humanities researchers[1].

Research has shown that these digital environments are not neutral repositories, but are shaped around the practices and expectations of their scholarly communities. More recent studies on research infrastructures and data archiving services further emphasise that humanities data emerge from dynamic collective processes, in which archives and service providers act not only as technical support structures, but also as mediators and co-constructors of research communities (Morselli et al., 2025). This perspective reinforces the idea that tools for discovery and analysis must remain sensitive to disciplinary practices rather than imposing external models of relevance.

Another distinctive feature of humanities research is the plurality of scholarly outputs and the central role of multilingualism as an epistemic and methodological dimension, as extensively discussed in multilingual Digital Humanities studies (Balula and Leão, 2021; Viola and Spence, 2024). Alongside journal articles, knowledge production includes datasets, curated corpora, digital editions, workflows, and forms of scholarly communication such as research blogs. The relevance of these outputs—and the need for appropriate citation, indexing, and valorisation practices—has been widely documented (Barbot et al., 2024; Mayeur, 2017). Any AI-enhanced discovery system that aims to serve SSH communities must therefore account for this diversity of formats, languages, and publication cultures.

In the context outlined above, the integration of LLM-based functionalities for bibliographic and documentary research in SSH cannot be reduced to generic applications. Language models need to be designed and evaluated in ways that reflect the specific research practices of the humanities, support interdisciplinary and transdisciplinary perspectives, and facilitate meaningful interaction among scholarly communities. The use case we propose, *ReSearch_SSH*, is a collaboration between French and Italian SSH research infrastructures within the LLMs4EU (Large Language Models for the European Union) project, funded by the European Commission and coordinated by ALT-EDIC (Alliance for Language Technologies – European Digital Infrastructure Consortium)[2]. It aims to enhance an existing search platform by integrating LLM-powered functionalities for advanced retrieval and for supporting the construction of state-of-the-art overviews grounded in curated SSH catalogs.

As noted by Fenlon (2017), Digital Humanities (DH) has witnessed the emergence of new forms of scholarly production, including research blogs, large-scale digital corpora, tools, and platforms that challenge traditional models of academic publishing. These outputs are well represented in the infrastructures mentioned above, and DH itself is a particularly dynamic research area, characterised by strong multilingualism and interdisciplinarity. For these reasons, DH constitutes a compelling testbed for the integration of LLM-based search functionalities. DH scholars, accustomed to working across methods, languages, and formats, are well positioned to act as expert users in the evaluation and co-construction of these new tools.

This paper is organized as follows. After framing the role of LLMs from the perspective of SSH research practices, we introduce the European LLMs4EU project within the broader context of the new ALT-EDIC infrastructure. We then present the ReSearch_SSH use case as a case study of model adaptation and fine-tuning for the SSH domain, focusing on data resources, methodology and finetuning strategies, and lastly on the foreseen evaluation strategies and legal issues. Finally, we suggest the involvement of research communities, especially within Digital Humanities, through the establishment of an expert panel, outlining possible forms of participation of the SSH communities in the development and evaluation of LLMs for research support.

2. ALT-EDIC and LLMs4EU

ALT-EDIC is a European Digital Infrastructure Consortium, a legal framework designed to enable several Member States to jointly develop and operate

[1] https://www.huma-num.fr/les-services-par-etapes/

[2] https://www.alt-edic.eu/

strategic digital infrastructures of common European interest. Specifically, ALT-EDIC aims to support European excellence in language technologies and to promote European linguistic diversity, following a cooperative model that brings together public institutions, industry, civil society and research. This is reflected in its close collaboration with CLARIN ERIC, the research infrastructure for language resources, and with several of its national nodes.

Established in 2024, ALT-EDIC carries out its activities through participation in several complementary European projects. Among these, OpenEuroLLM focuses on the development of open and multilingual language models, representing a key component of the European strategy for open, transparent, and value-aligned Artificial Intelligence.

Alongside this first pillar, the LLMs4EU project addresses the more applied dimension of Large Language Models across strategic sectors, adopting a strongly use-case-driven approach. The project is structured around five application domains, i.e. tourism, public services, telecommunications, energy, and science, and focuses on the fine-tuning and adaptation of open models to concrete contexts and needs. These use cases are not conceived as simple application demonstrators, but rather as experimental laboratories in which model adaptation requirements are defined, domain data are selected and prepared, fine-tuning and task-tuning strategies are tested, and evaluation protocols are designed. The overarching goal is to anchor the development of LLMs in real-world needs and in concrete communities of practice.

3. The *ReSearch_SSH* use case

Of particular interest to the European Digital Humanities community is the *ReSearch_SSH* use case, which belongs to the *Science* domain of the LLMs4EU project. It focuses on the development and experimental evaluation of generative language models adapted to advanced tasks for scholarly discovery and synthesis of scientific literature in the Social Sciences and Humanities. Particular attention is devoted to factuality, source traceability, and explainability of results. The initiative is grounded in close collaboration between major French and Italian research infrastructures,

Rather than building a new discovery system, the use case extends the existing ISIDORE platform[3], a widely used French search engine for humanities and social sciences (SSH) that aggregates multilingual scholarly resources. ISIDORE thus functions both as a large-scale documentary infrastructure and as an operational search environment.

The system follows a retrieval-augmented generation (RAG) architecture. Its retrieval layer is based on a domain-adapted model trained on historical ISIDORE queries, aligning retrieval with actual SSH search behaviour. This component is strengthened through graph-based techniques: contextualisation relies mainly on the Wikidata knowledge graph for entity disambiguation and semantic expansion, while relational structures derived from metadata support structured navigation across authors, themes, institutions, and citations. Document-level semantic graphs further enable fine-grained comparison between research outputs.

A multilingual large language model forms the generative layer, operating over the corpus indexed within ISIDORE and informed by metadata and graph-based signals. It supports natural-language querying, assisted literature synthesis, and the production of structured overviews grounded in explicitly retrieved sources. By combining adaptive retrieval with graph-aware contextualisation and generative reasoning, the system supports core SSH workflows such as literature review, thematic exploration, and comparative analysis, while ensuring that all generated outputs remain transparent and traceable to indexed documents.

The project is currently ongoing. In its present phase, it concentrates on enabling advanced querying in Italian and English over predominantly French scholarly materials aggregated within the ISIDORE ecosystem, as an initial step toward broader multilingual coverage.

Operationally, the use case is structured along four main stages.

- First, a phase of domain alignment and multilingual enrichment adapts the base model to SSH discourse using large-scale curated corpora.
- Second, retrieval and representation components are optimised in a retrieval-augmented generation (RAG) framework, leveraging ISIDORE's metadata and indexing infrastructure.
- Third, the model is fine-tuned for downstream scholarly tasks, including multilingual document retrieval, literature synthesis, and structured state-of-the-art construction.
- Finally, the system is evaluated through expert-based validation protocols designed to assess relevance, factual grounding, and usability in realistic SSH research scenarios.

In what follows we will illustrate the datasets, fine-tuning plan, and evaluation strategy.

[3] https://isidore-project.eu/

3.1. Data and disciplinary focus

The primary corpus for large-scale domain alignment is the SSH subset of the ISTEX database[4]. ISTEX aggregates more than 30 million scientific documents and provides full texts in XML/TEI format together with standardised and enriched metadata (de Salabert and Barreaux, 2020). For the purposes of ReSearch_SSH, an SSH-specific extraction (approximately 3 million documents) has been prepared based on disciplinary classifications and quality criteria (see Table 1).

Full texts are available in TEI/XML; metadata are provided in JSON format and include title, abstract, authors, affiliations, publisher, year, structured references, scientific classifications, and named entities.

Total documents	$\sim$ 3,000,000
Languages represented	60+
Full-text format	XML/TEI
Metadata license	Etalab (open)

Table 1: ISTEX SSH subset overview.

Language	Tokens (millions, approx.)
English	16,000–21,000
French	450–600
Portuguese	110–150
German	40–55
Spanish	40–50
Italian	10–15

Table 2: Estimated token distribution in ISTEX SSH subset.

Although English dominates in volume, French constitutes the main non-English component, which is particularly relevant for the ISIDORE ecosystem. The corpus therefore supports large-scale alignment to SSH discourse while preserving multilingual exposure (see Table 2).

To complement the French-dominant SSH material, targeted Italian and bilingual DH datasets are incorporated to strengthen multilingual and disciplinary alignment. These encompass the Italian Digital Humanities research papers from the annual AIUCD conference series and the Peer-reviewed Italian journal in Digital Humanities (266 articles, CC-BY license) (Tables 3 and 4).

These corpora provide high-quality DH-specific discourse, enabling: i) genre adaptation (conference papers, journal articles); ii) alignment with Italian DH terminology; and iii) cross-lingual bridging between Italian, French and English SSH traditions.

On top of this generalist alignment layer, the use case may introduce also scientific publications and

Language	Tokens (millions, approx.)
English	1.2
Italian	0.8
Total	2

Table 3: AIUCD Proceedings token distribution.

Language	Tokens (millions, approx.)
English	1.1
Italian	0.9
Total	1

Table 4: Umanistica Digitale token distribution.

academic blog posts from the Hypotheses[5] platform, together with research data and metadata from the Nakala repository[6], will be incorporated to expose the models to DH-specific practices, formats and epistemic communities.

3.2. Methodological approach and fine-tuning

The proposed approach adopts a Knowledge-Graph-based Retrieval-Augmented Generation architecture (GraphRAG; see, among others, Edge et al. 2024), designed to ensure that generated answers are explicitly grounded in the most relevant reference documents and that this grounding remains interpretable.

Model adaptation is organised in complementary stages. A first phase performs domain alignment through continued pre-training on SSH corpora in French and Italian, strengthening linguistic and terminological competence while preserving the general reasoning capacities of the base multilingual model.

A second phase distinguishes retrieval optimisation from generative task adaptation. The retrieval component is fine-tuned on historical query–interaction data from ISIDORE. Training signals are derived from user behaviour: documents effectively clicked and consulted are promoted, while systematically ignored results are ranked lower. This interaction-aware strategy aligns ranking with observed scholarly practices rather than relying solely on textual similarity.

In parallel, the generative component is instruction-tuned for research-oriented tasks within the GraphRAG pipeline. Particular emphasis is placed on the construction and extension of state-of-the-art overviews. Fine-tuning relies on examples of surveys and literature reviews combined with graph-based contextual signals encoding thematic, citation, and authorship relations. The objective is to enable the model to organise retrieved

[4] https://www.istex.fr/ressources-pour-le-fouille-de-textes/

[5] https://hypotheses.org/

[6] https://www.nakala.fr/

materials into coherent research syntheses rather than merely producing isolated summaries.

A central design constraint concerns query understanding. The system relies on a medium-scale multilingual model rather than very large reasoning-focused architectures, making robust interpretation of exploratory research queries essential. Fine-tuning therefore integrates panel-based analyses of user search practices in SSH, capturing how scholars formulate evolving and open-ended information needs.

The generation of synthetic fine-tuning data for the query understanding component will be investigated through multiple strategies, comparing both commercial models (e.g., GPT-5) and open European language models. Given the project's alignment with the OpenEuroLLM initiative and with the broader ALT-EDIC objective of promoting sovereign and transparent AI infrastructures, strong preference is given to open European models. Among these, the EuroLLM family (Martins et al., 2025), given their strong multilingual capabilities and open licensing terms. Closed commercial models will be used primarily as reference baselines to assess whether comparable data quality can be achieved through fully open pipelines, with the explicit goal of converging on a production workflow that operates entirely on open and auditable models, ensuring reproducibility and compliance with the project's governance principles. More broadly, the base models selected for domain-adaptive fine-tuning within the ReSearch_SSH use case will be drawn from the pool of open multilingual foundation models developed or endorsed within the European AI ecosystem. The final selection will be determined on the basis of multilingual coverage (with particular attention to French, Italian, and English), licensing compatibility with the intended research deployment, and baseline performance on SSH-relevant tasks. Specific model identifiers and version details will be reported in conjunction with the first experimental results.

Knowledge graphs — including those provided by Wikidata and OpenAIRE (Manghi et al., 2019), as well as the metadata and relational structures of ISIDORE and Nakala — enrich retrieval, strengthen document linking, and enhance the transparency of generated results. Structured intermediate representations may support internal reasoning, while explicit source attribution ensures traceability for end users.

An important consideration concerns the portability of the proposed architecture beyond the specific infrastructure on which it is developed. The GraphRAG pipeline can be decomposed into components with different degrees of infrastructure dependency. The core retrieval-augmented generation mechanism - comprising a retrieval model, a reranking layer, and a generative module operating over retrieved passages — is infrastructure-agnostic and can function with any document collection exposed through standard indexing interfaces. Similarly, the knowledge graph enrichment layer relies primarily on publicly available resources such as Wikidata and OpenAIRE, which are not tied to specific institutional environments and can be integrated into diverse system architectures. The most infrastructure-specific component is the interaction-aware retrieval fine-tuning, which exploits historical query–click data from ISIDORE. In the absence of comparable behavioural signals, this component can be replaced by purely semantic retrieval models or by lightweight relevance feedback mechanisms. Likewise, while the availability of full texts in TEI/XML format facilitates fine-grained document parsing and section-level retrieval, the system does not strictly require this format: plain-text corpora accompanied by structured metadata in standard formats (e.g., JSON, Dublin Core) can serve as an adequate alternative, albeit with some reduction in retrieval granularity.

3.3. Deployment and evaluation

The fine-tuned models will be deployed within a sandbox environment of the ISIDORE platform, conceived as an experimental "ISIDORE AI" workspace in which advanced research assistance functionalities can be tested in a controlled setting. The system will be integrated with the ISIDORE databases and knowledge graphs and will allow users to formulate complex queries and obtain organised and synthesised results accompanied by explicit references to the underlying sources. Multilingual querying will also be supported, potentially through the integration of a dedicated translation model.

System evaluation will be conducted in accordance with the LLMs4EU evaluation protocol, which mandates the involvement of an independent team external to the use case implementation. This team is responsible for defining realistic and scenario-based evaluation procedures grounded in the specific use case requirements and fine-tuning strategies. The independent evaluation plan, currently under development, will include quantitative assessment of retrieval performance, multi-document summarization quality, source traceability, and hallucination detection. Particular attention will be devoted to cross-lingual performance, reflecting the multilingual nature of the use case. The evaluation process may require the construction of ad hoc benchmark datasets tailored to SSH research scenarios.

A central component of the evaluation framework will consist of a structured qualitative assessment conducted by expert panels drawn from the French and Italian Digital Humanities communities. These panels will evaluate system outputs in terms of scholarly quality, epistemic reliability, methodological adequacy, and practical usefulness for research workflows. Beyond assessment, panel feedback will also inform preference optimisation processes (e.g., alignment tuning approaches), contributing to the refinement of the models so that they better reflect disciplinary standards, interpretative practices, and relevance criteria specific to Digital Humanities research.

4. Legal and regulatory compliance

ReSearch_SSH is designed in full compliance with the European regulatory framework on Artificial Intelligence, data protection, and intellectual property. In line with the governance structure of LLMs4EU, the use case is supported by both a Data Management Plan and a dedicated Legal and Ethics Compliance Assessment Plan, developed in coordination with the Legal and Ethics work package of the project. This structured governance framework ensures that data selection, model adaptation strategies, and deployment modalities are continuously assessed in light of applicable European legislation and of the principles of lawfulness, responsibility, transparency and proportionality guiding the project.

With respect to the EU AI Act, the system is conceived as a downstream modification of an existing general-purpose AI (GPAI) model. The use case does not qualify as a high-risk AI system as it operates as a research assistance tool without automated decision-making affecting individual rights. Nevertheless, transparency, documentation, and traceability requirements are treated as central design principles. In particular, the adoption of a KG-enhanced RAG architecture enforces source grounding and explicit citation mechanisms, thereby mitigating hallucination risks and strengthening epistemic reliability. Deployment is limited to a research-oriented sandbox environment within the ISIDORE infrastructure, and no unrestricted public release of licensed full texts or training data is foreseen. Human oversight is structurally embedded: system outputs are advisory, challengeable, and subject to expert validation.

From the perspective of personal data protection, the project operates in accordance with the General Data Protection Regulation (GDPR). Personal data may appear in scholarly corpora (e.g., author names, affiliations, citation metadata) and are processed only insofar as necessary for the purpose of the use case and to ensure scientific attribution, source traceability and hallucination mitigation. These data are retained until the end of the use case and are not used for profiling or behavioural modelling. Evaluation activities involving expert panels rely on informed consent procedures; panel data are limited to professional identifiers (e.g., institutional affiliation, disciplinary area) and are strictly separated from training data. In accordance with the principles of data minimisation and purpose limitation, panel data are used exclusively for evaluation and system validation, and are anonymised or deleted after the completion of the evaluation phase, though they may be made available in trusted repositories in aggregated and anonymised form.

Particular attention is devoted to copyright and database rights. The use case relies on heterogeneous data sources, including: (i) open-access SSH publications (e.g., under CC-BY or CC-BY-SA licences), (ii) metadata infrastructures released under open licences (e.g., Etalab), and (iii) licensed full texts from ISTEX, accessed under specific contractual conditions and Non-Disclosure Agreements. Licensed full texts are processed within controlled institutional environments, and retrieval-based responses are grounded in citation and excerpt linking mechanisms. The project acknowledges that the availability of content under open licences does not automatically imply unrestricted reusability for model fine-tuning, particularly when uses extend beyond the original dissemination purposes. In line with recent discussions (e.g., (Spichtinger, 2026)), tensions between open science principles and AI training practices require case-by-case legal assessment, including consideration of text and data mining (TDM) exceptions under Directive (EU) 2019/790.

Overall, ReSearch_SSH adopts a compliance-by-design approach, integrating legal analysis, risk assessment, and mitigation measures throughout the development lifecycle. This includes secure institutional storage, controlled access to licensed datasets, separation of personal identifiers from evaluation content, documentation of compute resources, bias and hallucination monitoring in coordination with the Evaluation team of the project, and explicit user disclaimers at deployment. This model reflects the broader LLMs4EU objective of developing multilingual language technologies that are not only performant, but also legally robust, ethically grounded, and aligned with European values.

5. Current state of the use case

The ReSearch_SSH use case is currently in an advanced preparatory phase. The multilingual corpus is being consolidated through the selection, harmonisation and verification of SSH datasets, including licensed and open-access resources. Subsequently, attention will devoted to metadata normalisation, knowledge graph integration and the preparation of domain-aligned subsets suitable for model adaptation. In parallel, the fine-tuning strategy, the evaluation framework and the Legal and Ethics Compliance Plan are being finalised in coordination with the relevant LLMs4EU work packages.

The first stages of domain alignment and retrieval-oriented fine-tuning are scheduled to begin shortly. These initial experiments will focus on controlled adaptation of an open foundation model to SSH scholarly discourse and on the optimisation of the KG-RAG pipeline. At this stage, particular care is taken to ensure that technical development proceeds in parallel with the independent evaluation protocol defined within LLMs4EU.

A central element currently under development is of course the constitution of an expert panel composed of scholars and professionals from the Social Sciences and Humanities, with a specific focus on Digital Humanities communities in France and Italy. Crucially evaluation activities will address not only model outputs, but the overall epistemic adequacy of the system in DH contexts.

The panel may also support experimentation with human-in-the-loop approaches and with evaluation tasks that go beyond generic benchmarks, privileging complex and context-sensitive research scenarios typical of humanities scholarship. From this perspective, the involvement of Digital Humanities communities aims to guide decisions related to data selection, annotation strategies and evaluation tasks, fostering the integration of disciplinary corpora that are representative of textual genres and research practices in the field.

6. Conclusions

This paper has presented ReSearch_SSH, a use case developed within the LLMs4EU project in the Science domain, aimed at adapting large language models to the specific epistemic and methodological requirements of Social Sciences and Humanities research. Rather than proposing yet another generic conversational interface, the initiative explores how domain-aligned models, integrated within a knowledge-graph-enhanced RAG architecture and deployed inside an existing research infrastructure (ISIDORE), can support scholarly discovery while preserving traceability, multilingualism and disciplinary specificity.

The approach adopted in ReSearch_SSH is grounded in three interconnected principles. First, adaptation is domain-driven: model alignment is guided by SSH corpora, research practices and multilingual requirements rather than by generic performance benchmarks. Second, evaluation is hybrid and independent: quantitative metrics are complemented by expert panel assessment, and evaluation protocols are defined by a team external to the use case implementation to ensure methodological robustness. Third, compliance is embedded by design: legal, regulatory and ethical considerations—ranging from GDPR and copyright to AI Act transparency obligations—are treated not as ex post constraints but as structural components of the system architecture and governance model.

At its current stage, the use case is moving from corpus consolidation and governance planning toward initial fine-tuning experiments and the constitution of a Franco-Italian expert panel. This transition marks a shift from conceptual design to empirical validation. The forthcoming phases will test whether the integration of domain adaptation, Graph-RAG grounding and participatory evaluation can effectively produce models that are not only technically performant, but also epistemically trustworthy and legally robust.

More broadly, ReSearch_SSH contributes to an emerging reflection within Digital Humanities on the role of large language models as infrastructural components of research ecosystems. By situating LLM adaptation within European open science infrastructures and within a coordinated legal and evaluation framework, the project aims to demonstrate that multilingual, domain-sensitive and regulation-compliant generative systems are both technically feasible and methodologically desirable. While the current implementation is anchored to a specific infrastructure, the modular design of the architecture is intended to facilitate transferability. Future work will investigate the deployment of individual components in institutional settings with varying levels of metadata quality and corpus structure, with the aim of identifying minimal infrastructure requirements for effective SSH-oriented language model support.

7. Acknowledgments

This work has been supported by *LLMs4EU "Large Language Models for the European Union"* project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) under the grant agreement 101198470.

8. Bibliographical References

Isuri Nanomi Arachchige, Francesca Frontini, Ruslan Mitkov, and Paul Rayson, editors. 2025. *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Ana Balula and Delfim Leão. 2021. Multilingualism within scholarly communication in ssh: A literature review. *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 12, 2, 2021*, pages 88–98.

Laure Barbot, Maja Dolinar, Edward J. Gray, Cristina Grisot, Klaus Illmayer, Michael Kurzmeier, and Barbara McGillivray. 2024. Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data*, 10(1).

Antonio Branco, Maria Eskevich, Francesca Frontini, Jan Hajič, Erhard Hinrichs, Franciska de de Jong, Pawel Kamocki, Alexander König, Krister Lindén, Costanza Navarretta, Maciej Piasecki, Stelios Piperidis, Olli Pitkänen, Kiril Simov, Inguna Skadiņa, Torsten Trippel, Andreas Witt, and Claus Zinn. 2023. The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*.

Camille de Salabert and Sabine Barreaux. 2020. Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*, pages 66–69, Nancy, France. ATALA.

Suzanne Dumouchel, Emilie Blotière, Laure Barbot, Gert Breitfuss, Yin Chen, Francesca Di Donato, Paula Forbes, Clara Petitfils, and Stefanie Pohle. 2020. TRIPLE project: building a discovery platform to enhance collaboration. *ITM Web of Conferences*, 33:03005.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Katrina S. Fenlon. 2017. *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities*. text, University of Illinois at Urbana-Champaign.

Alexander König, Laure Barbot, Cristina Grisot, Michael Kurzmeier, and Edward J. Gray. 2023. The SSH Open Marketplace and CLARIN. *Proceedings of the CLARIN Annual Conference*.

Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, and Pedro Principe. 2019. The OpenAIRE Research Graph Data Model. report.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Ingrid Mayeur. 2017. Imparting Knowledge in Humanities. About Some Practices of Scientific Blogging on Hypothèses. In *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*, pages 75–84. IOS Press.

Francesca Morselli, Jetze Touber, and Andrea Scharnhorst. 2025. Fostering Data Communities – perspective from a Data Archive Service Provider. ArXiv:2502.02321 [cs].

Daniel Spichtinger. 2026. *Perspective Chapter: Fit for Purpose? Creative Commons Licensing for Research Data in the Age of Artificial Intelligence*. IntechOpen.

Lorella Viola and Paul Spence, editors. 2024. *Multilingual Digital Humanities*. Routledge, London New York.

From One-Hot to Semantic Encoding: Entity Embedding for Small and Heterogeneous Digital Humanities Datasets

Isabelle Gribomont
CENTAL/UCLouvain
Place de l'Université 16
1348 Louvain-la-Neuve, Belgique
isabelle.gribomont@uclouvain.be

Abstract

This paper investigates the use of semantic encoding for the analysis of heterogeneous digital literature metadata. Drawing on two databases of Latin American digital literature, *Archivo de Literatura Digital en América Latina* and the *Atlas da Literatura Digital Brasileira*, we compare traditional one-hot encoding with a semantically enriched representation derived from feature-value descriptions embedded in a continuous vector space. In contrast to one-hot encoding, which treats categorical values as orthogonal, semantic encoding models accounts for similarity between features, thereby mitigating vocabulary mismatch across databases. We evaluate both approaches using between-group centroid distances, and normalized centrality measures. Our results show that semantic encoding clarifies structural differentiation across genres and might smooth arbitrary differences introduced by differing vocabularies across databases. The findings suggest that semantic representations provide a more interpretable embedding space for small and taxonomically heterogeneous datasets. Beyond technical performance, the study suggests that embedding-based methods can support critical inquiry in digital humanities, enabling the examination of database bias, categorical patterns, and diachronic evolution within a unified semantic framework. Code is available at https://github.com/isag91/semantic-encoding-DH.

Keywords: Categorical Data, Entity Embeddings, Digital Humanities

1. Introduction

Digital humanities research increasingly relies on structured databases to document, classify, and analyze cultural production. In the field of digital literature, initiatives such as the *Archivo de Literatura Digital en América Latina - ADLAL* (Archive of Digital Literature in Latin America) and the *Atlas da Literatura Digital Brasileira - ALDB* (Atlas of Brazilian Digital Literature) provide curated metadata describing works according to genre, publication format, technical requirements, hardware dependencies, artistic techniques etc. (Botero Betancur, 2023; Athayde and ROCHA, 2022). These databases are invaluable for scholarly inquiry, yet they pose methodological challenges.

First, they are relatively small. Digital humanities datasets often contain hundreds rather than thousands of entries. Therefore, patterns of category co-occurrence are often insufficient to reliably infer semantic relationships between labels (e.g. the close conceptual relationship between *hypertext* and *hypermedia*). Second, they are heterogeneous: distinct databases use different vocabularies, taxonomies, and degrees of granularity. Third, many features are multi-valued (e.g., a work may require a computer and loudspeakers to be consulted).

These difficulties limits the ways in which these datasets can be meaningfully leveraged within the field. One-hot encoding treats categorical values as equidistant, ignoring semantic relationships between them. This limitation creates what has been termed a semantic gap in categorical representation (Yang et al., 2026). Likewise, when different databases refer to similar practices using distinct terminologies, their underlying semantic proximity is obscured.

Recent work has proposed leveraging generative large language models (LLMs) to enrich categorical representations with external semantic knowledge and/or transformer semantic encoder models to represent categorical descriptions. The potential of these approaches for humanities datasets—characterized by sparsity, conceptual overlap, and evolving vocabularies—remains underexplored.

This paper investigates whether semantic encoding of categorical metadata can improve geometric representation of data points in the semantic space in the context of two digital literature databases. To do so, we compare a traditional one-hot baseline with a semantic encoding approach in which feature-value labels are described using a large language model and embedded using a pre-trained sentence transformer.

Our central hypothesis is that semantic encoding offers methodological advantages for humanities datasets in at least three respects:

1. It captures conceptual proximity between labels that are treated as independent symbols in one-hot

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 147–152
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

`representations.`

2. `It mitigates sparsity by importing external semantic knowledge.`
3. `It facilitates cross-database comparison despite divergent terminologies.`

In the case of the two digital literature databases examined here, a semantically informed encoding could enable analyzes that would otherwise be distorted by heterogeneous vocabularies and the absence of explicit relationships between categorical values. Such an approach would make it possible to investigate whether certain databases occupy only specific subregions of the semantic space; whether works by authors of different genders or countries of origin are distributed differently across this space of practices; and whether diachronic patterns can be observed, such as the emergence of new regions over time.

2. Related Work

Traditional one-hot encoding represents categorical variables as orthogonal vectors, implicitly assuming that all categories are equally dissimilar. In machine learning, this limitation has motivated the development of learned embeddings for categorical variables, particularly in tabular data contexts (Guo and Berkhahn, 2016; Gorishniy et al., 2021). These methods learn semantic relationships from co-occurrence patterns within the dataset. Such inferences require a large amount of data to be reliable.

Large language models (LLMs) trained on large-scale corpora encode extensive distributional knowledge and have demonstrated strong performance across a wide range of semantic tasks (Brown et al., 2020; Devlin et al., 2019; Feng et al., 2024). Beyond text generation, such models can serve as structured semantic resources capable of producing definitions, contextual descriptions, and attribute-level interpretations. Recent work has highlighted the potential of LLMs as general-purpose knowledge interfaces (Bommasani et al., 2021; Petroni et al., 2019; AlKhamissi et al., 2022).

Building on these observations, recent approaches have proposed to leverage LLMs as external knowledge bases to semantically enrich datasets, especially in cases where intra-dataset co-occurrence signals are too sparse to infer meaningful relationships between categorical values. By querying LLMs at the feature–value level and generating structured descriptions or embeddings, it becomes possible to inject semantic information that is not recoverable from the data alone (Yang et al., 2026; Huesmann and Linsen, 2025; Hegselmann et al., 2023).

Within digital humanities, embedding-based methods have primarily been applied to textual corpora, supporting large-scale stylistic, thematic, or historical analysis (Underwood, 2019; Piper, 2018). However, comparatively less attention has been devoted to the modeling of metadata structures themselves. By leveraging semantically enriched embeddings at the metadata level, this study extends vector-space modeling beyond textual analysis and into the domain of structured descriptive information. This method has the potential to improve the investigation of database bias, categorical patterns, and diachronic shifts within a unified semantic framework, particularly in small and heterogeneous datasets where statistical and deep learning methods relying on within-dataset co-occurrence may be limited.

3. Methodology

3.1. Data Sources

We use two curated databases of digital literature. The Archivo de Literatura Digital en América Latina indexes works from Latin America, except Brazil, and contains 180 items. The Atlas da Literatura Digital Brasileira indexes works from Brazil and contains 149 items. Both databases describe digital literature using very similar metadata fields such as *genre*, *publication type*, *access hardware*, *technical requirements* etc. The databases differ slightly in taxonomy and naming conventions. To enable comparison, feature names were mapped to a shared schema in English, but differences in the vocabulary itself were maintained, besides translation.

Missing values were encoded using an explicit *no_information* label to preserve structural consistency rather than discarding incomplete entries.[1]

3.2. Semantic Description of Feature-Value Pairs

Following a similar approach to the ARISE framework (Yang et al., 2026), for each unique feature-value pair (e.g., *genre_poetry*, *publication_type_software*), we generated a structured description using ChatGPT-5.3.[2] Each description followed a fixed template: [CORE]: general definition of the value. [INDICATOR]: what this value indicates in the context of digital literature.[3]

[1]The data, LLM prompt and code are available at `https://github.com/isag91/semantic-encoding-DH`

[2]GPT-5.1 was found to offer the best performance compared to Claude Opus 4.5, DeepSeek V3.2 and Gemini 3 Pro by Yang et al. (2026).

[3]Examples: *access_hardware_computer*: [CORE] A computer is an electronic device that processes data and

A potential limitation of LLM-based semantic encoding lies in the stochastic nature of text generation, which may introduce variability in the resulting representations. To mitigate this issue, descriptions were generated at the feature-value level rather than per instance, ensuring that identical categorical values are consistently mapped to the same semantic representation. In addition, structured prompting was employed to constrain the form and content of the generated descriptions, reducing variability and emphasizing discriminative information. Finally, descriptions generated by a LLM could also be replaced by expert definitions. Once the descriptions are set, the semantic encoding pipeline is stable and deterministic.

3.3. Embedding Strategy

Once structured descriptions were generated for each unique feature-value pair, these texts were transformed into dense vector representations using the pre-trained sentence transformer *all-mpnet-base-v2* used in Yang et al. (2026). To convert variable-length text into fixed-dimensional vectors, a pooling strategy is adopted to weight tokens according to their activation levels. Following (Yang et al., 2026), we extract the last hidden state of the transformer, which provides a contextualized embedding for each token. A scalar activation score is computed for each token as the mean across embedding dimensions, and these scores are normalized with a softmax to produce attention weights. The embedding is computed as an attention-weighted sum.

When a work contained multiple values for the same feature, their corresponding embeddings were averaged, resulting in one vector per feature for each work. After computing feature-level embeddings, vectors corresponding to all features of a given work were concatenated to form a unified semantic representation. Concatenation was chosen over summation to preserve feature-specific structure and avoid conflating semantically distinct metadata dimensions. In this way, each feature occupies a stable subspace within the overall representation.

runs programs, typically with a screen, keyboard, and mouse. [INDICATOR] This indicates the work is designed to run on a desktop or laptop system and may require precise cursor control, large displays, or locally executed software. *genre_poetry*: [CORE] Poetry is a literary form that emphasizes rhythm, sound, and condensed language. [INDICATOR] This indicates the work focuses on expressive language, structure, or visual arrangement rather than linear storytelling.

3.4. Baseline: One-Hot Encoding

As a baseline representation, categorical values were encoded using one-hot vectors. When a work was associated with a single value for a given feature, the corresponding dimension was set to one and all others to zero. In the case of multi-valued features (for example, a work categorized under multiple techniques), a multi-hot representation was used in which several dimensions could simultaneously take the value one. All feature-level vectors were subsequently concatenated to form a single high-dimensional sparse vector representing each work.

As one-hot encoding ignores conceptual relationships between categories, it provides a useful control condition against which to evaluate the contribution of semantic enrichment.

4. Comparative Analysis

To assess the representational differences between one-hot and semantic encoding, we do not rely on supervised clustering evaluation measures. Such metrics presuppose the existence of a ground-truth partition against which clustering results can be evaluated. This assumption does not hold in our context. Instead, as our objective is to determine whether semantic encoding yields a more meaningful representation, we adopt an intrinsic evaluation approach based on the geometry of the resulting vector spaces, informed by domain knowledge.

First, we computed centroid vectors for works associated with the genre labels *poetry*, *narrative*, and *poetry_and_narrative*. We use genre as a primary analytical category because, as a broad and conceptually encompassing dimension, it is expected to occupy more distinct and structurally differentiated regions in the semantic space than more specific features such as hardware requirements or reading processes. Moreover, we expect *poetry* and *narrative* to be more distant from one another than *poetry_and_narrative* is from either of them. In this sense, genre offers a simple but meaningful relational structure against which the ability of the representation to capture graded semantic proximity can be assessed.

For each label, the centroid was obtained by averaging the representations of all works assigned to that category. Pairwise cosine distances between these centroids were then measured in both the one-hot and semantic spaces (see table 1). In the one-hot representation, the distances between the works assigned with these three labels are uniform and fail to reflect conceptual overlap. In contrast, preliminary results in the semantic space indicate that *poetry_and_narrative* occupies an intermediate position. This suggests that semantic information

Label A	Label B	OH	Sem
poetry	narrative	0.184	0.105
poetry	poetry_narrative	0.175	0.058
narrative	poetry_narrative	0.182	0.064

Table 1: Pairwise cosine distances between genre centroids. The matrix captures the geometric organization of genres within the embedding space, with smaller distances reflecting closer semantic alignment.

is encoded in the embedding space in a manner consistent with interpretive intuition.

Second, we used database-level centroids to determine whether semantic encoding reduces artificial separation caused by divergent terminologies. The working hypothesis was that one-hot encoding may exaggerate differences due to mismatched vocabularies, whereas semantic representations may align conceptually similar categories even when labels differ, therefore allowing for more graded similarities to emerge. In principle, in the case of these two specific databases, there should not be any significant semantic differences between them, since they cover the same literary forms, unless Brazilian artists have distinct practices to other Latin American artists.

Because one-hot and semantic encodings differ substantially in their geometric properties, raw centroid distances are not directly comparable across representations. One-hot encoding produces sparse, orthogonal vectors that tend to inflate distances, whereas semantic embeddings generate denser, continuous spaces in which distances are typically compressed. To avoid conflating representational scale with structural differentiation, we therefore rely on normalized measures of centrality and displacement.

First, we computed the ratio to global dispersion, i.e. the cosine distance between a database centroid and the global centroid of the embedding space, divided by the mean distance of all works to the global centroid. Because the measure is normalized by overall dispersion, it allows comparison across embedding spaces that differ in scale. Lower values indicate a more central position of the database relative to the overall space.

Second, we computed the relative offset, which compares the distance between a database centroid and the global centroid to the average internal dispersion of that database (i.e. the mean distance of works within the database to their own centroid). Lower values indicate that the database does not form a distinct cluster and spread across the whole space (see table 2).

Both one-hot and semantic encoding position the databases close to the global centroid, indicating that neither database forms a strongly displaced subregion of the representational space. However, semantic encoding slightly reduces both measures, which aligns with the intuition that it should smooth the artificial orthogonality of one-hot encoding.

DB	RGD_OH	RGD_S	RO_OH	RO_S
ALDAL	0.079	0.065	0.087	0.070
ALDB	0.122	0.097	0.130	0.104

Table 2: Normalized centrality of database centroids in one-hot and semantic embedding spaces. The ratio to global dispersion (RGD) quantifies relative centrality within each representational geometry, while the relative offset (RO) controls for internal database heterogeneity. Together, these metrics distinguish structural displacement from global scale effects introduced by different encoding strategies.

5. Discussion and Conclusions

The purpose of this study is to examine semantic encoding as a methodological intervention in digital humanities research. Humanities datasets are often small, curatorially constructed, and conceptually heterogeneous. In such contexts, representation is not merely a technical preprocessing step but an epistemological decision that shapes analytical outcomes.

One-hot encoding reflects a structuralist view of categorical metadata in which labels function as discrete and unrelated symbols. In doing so, it suppresses nuance, hybridization, and gradience. Semantic encoding, by contrast, embeds categorical labels within a continuous vector space informed by external linguistic knowledge. As shown by the genre experiment, this transformation enables composite categories to occupy intermediate positions and allows conceptually related labels to cluster together. It also smooths the artificial differences brought in by the use of different terminologies, as suggested by the database experiment.

By incorporating semantic embeddings derived from large language models, we introduce a representational layer that captures conceptual similarity beyond local statistical evidence. In doing so, we mitigate sparsity effects, facilitate cross-database comparison, and create a space in which interpretive proximity can be examined quantitatively.

Ultimately, semantic encoding should be understood as a representational strategy that complements, rather than replaces, symbolic metadata. It provides a way to bridge the gap between qualitative interpretive categories and quantitative analytical methods, allowing digital humanities researchers to explore conceptual structure without discarding nuance.

Future research should extend both methodology and the analysis enabled by semantic encoding. We followed the ARISE methodology regarding the attention-weighted pooling, as well as the choice of LLM and encoding model. We could investigate other options, as well as combine semantic encoding with insights from co-occurence patterns, as proposed in Yang et al. (2026). In addition, regarding the evaluation, while centroid-based measures offer an interpretable first approximation of structural organization, additional diagnostics could be implemented. For instance, neighborhood-based metrics, such as k-nearest-neighbor, could further assess whether semantic encoding enhances local conceptual coherence.

In the context of this specific use case, the semantically aware representations made possible by semantic encoding open several future research avenues, not only as an exploratory tool but as a methodological framework for investigating both the structure of the field of digital literature and the practices of database curation. On the one hand, such representations can be used to address research questions about the field itself. Diachronic analyses may track the evolution of practices over time, while demographic metadata (e.g., gender or country of origin of the authors) could be mapped onto the embedding space to explore patterns of diverging practices. On the other hand, they provide a means to critically assess database construction by identifying representational biases, such as whether certain databases disproportionately foreground particular types of works or authors while neglecting others.

Extending this methodology to the numerous digital literature databases which exists would make it possible to move toward a more comprehensive, bird's-eye view of the field. However, such comparisons would likely require methodological adaptations beyond those presented here, as differing feature sets and levels of granularity introduce additional challenges for alignment and semantic integration.

Finally, combining geometric analysis with qualitative inspection of representative works would help ground embedding-based findings in close reading, reinforcing the interpretive relevance of semantic encoding within digital humanities research.

6. References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Manaíra Aires Athayde and REJANE ROCHA. 2022. Um arquivo para a literatura digital brasileira e algumas questões concretas. *RE-AUTO-META ARQUIVO*, page 84.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Constanza Botero Betancur. 2023. Gainza, c. y zúñiga, c. cartografía de la literatura digital latinoamericana. visualización, archivo y preservación de obras literarias digitales, 2018-2021. *publicaciones*, 4:e054.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 34, pages 1877–1901.

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.

Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, and Kam-Fai Wong. 2024. LLMEdgeRefine: Enhancing text clustering with LLM-based boundary point refinement. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18455–18462, Miami, Florida, USA. Association for Computational Linguistics.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 35, pages 18932–18943.

Cheng Guo and Felix Berkhahn. 2016. Entity embeddings of categorical variables. *CoRR*, abs/1604.06737.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Karim Huesmann and Lars Linsen. 2025. Large language models for transforming categorical data to interpretable feature vectors. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5754–5771.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Zihua Yang, Xin Liao, Yiqun Zhang, and Yiu ming Cheung. 2026. Bridging the semantic gap for categorical data clustering via large language models.

Design and methodological architecture of a multilingual corpus of interpreter-mediated public service telephone interactions

Daniel López Padilla, Raquel Lázaro Gutiérrez
Universidad de Alcalá, FITISPos-UAH, Nodo CLARIAH-CM
daniel.lopezp@uah.es , raquel.lazaro@uah.es

María José Vilella Sánchez, Fernando Espinoza Cuadros, Jorge Rico Ródenas
Sigma Cognition
mjvilella@sigmacognition.ai, fespinoza@sigmacognition.ai, jrico@sigmacognition.ai

Abstract

Multimodality in Social Sciences and Humanities (SSH) research is often associated with the integration of text and visual data. However, interpreter-mediated telephone interaction presents a different configuration of complexity, where acoustic, temporal, discursive, and pragmatic dimensions converge. This paper presents the design and methodological architecture of PRAGMACOR(Corpus Pragmatics and Telephone Interpreting: Analysis of Face-Threatening Acts, Ref. PID2021-127196NA-I00), a multilingual corpus of interpreter-mediated public service telephone interactions (Chinese–Spanish, English–Spanish, French–Spanish, German–Spanish), as a case study in multimodal and plurilingual SSH infrastructure.

The corpus integrates aligned audio recordings, orthographic transcriptions enriched with speech phenomena, temporal segmentation into speech acts, and multilayer pragmatic annotation of Face-Threatening Acts (FTAs), validated through a structured double-annotation and expert review process. Beyond textual data, the infrastructure captures prosodic overlap, turn-taking dynamics, and pragmatic mediation, enabling the study of cross-linguistic transfer and relational negotiation in asymmetrical institutional contexts. Datasets such as PRAGMACOR have proved essential to train LLMs for speech to speech translation (Sakai et al., 2024).

Attention is given to the ethical and technical design of the corpus, including local automatic transcription, systematic removal of personal identifiable information, and irreversible voice anonymization through spectral and temporal signal transformation. These procedures ensure both research usability and compliance with responsible data governance principles.

By conceptualising interpreter-mediated interaction as an acoustic-discursive multimodal object and plurilingual pragmatic process, this paper argues that PRAGMACOR provides a replicable model for the development of SSH-oriented infrastructures capable of supporting advanced research in multilingual communication, discourse analysis, and future evaluation of language technologies.

.

Keywords: Multimodal corpus design, Pragmatic annotation, Pragmatics, Multilingual Corpus

1. Introduction

Multimodality has become a central concern in Social Sciences and Humanities (SSH) research, particularly in fields engaging with digital archives, visual culture, and computational text analysis. In many cases, multimodality is conceptualised as the integration of text and image, or more broadly, the combination of heterogeneous data formats within digital infrastructures. However, less attention has been paid to forms of multimodality that emerge in spoken institutional interaction, especially in settings where communication is mediated across languages. Interpreter-mediated telephone encounters constitute a distinctive multimodal environment in which acoustic signals, temporal organisation, discursive structure, and pragmatic negotiation interact in complex ways.

Unlike face-to-face communication, telephone interaction lacks visual cues, yet it remains deeply multimodal. Meaning is co-constructed not only through lexical content but also through prosodic features, overlapping speech, turn-taking dynamics, and sequential organisation. When such interaction is mediated by an interpreter, an additional layer of complexity arises: linguistic transfer is intertwined with pragmatic mediation. Institutional asymmetries, face management, and relational positioning are negotiated across languages in real time, making these interactions particularly rich sites for examining multilingual and cross-cultural communication.

Despite this complexity, many multilingual speech corpora remain primarily text-oriented in their analytical design. Audio recordings are often treated as sources for transcription rather than as integral components of a layered multimodal object. Furthermore, the pragmatic dimension of interaction—such as the management of Face-Threatening Acts (FTAs), mitigation strategies, and relational effects—tends to be underrepresented in corpus infrastructures, even

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 153–157
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

though it plays a crucial role in institutional settings.

This paper presents PRAGMACOR, a multilingual corpus of interpreter-mediated public service telephone interactions (English–Spanish, French–Spanish, German–Spanish, and Chinese–Spanish), as a case study in the design of a multimodal and plurilingual research infrastructure. The corpus integrates aligned audio recordings, temporally segmented transcripts, multilayer pragmatic annotation, and rigorous validation procedures, all within a framework that prioritises ethical data governance and long-term reusability. Rather than conceptualising multimodality as the juxtaposition of separate media, PRAGMACOR operationalises it as the structured integration of acoustic, temporal, discursive, and pragmatic layers.

By foregrounding both multimodal architecture and plurilingual mediation, this paper argues that interpreter-mediated telephone interaction requires a reconceptualization of multimodal corpus design in SSH research. In doing so, it proposes an infrastructure model that is methodologically robust, ethically grounded, and adaptable to future developments in multilingual language technologies.

2. Conceptualising Multimodality in Interpreter-Mediated Telephone Interaction

Multimodality in SSH research has frequently been associated with the coexistence of distinct semiotic systems, most prominently text and image, within digital environments (Kress & van Leeuwen, 2006; Jewitt, 2014). These approaches have been particularly influential in media studies and visual communication. However, spoken institutional interaction—especially in telephone settings—calls for a broader understanding of modality. Rather than focusing on visual–textual combinations, interpreter-mediated telephone encounters reveal multimodality as the integration of acoustic, temporal, discursive, and pragmatic layers within a single interactional event.

From an acoustic perspective, prosodic variation, pauses, hesitations, speech rate, and intonation contours contribute to the construction of stance, affect, and relational positioning. Overlapping speech, which is frequent in institutional telephone communication, signals alignment, interruption, urgency, or resistance. These features are not peripheral embellishments to linguistic content; they are constitutive elements of interactional meaning.

At the temporal level, segmentation into speech acts and the organisation of turn-taking shape how institutional authority, accountability, and responsibility are negotiated. Telephone interaction unfolds in tightly structured sequences, where requests, clarifications, confirmations, and directives are embedded within institutional routines. When mediated by an interpreter, each segment is re-articulated across languages, introducing shifts in timing, emphasis, and pragmatic force.

The discursive layer further complicates this configuration. Interpreter-mediated interaction is not a linear transfer of propositional content but a process of multilingual recontextualisation. Speech acts produced in one language are reformulated in another within an asymmetrical institutional framework. This reformulation may preserve, attenuate, or intensify pragmatic force, affecting how Face-Threatening Acts (FTAs) are perceived and negotiated by participants.

In this sense, multimodality in interpreter-mediated telephone interaction is not reducible to multiple media formats; it emerges from the structured interplay of acoustic signal, temporal alignment, discursive segmentation, and pragmatic annotation. Capturing this interplay requires an infrastructure capable of integrating audio, transcription, alignment, and multilayer analytical categories within a single environment.

PRAGMACOR addresses this need by treating audio recordings not merely as sources for textual transcription, but as integral components of a multimodal research object. Through time-aligned segmentation and layered annotation, the corpus preserves the interactional architecture necessary for analysing how multilingual pragmatic mediation unfolds in institutional contexts. This design enables researchers to move beyond text-centric corpus analysis and engage with the dynamic, temporally situated nature of spoken communication.

3. Multilingual and Multilayered Corpus Architecture

PRAGMACOR has been designed as a multilingual and pragmatically annotated corpus of interpreter-mediated public service telephone interactions. The corpus currently includes interactions involving four language pairs—English–Spanish, French–Spanish, German–Spanish, and Chinese–Spanish—reflecting the linguistic diversity of institutional settings where telephone interpreting is deployed. This plurilingual configuration does not merely expand linguistic coverage; it enables the systematic study of cross-linguistic pragmatic mediation in comparable institutional scenarios.

3.1 Multilingual design and validation procedures

The corpus was developed through a structured annotation workflow involving professional transcribers and annotators who are Spanish speakers and native speakers of the respective foreign language. Each language pair was supported by a dedicated team responsible for transcription, annotation, and review.

To ensure methodological robustness, the annotation process followed a double-annotation protocol: each interaction was independently annotated by two bilingual annotators. In cases of discrepancy, annotations were merged through a semi-automated procedure that displayed divergent labels for adjudication. A third bilingual annotator validated the final decision, and an expert linguist supervised the process—particularly during the initial phases—to reinforce shared criteria and resolve interpretative ambiguities.

This layered validation structure reflects the interpretative nature of pragmatic annotation, where categorisation depends not only on formal linguistic features but also on contextual judgement and theoretical grounding (Ide & Pustejovsky, 2017). The use of double annotation and adjudication procedures aligns with established practices in corpus linguistics aimed at ensuring analytical reliability while acknowledging interpretative variability (Artstein & Poesio, 2008).

3.2 Multilayer annotation and segmentation

All transcription and annotation tasks were carried out using EXMARaLDA, a tool that enables the definition of multiple annotation tiers aligned to the same temporal axis. Each audio file was automatically converted into a structured template with predefined speaker roles and language layers. This configuration allowed annotators to segment speech into temporally bounded units corresponding, as far as possible, to single speech acts.

Segmentation plays a crucial role in preserving the interactional architecture of telephone communication. Annotators were instructed to avoid including more than one speech act per segment and to limit segment duration, thereby facilitating subsequent pragmatic annotation. Because each speaker occupies a separate channel, overlapping speech can be represented explicitly, preserving the simultaneity characteristic of institutional telephone interaction.

Beyond orthographic transcription—enriched with features such as false starts and filled pauses—the corpus incorporates multilayer pragmatic annotation, including the identification and classification of FTAs. These annotations are anchored to specific time-aligned segments, ensuring that pragmatic interpretation remains tied to the acoustic and sequential context in which it occurs.

3.3 Integration of automatic processing and ethical safeguards

To optimise efficiency while maintaining data security, automatic transcription was performed locally using a Whisper model deployed on internal servers, preventing the transfer of sensitive audio data to third-party platforms. Automatically generated transcripts were subsequently revised by human annotators, reinforcing accuracy and contextual interpretation.

Ethical design was central to the corpus architecture. All personal identifiable information (PII) was systematically replaced with dedicated tags in the transcripts. To ensure full anonymisation, a two-stage process was implemented at the audio level. First, time-aligned segments containing personal data were replaced with silence. Second, an irreversible voice anonymisation pipeline was applied, incorporating spectral and temporal signal transformations—including fundamental frequency modification and controlled perturbation—to prevent biometric re-identification.

By integrating automatic processing, multilayer annotation, bilingual validation, and irreversible anonymisation within a unified workflow, PRAGMACOR operationalises a model of multilingual and multimodal corpus design that balances methodological precision, interpretative depth, and responsible data governance.

4. Plurilingual Pragmatic Mediation and Analytical Potential

Interpreter-mediated institutional interaction is not merely multilingual communication; it is a process of plurilingual pragmatic mediation unfolding within asymmetrical institutional frameworks. In such settings, linguistic transfer is inseparable from relational negotiation. Requests, directives, explanations, and justifications are reformulated across languages in real time, often under conditions of urgency or vulnerability. Capturing this complexity requires analytical categories that move beyond lexical equivalence and address the management of face, authority, and accountability.

Face-Threatening Acts (FTAs), understood as acts that potentially challenge a participant's public self-image (Brown & Levinson, 1987), are particularly salient in institutional encounters, where obligations, refusals, procedural constraints, and evaluative statements are frequent. The annotation of FTAs within PRAGMACOR provides a structured entry point into this dimension. FTAs, understood as acts that potentially challenge a participant's public self-image, are particularly salient in institutional encounters, where obligations, refusals, procedural constraints, and evaluative statements are frequent. In interpreter-mediated telephone interaction, such acts are not simply reproduced across languages; they are re-articulated. The interpreter's rendition may preserve, mitigate, intensify, or subtly reframe the pragmatic force of the original utterance.

This plurilingual re-articulation has several analytical implications. First, it enables the study of cross-linguistic pragmatic transfer: how degrees of imposition, politeness strategies, or mitigation devices shift when speech acts move between linguistic systems. Second, it allows for the examination of relational positioning within institutional asymmetries. Public service encounters often involve unequal access to knowledge, authority, or resources; the interpreter's mediation can influence how these asymmetries are enacted discursively.

Because PRAGMACOR integrates time-aligned audio, segmented speech acts, and multilayer pragmatic annotation, researchers can examine these processes at different levels of granularity. For instance, it becomes possible to trace how a directive produced in one language is segmented, reformulated, and sequentially embedded in the target language. Overlaps, hesitations, or prosodic cues can be analysed alongside pragmatic labels, preserving the interactional context in which mediation occurs.

The plurilingual design of the corpus further expands its analytical scope. By including multiple language pairs involving Spanish as a common institutional language, PRAGMACOR facilitates comparative research across linguistic and cultural configurations. Patterns of mitigation, directness, or face management can be examined not only within individual interactions but also across language pairs, contributing to broader discussions on multilingual communication and cultural diversity in institutional contexts.

Importantly, this analytical potential extends beyond traditional discourse analysis. The structured, validated annotation framework positions the corpus as a foundation for future interdisciplinary research, including the evaluation of computational tools aimed at modelling multilingual pragmatic phenomena. While the present paper focuses on infrastructural design, the layered architecture of PRAGMACOR makes it adaptable to emerging methodologies in multilingual language technologies, without reducing its interpretative depth.

By conceptualising interpreter-mediated communication as plurilingual pragmatic mediation, PRAGMACOR demonstrates how multimodal corpus design can capture the relational and interactional dimensions of multilingual SSH research. Rather than treating language transfer as a neutral conduit of information, the corpus foregrounds the dynamic reconfiguration of meaning across languages and institutional roles.

5. Ethical, Infrastructural and Methodological Implications for SSH Research

The design of multilingual and multimodal corpora in institutional contexts entails specific methodological and ethical challenges. Telephone interpreting in public service settings frequently involves sensitive personal information, vulnerable populations, and asymmetric power relations. As a result, corpus development cannot be approached solely as a technical task; it must be embedded within a framework of responsible data governance and long-term research sustainability.

PRAGMACOR addresses these challenges through a layered anonymisation strategy that operates at both textual and acoustic levels. Personal identifiable information is systematically replaced with dedicated tags in transcripts, while audio segments containing such information are removed through time-aligned silencing. Crucially, voice anonymisation is implemented through irreversible spectral and temporal transformations of the signal, preventing biometric re-identification while preserving interactional properties relevant to research. This dual approach allows the corpus to remain analytically usable without compromising participant privacy.

From an infrastructural perspective, the integration of local automatic transcription tools with human validation procedures reflects a balanced model of technological support and expert interpretation. Automatic speech recognition accelerates processing while remaining embedded within a workflow that prioritises contextual judgement and theoretical grounding. Similarly, multilayer annotation in a time-aligned environment ensures that pragmatic categories remain anchored to the acoustic and sequential realities of interaction.

Methodologically, the corpus illustrates the importance of treating annotation not as a neutral labelling exercise but as an interpretative practice requiring training, calibration, and adjudication. The double-annotation and review system implemented in PRAGMACOR underscores the complexity of pragmatic classification, particularly when applied across languages and institutional contexts. This design foregrounds transparency and reproducibility by making the decision-making process traceable and structured.

Beyond its immediate analytical applications, the infrastructure proposed here has broader implications for SSH research. First, it offers a replicable model for integrating acoustic, temporal, discursive, and pragmatic layers within multilingual corpora. Second, it demonstrates how ethical safeguards can be embedded directly into the architecture of data processing rather than treated as external constraints. Third, it positions interpreter-mediated interaction as a valuable site for studying multilingual communication in contexts where relational and institutional dynamics are central.

As SSH research increasingly engages with multilingual data and computational tools, the need for robust, ethically grounded, and multimodally structured infrastructures becomes more pressing. By conceptualising telephone interpreting as a multimodal and plurilingual object of study, PRAGMACOR contributes to this emerging landscape, providing a framework that supports both fine-grained discourse analysis and future interdisciplinary collaboration.

6. Bibliographical References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Springer. https://doi.org/10.1007/978-94-024-0881-2

Jewitt, C. (Ed.). (2014). *The Routledge handbook of multimodal analysis* (2nd ed.). Routledge.

Kress, G., & van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (2nd ed.). Routledge.

Sakai, Y., Makinae, M., Kamigaito, H., Watanabe, T. (2024). Simultaneous Interpretation Corpus Construction by Large Language Models in Distant Language Pair. arXiv:2404.12299. https://doi.org/10.48550/arXiv.2404.12299

Toward Responsible and Epistemically Grounded Multilingual LLMs for Computational Social Science and Humanities

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large language models have rapidly evolved in multilingual competence and reasoning capacity, enabling their integration into Social Sciences and Humanities research workflows. Yet existing evaluation paradigms remain anchored in task-based NLP benchmarks and fail to address interpretive validity, cultural situatedness, and epistemic mediation. This paper reconceptualizes multilingual reasoning LLMs as hermeneutic instruments that actively structure meaning production across linguistic and cultural contexts. Drawing on hermeneutics, philosophy of technology, science and technology studies, multilingual NLP research, and computational social science methodology, we develop a theoretically grounded framework for evaluating multilingual reasoning in Social Sciences and Humanities (SSH) research. We articulate a rigorous experimental protocol with operationalized metrics for cultural alignment, cross-lingual stability, and reasoning faithfulness, along with transparency requirements tailored to interpretive research tasks. We illustrate the framework through a concrete application scenario involving multilingual political discourse analysis. The paper contributes a conceptual and methodological foundation for responsible integration of multilingual reasoning LLMs into computational social science infrastructures.

Keywords: multilingual LLMs, hermeneutic instruments, computational social science, reasoning faithfulness, cultural alignment, epistemic evaluation

1. Introduction

The development of transformer-based large language models has reconfigured the technical landscape of language processing. Models trained on large-scale multilingual corpora demonstrate zero-shot transfer, in-context learning, and chain-of-thought reasoning capacities across a growing number of languages (Brown et al., 2020; Chowdhery et al., 2023; Wei et al., 2022; Touvron et al., 2023). These advances have prompted increasing adoption of LLMs in domains beyond traditional NLP, including political science, sociology, communication studies, digital humanities, and cultural analytics.

Recent work demonstrates that LLMs can assist in text annotation, survey simulation, and qualitative coding tasks (Argyle et al., 2023; Gilardi et al., 2023; Ziems et al., 2024). The computational social science community has begun systematically examining how LLMs can transform research workflows, with emerging evidence showing that zero-shot LLMs can achieve fair levels of agreement with humans on taxonomic labeling tasks while producing explanations that sometimes exceed crowdworker quality (Ziems et al., 2024). Wilkerson and Casas (2017) documented the growing importance of large-scale computerized text analysis in political science, and LLMs represent the latest methodological frontier in this trajectory. However, performance improvements on benchmarks do not directly translate into epistemic adequacy for Social Sciences and Humanities (SSH) research. Interpretive analysis in social science is not merely a classification task. It involves contextualization, reflexivity, and engagement with culturally embedded meaning structures.

Multilingual reasoning models introduce additional complexity. Although models demonstrate cross-lingual competence, research consistently documents uneven performance across languages and structural inequalities in training data (Joshi et al., 2020; Blasi et al., 2022). The most widely used datasets in natural language processing currently represent only a handful of data-rich languages, and the datasets used for instruction fine-tuning are almost entirely focused on English (Singh et al., 2024). Cultural alignment and interpretive fidelity cannot be assumed on the basis of benchmark scores.

This paper advances the argument that multilingual reasoning LLMs should be conceptualized as hermeneutic instruments. Rather than treating them as neutral analytic tools, we frame them as mediators that shape interpretive horizons in computational social science. This reframing has methodological implications for evaluation design, transparency practices, and epistemic accountability. As Rockwell and Sinclair (2016) demonstrate in their foundational work on computer-assisted interpretation, computational tools do not simply process texts but actively structure interpretive possibilities. This perspective aligns with Moretti's (2013) notion of "distant reading," which acknowledges that computational approaches to text involve fundamental shifts in how texts are apprehended.

The contribution is fourfold. First, we articulate a philosophy of technology-grounded account of

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 158–169
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

multilingual LLM mediation in SSH research. Second, we provide an empirically informed analysis of cultural bias in multilingual models, with particular attention to non-Western language contexts. Third, we propose a detailed methodological framework with operationalized metrics for evaluating multilingual reasoning models in interpretive contexts, illustrated through a concrete application scenario. Fourth, we identify epistemic and ethical implications that extend beyond technical evaluation and offer concrete recommendations for the SSH research community.

2. Hermeneutics and Computational Mediation

Hermeneutic philosophy offers a foundational lens for understanding interpretation as historically and culturally situated. Gadamer (1975) emphasizes that understanding occurs through a fusion of horizons between interpreter and text. Ricoeur (1976) similarly frames interpretation as a dialectical process that unfolds between explanation and understanding. In hermeneutic thought, interpretation is never extraction of fixed meaning but a productive act shaped by presuppositions and contextual frames.

The interpretive act is structured by mediating conditions, including language, tradition, and institutional practice. As scholars in digital humanities have argued, computational text analysis necessarily involves hermeneutical operations: the creation of vector space models, topic models, or neural representations should be understood as interpretations that reinscribe texts into new analytical forms (Kuhn, 2019; van Zundert, 2016). When computational systems are introduced into interpretive workflows, they become part of these mediating conditions. Their architectures encode statistical regularities derived from training data, reflecting historically contingent distributions of language use, cultural representation, and epistemic dominance. Kuhn (2019) identifies a central tension in digital humanities between hermeneutic traditions of text interpretation and method-oriented research strategies in computational linguistics.

The philosophical insight that instruments mediate rather than passively transmit phenomena is reinforced in science and technology studies. Latour (1987) argues that scientific instruments transform the entities they measure. Ihde (1990) develops this insight through the concept of "technological intentionality," suggesting that instruments structure perception and interpretation in ways that are not fully transparent to users. Winner (1980) further demonstrates that technological artifacts embody political and social values.

Applying this perspective to multilingual reasoning LLMs shifts the analytical focus. The central question becomes not only whether the model performs accurately but how it structures interpretive possibilities. When an LLM summarizes political discourse in Arabic, reconstructs an argument in French, or attributes causality in English, it does so through latent statistical priors shaped by its training distribution. These priors influence what counts as salient, coherent, or plausible. What Gadamer called the interpreter's "effective-historical consciousness" finds a distorted analogue in the model's training distribution: the model interprets through a horizon constituted not by lived experience but by the statistical regularities of its corpus, weighted by the demographic and linguistic composition of its data sources.

3. Multilingual Representation and Structural Inequality

The multilingual capacity of LLMs is often framed as evidence of inclusivity and global reach. However, empirical research in NLP highlights persistent structural asymmetries. Joshi et al. (2020) document severe disparities in resource availability and benchmark inclusion across languages, categorizing the world's languages into a taxonomy where the vast majority remain "left-behind" in NLP research. Blasi et al. (2022) demonstrate that English plays a disproportionate role in shaping multilingual model performance, leading to what they describe as the unreasonable effectiveness of English in cross-lingual transfer.

Bender et al. (2021) argue that large-scale language models reflect the biases and imbalances of their training data. The Aya initiative represents an important effort to address these gaps through community-driven multilingual instruction tuning across 101 languages, yet significant disparities persist (Singh et al., 2024; Üstün et al., 2024). Community-driven efforts such as Masakhane, which focuses on NLP for African languages, demonstrate the importance of participatory approaches that involve native speakers as researchers rather than merely data providers (Orife et al., 2020; Adelani et al., 2021).

In multilingual reasoning contexts, these structural asymmetries have interpretive implications. Cross-lingual reasoning tasks often rely on translation or transfer learning. Yet semantic equivalence does not guarantee cultural equivalence. Research on cultural alignment demonstrates that multilingual capability does not imply multicultural understanding: LLMs trained primarily on English data consistently align with Western, particularly US-centric, cultural values even when generating content in other languages (Rystrøm et al., 2025; Tao et al., 2024). Li et al. (2024) find that LLMs inherit and amplify cultural patterns present in their training data, replicating cross-cultural personality differences while overrepresenting Western perspectives due to English-dominant corpora.

For computational social science, this is particularly significant. Comparative political analysis depends on detecting differences in framing, narrative structure, and causal attribution. Following Entman's (1993) influential framework, framing involves selecting and making salient certain aspects of perceived reality to promote particular problem definitions, causal interpretations, moral evaluations, and treatment recommendations. Chong and Druckman (2007) further demonstrate how framing shapes public opinion formation. If multilingual reasoning models encode dominant cultural frames more strongly than minority ones, cross-national comparison may inadvertently reflect model priors rather than empirical reality. Prior work on annotating stance, sentiment, and framing in Arabic social media discourse (Laabar and Zaghouani, 2024) illustrates how culturally specific interpretive categories can diverge substantially from categories developed in English-language contexts.

Existing multilingual benchmarks such as XNLI (Conneau et al., 2018) provide valuable infrastructure for evaluating cross-lingual transfer but focus primarily on natural language inference rather than culturally situated interpretation. The MBBQ benchmark demonstrates that bias patterns differ substantially across languages (Neplenbroek et al., 2024). SSH-specific evaluation requires extending these resources to capture interpretive validity.

4. Cultural Bias as Epistemic Distortion

The abstract concern about cultural homogenization in multilingual models is supported by a growing body of empirical work that deserves careful attention from the SSH community. Naous et al. (2024) introduce CAMeL, a benchmark of naturally occurring prompts and entities contrasting Arab and Western cultures, and demonstrate that both multilingual and Arabic monolingual language models exhibit systematic bias toward entities associated with Western culture. When prompted in Arabic to complete sentences about food, beverages, or personal names, models disproportionately generate Western-associated responses rather than culturally appropriate Arab alternatives. This bias persists even in models specifically fine-tuned for Arabic, suggesting that the issue runs deeper than surface-level language competence.

The origins of these biases have been further investigated by Naous and Xu (2025), who find that language models struggle particularly with Arabic entities that appear at high frequencies in pre-training data, where such entities exhibit strong word polysemy. Their analysis reveals that frequency-based tokenization contributes to this problem, and that performance gaps between Arab and Western cultural entities are smaller when models are tested in English compared to Arabic. This finding has a striking implication for SSH researchers: a model may appear culturally competent when evaluated in English but reveal significant cultural blind spots when operating in the target language of analysis.

Large-scale evaluations reinforce these findings. Sukiennik et al. (2025) conduct the first comprehensive assessment of cultural value alignment across 20 countries and 10 LLMs using Hofstede's Values Survey Module, finding that model outputs converge toward a moderate cultural middle ground and that the United States is consistently the best-aligned country across models. Critically, models regardless of their country of origin align better with US cultural values than with the values of their home countries, suggesting that the dominance of English-language training data creates a structural gravitational pull toward Western cultural norms.

These empirical findings have direct methodological consequences for SSH research. A researcher using an LLM to analyze political discourse across Arabic-speaking countries may encounter a model that systematically underweights culturally specific concepts, prioritizes Western-normative framings, or generates interpretive outputs that obscure genuine cross-cultural variation. Work on multidialectal Arabic hate speech detection (Charfi et al., 2024a) and cross-domain stance analysis (Charfi et al., 2024b) has shown that even within a single language, dialectal and cultural variation can substantially affect model performance and annotation validity. For disciplines such as comparative politics, area studies, and cultural sociology, where the detection of culturally specific meaning is the primary research objective, these biases represent not merely technical limitations but epistemic threats to the validity of findings.

5. Reasoning, Explanation, and Epistemic Authority

Chain-of-thought prompting has been shown to improve reasoning performance in large language models (Wei et al., 2022). Generating intermediate steps appears to enhance accuracy on arithmetic and logical tasks. However, Turpin et al. (2023) demonstrate that reasoning traces may not reliably correspond to internal inference processes. Their experiments reveal that LLMs can produce chain-of-thought explanations that are systematically unfaithful, influenced by biasing features in inputs that models fail to mention in their explanations. When models are biased toward incorrect answers through manipulated prompts, they frequently generate plausible-sounding reasoning that rationalizes those incorrect answers, causing accuracy to drop by as much as 36% on benchmark tasks.

This finding has been corroborated by Lanham et al. (2023), who develop multiple metrics for assessing chain-of-thought faithfulness and find substantial variation across tasks in how strongly models condition on their stated reasoning when predicting answers. Their experiments introduce perturbations to chain-of-thought outputs, such as adding mistakes or paraphrasing, to measure whether models genuinely rely on their stated reasoning. Critically, as models become larger and more capable, they sometimes produce less faithful reasoning on certain tasks, raising concerns about the inverse scaling of interpretability.

For SSH research, reasoning traces may acquire epistemic authority. When a model provides a structured explanation of why a political actor adopts a particular stance, researchers may treat the reasoning as analytically meaningful. Yet if reasoning traces are post-hoc constructions optimized for plausibility rather than grounded inference, their interpretive status must be critically examined. Zheng et al. (2023) further demonstrate that LLM-based evaluators can introduce systematic biases in judging outputs, including position bias, verbosity bias, and self-enhancement bias. Chen et al. (2024) and Gu et al. (2024) identify twelve distinct types of bias that can undermine LLM-as-judge reliability.

The faithfulness problem intersects with multilingual reasoning in ways that remain underexplored. When a model reasons in a language other than English, it may rely more heavily on transferred English-language priors, producing reasoning traces that appear coherent in the target language but are actually anchored in English-centric conceptual structures. This creates a particularly insidious form of epistemic distortion: the model may generate culturally inappropriate interpretations accompanied by fluent, seemingly well-reasoned explanations that mask the underlying cultural misalignment. SSH researchers lack adequate tools to detect when this form of cross-lingual reasoning contamination is occurring, making the development of language-specific faithfulness metrics an urgent priority.

6. LLMs in Computational Social Science

Before articulating an evaluation framework, it is important to survey how LLMs are currently being deployed in SSH research. The use of LLMs for text annotation has expanded rapidly, with studies demonstrating that GPT-4 can achieve annotation accuracy comparable to or exceeding human crowdworkers across multiple tasks (Gilardi et al., 2023; Heseltine and Clemm von Hohenberg, 2024). Tornberg (2023) provides evidence that ChatGPT-4 achieves higher accuracy, higher reliability, and equal or lower bias than human classifiers. Rathje et al. (2024) demonstrate that GPT is effective for multilingual psychological text analysis across 12 languages.

Egami et al. (2024) develop a rigorous statistical framework for using LLM annotations in downstream social science analysis, showing that ignoring prediction errors from automated annotation can lead to substantial bias, invalid confidence intervals, and inaccurate p-values. Carlson and Burbano (2025) extend this line of work by developing foundational guidelines for using LLMs to annotate data in management research, demonstrating that subtle implementation choices, including prompt wording, model version, and parameter settings, can significantly affect not only annotation accuracy but also downstream research conclusions.

However, research also reveals that LLMs exhibit party cue biases when annotating political content (Vallejo Vera et al., 2025). For qualitative research, Hayes (2025) argues that LLMs enable researchers to “converse” with qualitative data in unprecedented ways, but this capability comes with risks: LLMs may impose interpretive frameworks that do not align with the cultural contexts being studied. The codebook-following capabilities of LLMs have been systematically examined by Halterman and Keith (2024), who find that providing detailed social science codebooks significantly improves classification performance. Ollion et al. (2023) urge researchers to “mind the hype,” noting that performance varies considerably across tasks and contexts.

An emerging concern involves the propagation of methodological choices through the research pipeline. When LLM annotations serve as inputs to regression models, hypothesis tests, or causal inference procedures, even small systematic biases can compound. Egami et al. (2024) propose design-based supervised learning (DSL) as a correction mechanism, but this requires a subsample of gold-standard human annotations. For multilingual SSH research, the practical challenge is acute: obtaining high-quality human annotations in multiple languages and cultural contexts is precisely the bottleneck that motivates LLM adoption in the first place. This creates a methodological circularity that the field must address through creative experimental designs, such as stratified validation sampling that ensures adequate representation of culturally distinctive categories.

7. Methodological Framework with Operationalized Metrics

This section outlines an experimental protocol for evaluating multilingual reasoning LLMs in computational social science contexts, with concrete operationalized metrics addressing cultural alignment, cross-lingual stability, and reasoning faithfulness.

Corpus Construction and Documentation. A valid multilingual evaluation requires corpora composed of native language texts rather than translated benchmarks. The distinction between native and translated evaluation materials is critical: translated benchmarks inherit the conceptual categories and pragmatic assumptions of their source language, introducing systematic confounds into cross-lingual evaluation. Documents should be sampled from comparable genres across linguistic contexts, where "comparable" is defined by functional equivalence (texts serving similar communicative purposes in their respective societies) rather than semantic equivalence (texts expressing the same propositional content). Following Bender and Friedman (2018), researchers should provide comprehensive Data Statements documenting: (a) curation rationale, (b) language variety with dialect specification, (c) speaker demographics, (d) annotator demographics, (e) speech situation context, and (f) text characteristics. This documentation standard, along with Model Cards (Mitchell et al., 2019), ensures transparency about the populations represented in evaluation data.

Annotation Schema. Interpretive tasks must be grounded in established SSH theory. For political discourse analysis, tasks may include frame identification following Entman's (1993) four functions. Annotation guidelines should be developed collaboratively with native speaker experts. The distinction between "universal label assumptions" and "codebook-contextual label assumptions" is crucial (Halterman and Keith, 2024). Universal label assumptions treat annotation categories as cross-culturally stable, whereas codebook-contextual assumptions recognize that categories may require adaptation to local meaning systems. Multiple annotators per language should independently code a subset of the corpus to establish reliability. Following Krippendorff (2004), we adopt the standard thresholds of Krippendorff's alpha $\geq$ 0.67 for tentative or exploratory conclusions and $\geq$ 0.80 for confirmatory research where findings will inform substantive claims.

Cultural Alignment Metrics. To operationalize cultural alignment, we propose measuring the distance between model-predicted attitude or value distributions and population benchmarks from validated cross-cultural surveys such as the World Values Survey (WVS) or European Social Survey (ESS). Specifically:

- *KL Divergence*: $D_{KL}(P_{population}||P_{model})$ where $P_{population}$ represents the distribution of responses on a value dimension (e.g., traditional vs. secular-rational values) from WVS respondents in the target culture, and P_{model} represents the distribution of model outputs on comparable items. Lower values indicate better alignment. We propose a threshold of $D_{KL} < 0.1$ nats as a starting point; this value should be calibrated empirically through pilot studies, as the appropriate threshold will vary by task and domain.
- *Earth Mover's Distance (EMD)*: For ordinal scales, EMD provides an interpretable metric of the "work" required to transform model distributions into population distributions.
- *Cultural Bias Score*: Following Naous et al. (2024), researchers should additionally compute entity-level bias scores comparing model performance on culturally specific entities (e.g., Arab vs. Western food items, names, or locations) to detect systematic cultural skew that aggregate distributional metrics may obscure.

Cross-Lingual Stability Metrics. Cross-lingual stability measures whether model performance and interpretive outputs remain consistent across languages for semantically equivalent inputs. We propose:

- *Variance decomposition*: Using mixed-effects models with language as a random effect, decompose total variance into between-language ($\sigma^2_{language}$) and within-language ($\sigma^2_{residual}$) components. The intraclass correlation coefficient (ICC) = $\sigma^2_{language}/(\sigma^2_{language} + \sigma^2_{residual})$ quantifies the proportion of variance attributable to language. Following conventions in reliability research (Cicchetti, 1994), ICC $>$ 0.10 suggests that language identity explains a non-trivial share of variance, warranting investigation of language-specific performance differences.
- *Pairwise agreement*: For each language pair, compute Cohen's kappa between model outputs on parallel test items. Mean pairwise kappa $\geq$ 0.60 indicates acceptable stability.
- *Language-direction asymmetry*: For each language pair, compare model performance when the task is formulated in language A versus language B. Asymmetric performance (e.g., consistently higher accuracy when prompting in English than in Arabic for the same underlying task) may indicate reliance on English-language priors rather than genuine multilingual competence.

Reasoning Faithfulness Assessment. Following Turpin et al. (2023) and Lanham et al. (2023), we recommend three perturbation families:

- *Bias injection*: Introduce irrelevant features (e.g., suggested answers, social cues) into prompts. If model answers change but explanations do not acknowledge the influence, faithfulness is compromised. Acceptance criterion: answer stability $\geq$ 90% under bias injection.
- *Reasoning corruption*: Introduce errors into chain-of-thought traces and measure whether final answers change correspondingly. If answers remain stable despite corrupted reasoning, the model is not genuinely conditioning on its explanations.
- *Cross-lingual reasoning transfer*: Present the same reasoning task in multiple languages and compare not only final answers but the structure and content of reasoning traces. Divergent reasoning paths for semantically equivalent inputs may reveal language-dependent reasoning strategies that compromise cross-lingual interpretive consistency.

Relationship to Existing Frameworks. Our framework complements rather than replaces existing evaluation paradigms. The Holistic Evaluation of Language Models (HELM) framework (Liang et al., 2023) provides infrastructure for multi-metric evaluation including fairness and robustness; we extend this by adding SSH-specific interpretive dimensions. XNLI (Conneau et al., 2018) and related benchmarks provide cross-lingual NLU baselines; our framework adds culturally grounded interpretive tasks. The CAMeL benchmark (Naous et al., 2024) and the cultural alignment evaluations by Sukiennik et al. (2025) provide complementary tools for assessing entity-level cultural bias and value-level cultural alignment respectively; our framework integrates both granularity levels within a unified SSH evaluation pipeline.

Transparency and Reproducibility. Researchers must log prompts, outputs, timestamps, model versions, and decoding parameters. For proprietary models, researchers should test for consistency across sessions and document any detected model updates (Linegar et al., 2023). Following Abdurahman et al. (2025), researchers should report disaggregated results across languages and demographic categories represented in texts. Carlson and Burbano (2025) recommend systematic sensitivity analysis across prompt formulations; we adopt this as a core principle: SSH researchers should routinely test at least three prompt variants per annotation task and report the range of results obtained.

8. Illustrative Application: Multilingual Political Discourse Analysis

To demonstrate how the proposed framework can be applied in practice, we outline a concrete research scenario involving multilingual political framing analysis, a common task in computational social science.

Research question. How do news media in Arabic, English, and French frame immigration policy debates, and can LLMs reliably identify these frames across languages?

Step 1: Corpus construction. The researcher assembles native-language editorial articles from major outlets in each language (e.g., Al Jazeera Arabic, The Guardian, Le Monde). Crucially, these are not translations but independently authored texts addressing immigration in their respective national contexts. A Data Statement (Bender and Friedman, 2018) documents language variety (e.g., Modern Standard Arabic vs. Gulf dialect), publication period, editorial stance distribution, and any known ideological affiliations.

Step 2: Annotation schema development. Frame categories are developed collaboratively with area specialists in each language community, following Entman's four framing functions (problem definition, causal attribution, moral evaluation, treatment recommendation). The schema is piloted with native-speaker annotators. Suppose inter-annotator reliability yields Krippendorff's $\alpha = 0.74$ for Arabic, $\alpha = 0.82$ for English, and $\alpha = 0.78$ for French. Following Krippendorff (2004), the Arabic and French scores fall in the tentative range (0.67–0.80) and are acceptable for exploratory analysis but would require refinement for confirmatory claims. The English score exceeds 0.80 and supports stronger conclusions.

Step 3: Cultural alignment check. Before deploying LLMs for full-corpus annotation, the researcher tests cultural alignment on a probe set of 50 items per language containing culturally specific entities and concepts. Using the Cultural Bias Score method from CAMeL (Naous et al., 2024), the researcher discovers that the model assigns Western-normative immigration framings (e.g., "economic burden") at higher rates in Arabic than Arabic-speaking annotators do, revealing a systematic cultural skew that would contaminate cross-national comparisons if left uncorrected.

Step 4: Cross-lingual stability assessment. The researcher identifies a subset of 30 parallel items (events covered in all three languages) and computes pairwise Cohen's κ between LLM annotations across languages. If English-French $\kappa = 0.72$ but English-Arabic $\kappa = 0.51$, this asymmetry signals that the model's Arabic frame identification is substantially less stable, likely reflecting weaker Arabic-language priors.

Step 5: Reasoning faithfulness test. For a subsample of items, the researcher requests chain-of-thought explanations and applies bias injection (prepending a suggested frame label). If the model's frame assignments shift in 25% of Arabic cases but only 8% of English cases under bias injection, this reveals language-dependent faithfulness, a red flag for cross-lingual interpretive validity.

Step 6: Reporting. The researcher documents all findings in the Epistemic Risk Register (Section 9), reports disaggregated metrics, and qualifies cross-lingual conclusions appropriately. The Arabic frame analysis is presented with explicit caveats about reduced stability and cultural alignment, while the English-French comparison receives stronger interpretive weight.

This scenario illustrates that the framework is implementable with modest resources: the probe sets and parallel items are small (50 and 30 items respectively), the statistical tests are standard, and the primary investment is in recruiting area specialists for schema development and cultural probing rather than in large-scale annotation.

9. Epistemic and Ethical Implications

The integration of multilingual reasoning LLMs into SSH research introduces epistemic and ethical considerations that extend beyond technical evaluation.

Cultural homogenization risk arises when dominant language priors shape interpretation across contexts. Models trained primarily on English data may impose Western conceptual frameworks on texts from other cultural traditions. Research on cultural alignment reveals that even when generating content in non-English languages, LLMs often reflect the value systems of English-speaking countries (Tao et al., 2024). The empirical findings from Naous et al. (2024) and Sukiennik et al. (2025) demonstrate that this risk is not hypothetical but measurable.

Authority displacement may occur if model outputs are treated as objective rather than mediated. Hayes (2025) cautions that while LLMs offer powerful capabilities for engaging with qualitative data, researchers must maintain critical awareness that model outputs reflect training distributions rather than unmediated access to textual meaning. The risk is amplified when LLMs are used both to generate and evaluate interpretive claims, creating closed loops that may reinforce rather than interrogate model priors (Zheng et al., 2023). Experience from large-scale annotation projects involving culturally sensitive content, such as hate speech annotation, has shown that even human annotators are affected by emotional toll and interpretive fatigue (Al Emadi and Zaghouani, 2024), underscoring the need for careful oversight whether annotation is performed by humans or machines.

Reproducibility challenges emerge when proprietary models are updated without transparency (Bommasani et al., 2021). Open-source models offer greater stability and control, but even these are subject to community-driven updates and version proliferation.

The reliability of survey-based methods for assessing cultural alignment itself warrants critical examination. Recent work has challenged assumptions that cultural alignment is a stable property of models rather than an artifact of evaluation design, and that alignment on one set of cultural dimensions predicts alignment on others. Empirical tests reveal significant instability across presentation formats and incoherence between evaluated and held-out cultural dimensions, reinforcing the need for multi-method evaluation approaches.

We propose an **Epistemic Risk Register** that SSH researchers should complete when deploying multilingual LLMs:

1. *Codebook provenance*: Who defined the annotation categories? Were native speakers from target cultures involved?

2. *Cultural stakeholder engagement*: How were affected communities consulted in evaluation design?

3. *Disaggregation requirements*: What subgroup analyses are required to detect differential performance?

4. *Model update monitoring*: What procedures exist to detect and document model changes over time?

5. *Interpretive authority*: How are model outputs positioned relative to human expert judgment?

6. *Prompt sensitivity documentation*: Were multiple prompt formulations tested, and what was the range of variation in results?

7. *Cultural alignment verification*: Were model outputs assessed for systematic cultural bias using established benchmarks or domain-appropriate cultural probes?

10. Toward SSH-Specific Benchmarks

The framework proposed here highlights a significant gap in the current evaluation landscape: the absence of benchmarks specifically designed for SSH interpretive tasks. Existing multilingual benchmarks primarily assess factual knowledge retrieval, natural language inference, or commonsense reasoning. While these capabilities are necessary for SSH applications, they are insufficient.

We identify four properties that SSH-specific benchmarks should exhibit. First, *interpretive pluralism*: tasks should admit multiple defensible answers rather than a single correct response. Evaluation metrics should reward appropriate uncertainty and penalize unwarranted confidence. Second, *cultural grounding*: benchmark items should be developed in their target languages by domain experts embedded in the relevant cultural contexts, not translated from English-language originals. Third, *theoretical anchoring*: annotation categories should derive from established SSH theoretical frameworks rather than ad hoc classification schemes. Fourth, *ecological validity*: test materials should be drawn from the actual text genres that SSH researchers analyze (parliamentary debates, news editorials, social media discourse, interview transcripts) rather than synthetic inputs.

The development of such benchmarks requires sustained collaboration between NLP researchers and SSH scholars. Kuhn (2019) identifies the scheduling dilemma that arises when computational methods require early specification while hermeneutic approaches prefer late specification as understanding develops. We propose that benchmark development adopt an iterative co-design methodology. Community-driven approaches exemplified by Masakhane (Orife et al., 2020) and the Aya initiative (Singh et al., 2024; Üstün et al., 2024) should inform how native speakers and SSH scholars from underrepresented language communities participate in benchmark development as researchers rather than merely as annotators.

11. Practical Guidance for Resource-Constrained Settings

A realistic concern, raised in the evaluation of this framework, is that the full protocol may be infeasible for researchers lacking access to large multilingual annotator pools, validated cultural survey data, or extensive computational resources. We address this by proposing a tiered implementation approach.

Minimum viable evaluation (Tier 1). Even with limited resources, researchers can implement three basic checks: (a) test at least three prompt variants per task and report variance across formulations, following Carlson and Burbano (2025); (b) compute cross-lingual agreement on a small parallel set (as few as 20 items) to flag gross instability; and (c) inspect a sample of reasoning traces for obvious cultural misalignment using researcher domain expertise. These steps require no specialized infrastructure and can be completed in hours.

Standard evaluation (Tier 2). With moderate resources, researchers add: (a) formal inter-annotator reliability with native speakers (minimum two annotators per language, 100 item subsample); (b) cultural bias probing using entity substitution (comparing model behavior on culturally specific vs. Western entities); and (c) bias injection tests on a subsample of reasoning tasks. This tier requires collaborators in each target language community but no large-scale data collection.

Full protocol (Tier 3). The complete framework as described in Section 6, including population-level cultural alignment metrics using WVS or ESS data, formal variance decomposition, and comprehensive faithfulness testing. This tier is appropriate for high-stakes research projects with dedicated evaluation budgets.

This tiered approach ensures that even researchers with minimal resources can incorporate epistemic accountability into their workflows, while providing a clear path for scaling evaluation rigor as resources allow.

12. Implications for the SSH Research Agenda

Realizing the potential of multilingual reasoning LLMs for SSH research requires developing shared infrastructure for documenting model configurations, logging experimental procedures, and archiving annotation materials. We recommend that SSH researchers adopt Data Statements (Bender and Friedman, 2018) and Model Cards (Mitchell et al., 2019) as minimum documentation standards, extended with the epistemic risk register proposed above.

The field would benefit from establishing multilingual LLM evaluation consortia that pool expertise across language communities and SSH disciplines. Such consortia could maintain living benchmarks that evolve alongside model capabilities, conduct regular cross-model comparative evaluations on SSH-relevant tasks, and develop shared annotation resources that reduce the per-project cost of multilingual validation.

We further recommend that SSH journals and conferences develop reporting standards for studies that use LLM-generated annotations or LLM-assisted analysis. At a minimum, researchers should report the specific model name, version, and access date; the complete prompt text used for each task; the decoding parameters (temperature, top-p); the number of prompt variants tested and the sensitivity of results to prompt choice; disaggregated performance metrics across languages and cultural subgroups; and any detected changes in model behavior over the course of data collection.

13. Conclusion

Multilingual reasoning LLMs function as hermeneutic instruments that mediate interpretation in computational social science. Recognizing this role requires moving beyond benchmark accuracy toward epistemically grounded evaluation frameworks that attend to cultural situatedness, reasoning faithfulness, and the structural inequalities embedded in training data. By combining hermeneutics, philosophy of technology, and technical methodology with operationalized metrics, this paper provides a foundation for responsible multilingual reasoning assessment aligned with SSH values. The empirical evidence reviewed here, from cultural bias in Arabic-language models to the unreliability of chain-of-thought reasoning and the instability of cultural alignment evaluations, demonstrates that these concerns are not merely theoretical but have concrete, measurable consequences for the validity of SSH research. The illustrative application in Section 8 demonstrates that meaningful evaluation is achievable even with modest resources when guided by a principled framework. Future work should operationalize these principles in large-scale empirical studies, and we particularly encourage pilot studies that implement the tiered evaluation protocol on small multilingual corpora to demonstrate feasibility and refine metric thresholds. The research community is well-positioned to advance this agenda through sustained interdisciplinary collaboration.

14. Limitations

This position paper presents a conceptual and methodological framework for evaluating multilingual reasoning large language models in Social Sciences and Humanities research. The paper is primarily theoretical. The suggested measures for cultural alignment, cross-lingual stability, and reasoning faithfulness have not been validated through large-scale empirical testing across many different models. This is especially true for models that work with low-resource or typologically diverse languages.

The cultural alignment metrics depend on population-level cultural datasets from validated surveys such as the World Values Survey. These datasets may be unavailable, outdated, or insufficiently representative for many non-Western and indigenous groups. The proposed KL divergence threshold of 0.1 nats and the ICC threshold of 0.10 are offered as starting points informed by standard statistical practice, but they require empirical calibration through pilot studies in specific SSH domains before they can be treated as firm benchmarks.

The framework focuses on text-based tasks and does not yet address multimodal analysis or real-time interactive settings. Language-specific challenges such as diverse writing systems, dialectal variation, code-switching, and uneven tokenization may produce larger performance disparities than the current stability metrics capture. The illustrative scenario in Section 8, while designed to be realistic, has not been executed as a full empirical study. Finally, the framework assumes some degree of model access or adaptability, which may not be available with closed commercial models. These points underscore the need for iterative, community-based validation before wide adoption.

15. Ethical Considerations

Integrating multilingual reasoning LLMs into SSH research raises ethical concerns beyond epistemic issues. A primary risk is cultural homogenization from English-dominant training data, which perpetuates Western normative interpretations and marginalizes non-Western epistemologies, indigenous knowledge, and minority voices. This reinforces colonial legacies in global knowledge production. Other concerns include deskilling researchers through over-reliance on automation, privacy risks with sensitive qualitative data, misuse of outputs in high-stakes interpretive work without human oversight, and accountability problems from hallucinations or inconsistent cross-lingual performance.

To address these we advocate participatory design with native speakers, Global South scholars, and community stakeholders as co-creators of benchmarks, protocols, and transparency standards, following models like Masakhane (Orife et al., 2020) and Aya (Singh et al., 2024). Researchers must document model provenance, prompts, detected biases, and limitations while keeping human reflexivity central. Responsible adoption requires ongoing ethical reflection to ensure LLMs support rather than replace culturally situated human understanding.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

16. Bibliographical References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., and Dehghani, M. 2025. A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science*, 8(2).

Adelani, D. I., Abbott, J., Neubig, G., et al. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Al Emadi, M. M. and Zaghouani, W. 2024. Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, 66–72.

Argyle, L. P., Busby, E., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bender, E. M. and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021*, 610–623.

Blasi, D. E., Anastasopoulos, A., and Neubig, G. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of ACL 2022*.

Bommasani, R., Hudson, D. A., Adeli, E., et al. 2021. On the opportunities and risks of foundation models. arXiv:2108.07258.

Brown, T. B., Mann, B., Ryder, N., et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33.

Carlson, N. A. and Burbano, V. 2025. The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*.

Charfi, A., Bessghaier, M., Akasheh, R., Atalla, A., and Zaghouani, W. 2024a. Hate Speech Detection with ADHAR: A Multi-Dialectal Hate Speech Corpus in Arabic. *Frontiers in Artificial Intelligence*, 7, Article 1391472.

Charfi, A., Bessghaier, M., Atalla, A., Akasheh, R., Al-Emadi, S., and Zaghouani, W. 2024b. Stance Detection in Arabic with a Multi-Dialectal Cross-Domain Stance Corpus. *Social Network Analysis and Mining*, 14(1), 161.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. 2024. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of EMNLP 2024*, 8301–8327.

Chong, D. and Druckman, J. N. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Chowdhery, A., Narang, S., Devlin, J., et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Cicchetti, D. V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, 2475–2485.

Egami, N., Hinck, M., Stewart, B., and Wei, H. 2024. Using large language model annotations for the social sciences: A general framework with applications. Working paper.

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Gadamer, H. G. 1975. *Truth and Method*. Seabury Press.

Gilardi, F., Alizadeh, M., and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Gu, J., Jiang, X., et al. 2024. A survey on LLM-as-a-Judge. arXiv:2411.15594.

Halterman, A. and Keith, K. A. 2024. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. arXiv:2407.10747.

Hayes, A. S. 2025. Conversing with qualitative data: Enhancing qualitative research through large language models. *International Journal of Qualitative Methods*.

Heseltine, M. and Clemm von Hohenberg, B. 2024. Large language models as a substitute for human experts in annotating political text. *Research and Politics*, 11(1):1–10.

Ihde, D. 1990. *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020*, 6282–6293.

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage Publications.

Kuhn, J. 2019. Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, 53(4):565–602.

Laabar, S. and Zaghouani, W. 2024. Multi-Dimensional Insights: Annotated Dataset of Stance, Sentiment, and Emotion in Facebook Comments on Tunisia's July 25 Measures. In *Proceedings of the Second Workshop on NLP for Political Sciences @ LREC-COLING 2024*, 22–32.

Lanham, T., Chen, A., Radhakrishnan, A., et al. 2023. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702.

Latour, B. 1987. *Science in Action*. Harvard University Press.

Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. 2024. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024).

Liang, P., Bommasani, R., Lee, T., et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Linegar, M., Kocielnik, R., and Alvarez, R. M. 2023. Large language models and political science. *Frontiers in Political Science*, 5:1257092.

Mitchell, M., Wu, S., Zaldivar, A., et al. 2019. Model cards for model reporting. In *Proceedings of FAccT 2019*, 220–229.

Moretti, F. 2013. *Distant Reading*. Verso Books.

Naous, T., Ryan, M. J., Ritter, A., and Xu, W. 2024. Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of ACL 2024*, 16366–16393.

Naous, T. and Xu, W. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. In *Proceedings of NAACL 2025*.

Neplenbroek, V., Bisazza, A., and Fernández, R. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *Proceedings of COLM 2024*.

Ollion, E., Shen, R., Macanovic, A., and Chatelain, A. 2023. ChatGPT for text annotation? Mind the hype! SocArXiv.

Orife, I., Kreutzer, J., Sibanda, B., et al. 2020. Masakhane – Machine translation for Africa. arXiv:2003.11529.

Rathje, S., Mirea, D. M., Sucholutsky, I., et al. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Ricoeur, P. 1976. *Interpretation Theory*. Texas Christian University Press.

Rockwell, G. and Sinclair, S. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.

Rystrøm, J., Kirk, H. R., and Hale, S. 2025. Multilingual != Multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. arXiv:2502.16534.

Singh, S., et al. 2024. Aya Dataset: An open-access collection for multilingual instruction tuning. arXiv:2402.06619.

Sukiennik, N., et al. 2025. An evaluation of cultural value alignment in LLM. arXiv:2504.08863.

Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Tornberg, P. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv:2304.06588.

Touvron, H., Lavril, T., Izacard, G., et al. 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.

Üstün, A., Aryabumi, V., Yong, Z. X., et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. arXiv:2402.07827.

Vallejo Vera, S., et al. 2025. LLMs as annotators: The effect of party cues on labelling decisions by large language models. *Humanities and Social Sciences Communications*.

van Zundert, J. 2016. Screwmeneutics and hermenumericals: The computationality of hermeneutics. In *A New Companion to Digital Humanities*, Wiley-Blackwell.

Wei, J., Wang, X., Schuurmans, D., et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wilkerson, J. and Casas, A. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544.

Winner, L. 1980. Do artifacts have politics? *Daedalus*, 109(1):121–136.

Zheng, L., Chiang, W. L., Sheng, Y., et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Ziems, C., Held, W., Shaikh, O., et al. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Automatic Evaluation of Multiple-Choice Items for Reading Comprehension: Effects of Question and Distractor Categories

John S. Y. Lee[1], **Yin Poon**[1], **Shunjie Wang**[1], **Chu Kai Wah Samuel**[2]
[1]Department of Linguistics and Translation, City University of Hong Kong
[2]Graduate School, The Education University of Hong Kong
jsylee@cityu.edu.hk, poonyin@outlook.com, ShunjieWang@hotmail.com, chukaiwah@eduhk.hk

Abstract

Automatic generation of multiple-choice (MC) items for reading comprehension can support language learning by providing large amounts of practice materials. To enable rapid development of MC generation models, automatic assessment is essential since it is time-consuming to manually evaluate question and distractor quality. Although Text Informativity (TI) has been adopted as an automatic evaluation metric, the ability of Large Language Models (LLMs) to estimate the TI scores of different categories of questions and distractors has not yet been thoroughly analyzed. This paper investigates LLM performance in calculating TI scores for the range of questions and distractors defined in the PIRLS (Progress in International Reading Literacy Study) and STARC (Structured Annotations for Reading Comprehension) frameworks. We show that automatically estimated TI scores may result in systematic preferences for some question and distractor categories, and recommend that TI scores be used for within-category comparisons only.

Keywords: multiple-choice items, text informativity, question generation, distractor generation

1. Introduction

A multiple-choice (MC) item for reading comprehension consists of a passage, a question, and a number of *answer options*, which must include one *key* (i.e., the correct answer) and several *distractors*. Automatic generation of MC items not only reduces teachers' workload (Cheung et al., 2023), but also provides language learners with large amounts of exercises and practice materials. Advances in artificial intelligence have led to generation models that can produce high-quality MC items for various text genres (Elkins et al., 2024; Kalpakchi and Boye, 2023; Xiao et al., 2023; Wang et al., 2022). However, evaluation methodology has mostly relied on manual assessment of item quality, which is extremely time consuming.

To facilitate rapid development of MC generation models, Text Informativity (TI) has been proposed as an automatic evaluation metric (Säuberli and Clematide, 2024). Using a Large Language Model (LLM) for TI calculation assumes the model's competence in reading comprehension, but this assumption may not hold in the face of sophisticated questions and distractors, such as those recommended in reading comprehension assessment research (King et al., 2004; Mullis and Martin, 2019).

This paper investigates LLM performance in estimating TI scores for the range of questions and distractors defined in the PIRLS (Progress in International Reading Literacy Study) (Mullis and Martin, 2019) and STARC (Structured Annotations for Reading Comprehension) (Berzak et al., 2020) frameworks. We show that the estimated TI scores may lead to systematic preferences for some question and distractor categories. To mitigate this bias, it is recommended that TI scores be used for within-category comparisons only.

2. Question dataset

According to the International Association for the Evaluation of Educational Achievement, reading comprehension questions should address four comprehension processes, as defined in the PIRLS standards (Mullis and Martin, 2019):

Retrieval The answer is explicitly given in a text span in the passage.

Inferencing Answering the question requires inferences about ideas or information that is not explicitly stated in the text.

Integrating Answering the question requires comprehension of the entire passage, or at least significant portions of it.

Evaluation The answer involves judgement about some aspect of the text, and is not necessarily found in the passage.

The three latter categories, which require deeper understanding of the text, are known as *higher-order* questions.

While large-scale MC datasets are available for evaluating LLM performance (Hendrycks et al., 2021; Li et al., 2024), they do not focus on reading comprehension and are not annotated with question or distractor categories. We harvested a set of 390 questions for reading comprehension in Chinese that have been manually annotated with

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 170–174
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

PIRLS		STARC	
Question category	# questions	Distractor category	# distractors
Retrieval	103	In-span	486
Inferencing	147	Out-of-span	486
Integrating	69	Out-of-text	486
Evaluation	71		

Table 1: Breakdown of PIRLS question categories (Section 2) and STARC distractor categories (Section 3) in our dataset of human-crafted MC items

PIRLS categories.[1] They consist of 100 questions from public examinations and 290 questions taken from online PIRLS exercises.[2]

3. Distractor dataset

Educators have advocated the use of distractors that reveal the test-taker's misunderstanding of the passage, which can provide more informative assessment (King et al., 2004). To this end, the STARC framework proposes a taxonomy of distractors that reflect increasing degrees of miscomprehension (Berzak et al., 2020):[3]

Category B (In-span) is a distractor that is based on a misinterpretation of the *critical span*, i.e., the text span in the passage that is relevant to the question.

Category C (Out-of-span) is a distractor derived from a text span that is outside the critical span and is irrelevant to the question.

Category D (Out-of-text) is a distractor based on external knowledge or common sense, without textual support in the passage.

We used the OneStopQA dataset (Berzak et al., 2020), which contains 163 short passages and 3 questions per passage. For each of the 486 questions, besides the key, one In-span distractor, one Out-of-span distractor and one Out-of-text distractor are provided.

4. Evaluation metric

Text Informativity (TI) estimates the degree to which an MC item measures the test-taker's comprehension of the given passage (Säuberli and Clematide, 2024). Two scores need to be computed:

Answerability Given a passage, a question and an answer option, whether the human judge (or the LLM in automatic evaluation) can correctly label the answer option as `true`/`false`. Answerability should be *high* for well-designed MC items with clear questions about the passage and unambiguous answer options.

Guessability Same as above, except that the judge is not given the passage. Guessability should be *low* for well-designed MC items, since it should be impossible to determine whether an answer option is `true`/`false` without the passage. Questions that are not tied to the passage, and distractors that are obviously false, would lead to higher guessability.

The TI score is defined as the difference between the answerability score and the guessability score. Hence, a high TI score means the MC item has answer options whose correctness can be determined (high answerability), but only based on the content of the passage (low guessability).

5. Research Questions

Automatic evaluation is critical for rapid development of MC generation algorithms. According to a study on a German dataset (Säuberli and Clematide, 2024), GPT-4 was able to assign higher TI scores to human-crafted MC items than machine-generated items. However, the study was agnostic to the nature of the underlying questions and distractors. Research in reading comprehension assessment has emphasized the importance of using questions that require a variety of skills (Section 2) and distractors that reveal a range of comprehension levels (Section 3). It is therefore important to ascertain whether LLMs can reliably calculate TI scores for all categories of questions and distractors. To address this issue, this paper seeks to answer the following research questions:

RQ1 *How well can LLMs estimate the TI score (answerability and guessability) of MC items with different question and distractor categories?* An LLM that is not competent with higher-order questions (Section 2), for example, would underestimate their answerability.

RQ2 *Does TI systematically favor any question category or distractor category?* Some distractors are designed to appeal to common sense rather than to the content of the passage (Section 3). They may have lower guessability since they would likely appear to be true when considered without the passage.

[1] Our question dataset can be downloaded from https://github.com/pypoon/PIRLS_ZH_Dataset

[2] These exercises were accessed from: https://read.smes.tyc.edu.tw/smes/PIRLS/ ; https://www.cacler.hku.hk/hk/content/basic/5675 ; https://drive.google.com/file/d/1QTTaarqMJRvy7wU _SzOmw RxE2raAMzgt/view

[3] The original scheme uses only the letters (B, C, D). The descriptive names are given to facilitate discussion.

Dataset	Answer-ability	Guess-ability	Text Informativity
PIRLS	0.9455	0.6468	0.2987
STARC	0.9534	0.5295	0.4239

Table 2: Automatically computed TI scores on human-crafted MC items

These two questions must be resolved in order to apply TI as an automatic metric for MC item evaluation. If TI penalizes challenging questions because of lower answerability (RQ1), or if it favors distractors with inherently lower guessability (RQ2), then it would not facilitate the selection of MC items with diverse question and distractor categories.

6. Experiments and Analysis

The Text Informativity (TI) scores of MC items were calculated with GPT-4o.[4] TI is based on the answerability score and guessability score. To determine the answerability and guessability of a triplet {<passage>, <question>, <answer>), the LLM was prompted to label the answer option as true or false. In our experiments, we used the prompt in Table 3 for answerability, and the prompt in Table 4 for guessability.

```
Text:  <passage>
Question:  <question>
Answer:  <answer>

Based on the text above, is this
answer correct (T) or incorrect
(F)? Indicate only the letter T
or F.
```

```
文本: <passage>
問題： <question>
答案： <answer>

根據上面的文本，這個答案是正確的(T)還是
錯誤的(F)？僅輸出字母T或F。
```

Table 3: Prompt for answerability: the LLM is to label an answer option as true/false when given the passage and question in English (top) and in Chinese (bottom)

6.1. Overall results

Table 2 shows the overall answerability and guessability scores. Since all MC items in our datasets are manually crafted, the gold answerability should in principle be 100%. GPT-4o performed well on both the PIRLS (94.55%) and STARC (95.34%) datasets, achieving relatively high answerability.

When not given the passage, GPT-4o was less accurate in judging the answer options (64.68% for PIRLS and 52.95% for STARC). The substantial gap between the answerability and guessability scores suggests that GPT-4o did indeed understand the passages when tackling the MC items. While these results suggest that the LLM can approximate the ability of a human reader, the next sections will reveal variations in its competence for different question and distractor categories.

```
The following question and answer
are from a multiple-choice com-
prehension task about an unknown
text.

Question:  <question>
Answer:  <answer>

Without knowing the text, only
based on general knowledge, is
this answer more likely to be
correct (T) or incorrect (F)? In-
dicate only letter T or F.
```

```
以下問題和答案來自一篇未知文本的多項選
擇閱讀理解題。

問題： <question>
答案： <answer>

在不知道文本的情況下，僅根據一般知
識，這個答案更有可能是正確的(T)還是錯誤
的(F)？僅輸出字母T或F。
```

Table 4: Prompt for guessability: the LLM is to label an answer option as true/false when given the question but not the passage, in English (top) and in Chinese (bottom)

PIRLS categories	Answer-ability	Guess-ability	Text Informativity
Retrieval	**0.9587**	0.6335	0.3252
Inferencing	0.9473	**0.5833**	**0.3639**
Integrating	0.9275	0.6812	0.2464
Evaluation	0.9401	0.7641	0.1761

Table 5: Automatically calculated TI scores on human-crafted MC items with various PIRLS question types (Section 6.2)

6.2. Effect of Question Categories

Table 5 shows the automatically calculated TI scores for questions belonging to each PIRLS category, using the dataset described in Section 2.

[4] Version 2024-05-13, via Azure OpenAI API. The temperature was set to 0.

STARC categories	Answer-ability	Guess-ability	Text Informativity
In-span	0.9527	0.4074	0.5453
Out-of-span	0.9877	0.4753	0.5123
Out-of-text	**0.9897**	**0.3786**	**0.6111**

Table 6: Automatically calculated TI scores for human-crafted MC items with various STARC distractor categories (Section 6.3)

Answerability. GPT-4o performed best on the Retrieval questions, accurately labeling 95.87% of the answer options as `true`/`false`. These questions should be easier than the higher-order questions, which require more advanced comprehension skills. In line with this expectation, the LLM's performance degraded to 94.73% for Inferencing, and further down to 92.75% for Integrating. Performance on the Evaluation questions was slightly higher (94.01%). These questions may require subjective judgment based on general knowledge (Section 2), to which GPT-4o is well exposed.

Guessability. For similar reasons, the LLM was most capable in judging the answer options for the Evaluation questions (76.41%) without reading the passage. Retrieval and Inferencing questions, which are most directly related to the content of the passage, had the lowest guessable scores.

Implications. To address RQ1, GPT-4o was most accurate in calculating the answerability of Retrieval and Inferencing questions. It underestimates the answerability of Integration questions because of their more challenging nature. Evaluation questions were similarly penalized because of their higher guessability.

To answer RQ2, we evaluated direct use of TI in question selection for MC items. For the 85 passages containing questions of all four PIRLS categories, we ranked the questions in each passage by TI score. Consistent with the observations above, there was a preference for the Retrieval and Inferencing categories: 56.5% of the highest-scoring questions were either Retrieval (23 passages) or Inferencing (25 passages). In contrast, there appeared to be a bias against the other two categories, with only 43.5% of the highest-scoring questions representing the Evaluation category (15 passages) or Integrating category (22 passages). This bias can hinder the construction of a well-balanced assessment item set with a variety of question categories (Mullis and Martin, 2019). To mitigate this issue, the TI score should be used for selecting questions only within the same PIRLS category, so that score differences could be attributed solely to question quality and not to question category.

6.3. Effects of Distractor Categories

Table 6 shows the automatically calculated TI scores attained by distractors in each STARC category, using the dataset described in Section 3.

Answerability. While GPT-4o succeeded in labeling most distractors as `false`, its performance varied according to distractor categories. For a competent reader, an Out-of-text distractor should be the least plausible, since it is not supported by the content in the passage; in contrast, an In-span distractor, which is based on the critical span, should be the most plausible. Dovetailing with this expectation, GPT-4o was most capable of recognizing Out-of-text distractors as `false` (98.97%), followed by Out-of-span distractors (98.77%). It was most often misled by In-span distractors (95.27%).

Guessability. Out-of-text distractors are aimed at low-proficiency students who, unable to understand the passage, judge the answer option based on general knowledge rather than the passage. Consistent with this design, when not allowed to read the passage, GPT-4o most often failed to label Out-of-text distractors as `false`, leading to low guessability (37.86%). It had greater success in judging Out-of-span distractors as `false` (47.53%) since they are derived from irrelevant text spans and often contain implausible content.

Implications. To address RQ1, on the one hand, GPT-4o was least accurate in calculating answerability for In-span distractors, which are designed to be plausible even for a competent reader. On the other hand, it may favor Out-of-text distractors by assigning them lower guessability (labeling them as `true`), since they are designed to appeal to students who judge them based on common sense.

To answer RQ2, we investigated the degree to which TI may penalize In-span distractors and favor Out-of-text ones using a TI-based selection criterion. This criterion requires a distractor to be both answerable, i.e., labeled by the LLM as `false` when the passage is available; and not guessable, i.e., labeled by the LLM as `true` when the passage is not available. Among the distractors in our STARC dataset, 61.11% of the Out-of-text distractors met this criterion, compared to only 55.56% of the In-span distractors and 51.44% of the Out-of-span distractors. This shows that naive application of answerability and guessability could lead to an overuse of Out-of-text distractors in MC items, at the expense of the other distractor categories. These results reinforce our recommendation to use TI scores only for within-category comparisons.

7. Conclusions

We have presented an in-depth analysis on automatic evaluation of MC items using Text Informativity (Säuberli and Clematide, 2024) on English and Chinese datasets annotated in the PIRLS and STARC frameworks. Results showed that GPT-4o underestimates the answerability of question and distractor categories that are more challenging, and that some categories have inherently lower guessability. Naive use of TI can lead to systematic preferences for more straightforward questions and Out-of-text distractors. It is recommended that practitioners use TI scores to compare only questions and distractors that are within the same category.

8. Bibliographical References

B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, D. Kulkarni, C. S. Seow, R. Wong, and M. T.-H. Co. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*, 18(8):Article e0290691.

Sabina Elkins, Ekaterina Kochmar, Jackie Chi Kit Cheung, and Iulian Vlad Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proc. 14th Symposium on Educational Advances in Artificial Intelligence*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proc. ICLR*.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proc. 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, page 477–491.

K.V. King, D. A. Gardner, S. Zucker, and M. A. Jorgensen. 2004. *The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics*. Pearson.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, page 11260–11285.

Ina V. S. Mullis and Michael O. Martin. 2019. *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement.

Andreas Säuberli and Simon Clematide. 2024. Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models. In *3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, page 22–37.

Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk. 2022. Towards Human-Like Educational Question Generation with Large Language Models. *Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science*, 13355.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page 610–625.

9. Language Resource References

Yevgeni Berzak and Jonathan Malmaud and Roger Levy. 2020. *STARC: Structured Annotations for Reading Comprehension*. Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Reflexive Research with LLMs: Considering the positionality of users and systems

Eleanor Smith, Luis Morgado da Costa, Antske Fokkens
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
e.l.t.smith, l.g.de.passos.morgado.da.costa, antske.fokkens@vu.nl

Abstract

Previous work has found that people often perceive computational systems as neutral tools (van Es, 2023), and yet these systems are not developed or deployed within a vacuum. As the popularity of Large Language Models (LLMs) in digital social science and humanities (DSSH) research increases, it is important that we reflect both on our positionality as researchers regarding how we are primed to interact with these systems and the positionality of the systems themselves as defined by their design and training. This paper presents a model of factors and interactions affecting the use of LLMs in DSSH research and argues that explicit discussion of both human biases, which affect how we interact with systems, and the potential biases encoded in systems are needed in conjunction with strong case specific system evaluation when developing methodologically sound applications of LLMs.

Keywords: Generative LLMs, Positionality, Bias

1. Introduction

The question of methodology in DSSH research is not new. Rieder and Röhle (2012) provide a strong reminder of this, showing that there has been reflexive work on digital methodologies for many decades. The article quotes the 1966 inaugural issue of the journal *Computers and the Humanities*, stating that "we need never be hypnotized by the computer's capacity to count into thinking that once we have counted things we understand them" (Rieder and Röhle, 2012, p.71). Despite their hype and anthropomorphization, Large Language Models (LLMs) should not bypass this consideration.

Ries et al. (2024) observe that exploring and describing the biases of computational models is a humanities question. There is a long tradition among SSH disciplines of acknowledging bias originating from the context of the researcher by considering their positionality (Selka, 2022).

DSSH also has a tradition of evaluating information and its encoded and contextual biases through source and tool criticism (Koolen et al., 2018). Source criticism is the practice of analyzing the credibility, reliability and authenticity of information sources (Backerra, 2024). Tool criticism extends this to reflect on the impact of tools on the data they interact with (Koolen et al., 2018), with some also including the interaction between user and tool within the practice (van Es et al., 2018).

The advent of chat based interaction with LLMs has increased the accessibility of applying digital methods to SSH data, while also obscuring the nature of these methodologies behind anthropomorphized, and in several cases proprietary, black box models. This, coupled with the framing of technology generally and AI specifically as neutral, contributes to the growing hype around using LLMs in research contexts. But just like other tools, these systems encode biases, as do the ways we interact with them.

Sections 2 and 3 discuss the different biases at play in the positionality of both the user and the LLM. These biases interact with one another to influence the contexts and outcomes of system interactions. Figure 1 provides a representation of these interactions. In this paper, we argue that the already developed practices of positionality, source, and tool criticism from previous SSH research should continue to be carefully applied to both ourselves as researchers and the systems that are used in digital methodologies. These practices, coupled with clearly reported system evaluation, constitute an effective way of impacting user level factors in our proposed model. The upward arrows presented in Figure 1 illustrate that through changes in user level factors, it is possible to influence the wider system of interactions involved in research with LLMs.

We first outline prior work on how AI systems are viewed by users, followed by an overview of known biases of language models. Taking user positionality as a starting point clearly illustrates the risks of LLMs becoming more prominent tools in research. We end with recommendations on how to mitigate these risks and increase methodological soundness.

2. Positionality of the User

Considerations of researcher positionality are well established. We aim to highlight in this paper that human-computer interaction is a specific context

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 175–186
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

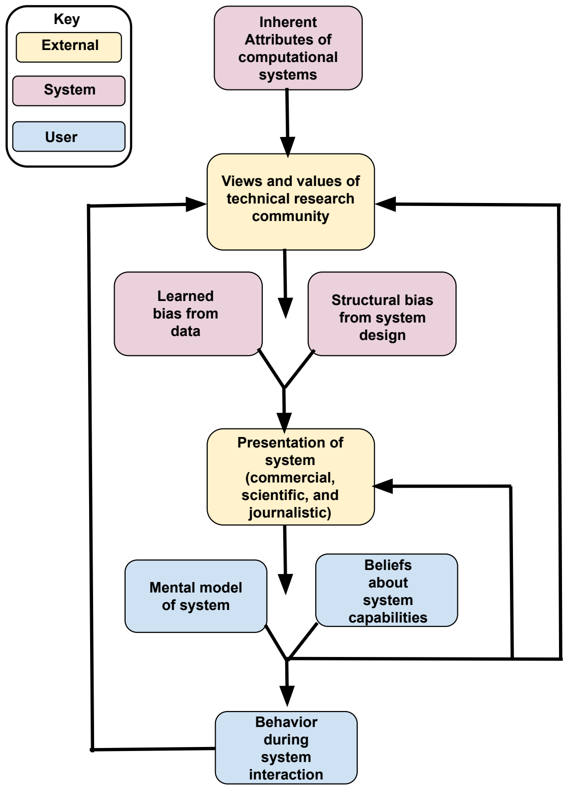


Figure 1: Factors and interactions affecting the use of LLMs in DSSH research

which has the ability to prime certain unconscious behaviors and biases in users. We argue that digital research should reflect on the position that the researcher holds as a user of an automated system. The concepts discussed in this section relating to users are operationalized in Figure 1 as user level factors, shown in blue.

2.1. People's Beliefs about Systems

Carlson (2019) defines the term *mechanical objectivity* as a belief in technology capable of rendering a particular output in a manner that overcomes the limits of human subjectivity. This belief is not new. Carlson discusses how ideas of *mechanical objectivity* were applied to photographs when they first replaced sketches in journalistic publications. At the time, people lauded this technological development as providing 'objective images'. Although photographs provide to some degree a more faithful rendering of events than an artist's sketch, there are still many decisions that contribute to the final version of an image in a newspaper, all of which are obscured by ideas of *mechanical objectivity*. A contemporary version of this belief can be seen in the ideas of *mechanical objectivity* applied to the concepts of big data and by extension AI. Carlson cites Wired magazine's editor in chief from a 2008 interview stating that "[w]ith enough data, the numbers speak for themselves" (Carlson, 2019, p.1125), this view that large volumes of data equate to truth is prevalent. Barocas and Selbst (2016) discuss how this false equation of data and truth can lead not just to discrimination against groups unrepresented in these data sets, but also how the emergent quality of this discrimination obscures its presence and makes legal arguments against it particularly difficult.

Findings from research in the Netherlands illustrated that generally decisions made by automatic systems were seen as equivalent or better than those made by human experts (Araujo et al., 2019). Respondents indicated both a potential distrust in human decision making and the application of ideas of *mechanical objectivity* to automatic decision making systems which they believed were in some way free of the biases of the human experts.

Sundar (2008) discusses how ideas of *mechanical objectivity* can be triggered by what they call the *machine heuristic*. Sundar argues that the *machine heuristic* is the mental shortcut triggered by interacting with an interface which appears machine-like leading to the attribution of characteristics associated with *mechanical objectivity* to its performance. Sundar notes that triggering the *machine heuristic* may lead to more positive credibility judgments.

A person's *mental model* of an LLM is their personal understanding of how the system works based on their own interactions with both LLMs as well as discourse on the topic. Generally people expect machines to be highly predictable, precise and consistent, lacking in both emotional and general understanding (Schneider, 2025). The anthropomorphized chat functions of many LLMs simulate human-like output potentially influencing users' *mental models* of how the system functions and what it is capable of (Sharkey and Sharkey, 2007). Schneider analyzes 200,000 human-LLM interactions and finds that users increasingly adopt conversational behaviors typical of human-to-human communication, starting from their second conversational turn. This suggests a transition in the way users perceive the system from a *mental model* of the system as a machine to a *mental model* closer to that of a human interlocutor as the interaction progresses. Cabrero-Daniel and Sanagustín Cabrero (2023) found that users had mixed *mental models* of LLMs, with some appearing to apply ideas of *mechanical objectivity* labeling the system as an impartial entity due to their *mental model* of the system as machine-like, while others expected LLMs to understand the emotional landscape of the user, something generally seen in *mental models* of other humans (Schneider, 2025). This potentially shows the influence of anthropomorphization on user *mental models*. Wang et al. (2023) report that there is a general lack of understanding and awareness of how bias affects LLMs, with system output being seen as potentially more accurate and 'up to date' than other online sources such as Wikipedia showing a lack of users' understanding of the quality and provenance of the information in LLMs. Cabrero-Daniel and Sanagustín Cabrero reported that although most participants in their study agreed that gender bias was present in the output of LLMs, half of them thought that gender bias was inherent to natural language generation and that the system was as good, or better, than humans at avoiding gender bias. This shows the reluctance of users to categorize this bias as error in their *mental model* of the system. This matters, because an inaccurate *mental model* of the system can damage users' ability to interpret and evaluate LLM responses accurately and effectively, making it more difficult for them to recognize errors and inconsistencies in output, leading to inappropriate levels of trust in the system (Eigner and Händler, 2024).

2.2. What Feeds into these Beliefs

Beer (2017) discusses how inherent attributes of the system feed into the perception of systems as more powerful than humans. This is a system level factor and is shown in Figure 1 in red. While LLMs surpass certain human capabilities in terms of speed and volume of information processing, these attributes do not ensure that systems are more accurate than humans (Rieder and Röhle, 2012). As the technical report of GPT4 (Achiam et al., 2023) acknowledges: the system is capable of making simple reasoning errors and is often "confidently wrong" (Achiam et al., 2023, p.10). Williams and Huckle (2024) develop a benchmark test focused on known limitations of LLM capabilities and test several widely used models. Their findings show that LLMs struggle with logical reasoning, spatial intelligence, mathematical reasoning, linguistic understanding, knowledge of popular science, and relational perception. This highlights that some tasks which are easy for humans are still very difficult for LLMs.

The way that a system is presented to its users also contributes to the *mental models* that users build. This factor is external to both the user and the system and is shown in orange in Figure 1. LLMs are positioned within a consistent hype around their capabilities with systems often being portrayed as objective and impartial (Araujo et al., 2019). Bender and Koller (2020) discuss the use of misleading language in both academic and journalistic publications which frame LLMs as being able to 'understand' or 'comprehend' the meaning of text, potentially compounding the effect of anthropomorphization and leading users to conceptualize systems as more human-like. Bender and Koller argue that LLMs work strictly with form and never meaning and thus cannot be seen to 'understand' any form of textual data. While it can be difficult to find the right words to convey what models do, leaning too heavily on metaphorical uses of verbs like 'understand' and 'comprehend' which align closely with inconsistencies already present in users' *mental models* of LLMs may further obscure the methods of the system.

Regarding hype, Narayanan and Kapoor's (2024) review of news journalism on the topic of AI found that articles often repeat PR statements, use images of robots, and downplay limitations. The research of Kapania et al. investigates attitudes towards AI in India, where the use of AI is seen as aspirational and technology is generally discussed optimistically. Their findings indicate that overly optimistic narratives played a key role in legitimizing *AI authority*. *AI authority*, defined as the power of AI to influence human actions without adequate evidence of system capabilities, was also linked to a higher tolerance for harm and lower recognition of bias. This illustrates the strength of potential knock on effects of overly positive, unbalanced reporting on AI.

The lack of clear and accessible reporting on LLMs also leads to increased uncertainty and reliance on the idea of these models as unexplainable in general discourse, further feeding ideas of AI as a mythologized entity (Doherty, 2024). Spatola and Urbanska (2019) investigate semantic representations of AI and robots in comparison to natural entities such as humans and animals and divine entities such as gods. Findings showed that at both an explicit and implicit level participants had semantic overlap between concepts relating to AI and robots and concepts relating to divine entities. Work by Karataş and Cutright (2023) adds more weight to the association of AI and divine entities, finding that thinking about God and religion directly before a task leads participants to be "more willing to consider AI-based recommendations" (Karataş and Cutright, 2023, p.1). The overlap in conceptualizations of AI with divine entities is problematic as presenting systems as unknowable reduces human agency and makes systems difficult to challenge (Narayanan and Kapoor, 2024).

2.3. The Kind of Behavior this Leads to

Kapania et al. (2022) discuss the power that humans give to systems. The power of LLMs has been legitimized through their broad deployment and the their high adoption rate giving them strong social capital. This section provides more detail on how users over-rely on system output, allowing the advice of the system to override their own judgment, crowd sourced advice, and the advice of experts.

In their experiments, Logg et al. (2019) find that participants consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic as opposed to human source. They label this behavior *algorithmic appreciation*. Multiple experiments have found evidence for *algorithmic appreciation* across different contexts and groups. In turn, Klingbeil et al. (2024) define *Over-reliance* as the behavior of following machine outputted advice even when it contradicts clearly available context information and when this leads to inferior outcomes. Klingbeil et al. show in their research that labeling advice as being generated by an AI system was enough to cause over-reliance. Gunaratne et al. (2018) compare the persuasive power of advice labeled as algorithmic to advice labeled as crowd sourced, finding that algorithmic advice is significantly more persuasive than advice based on the aggregation of peer behavior. Eric Bogert and Watson's (2021) results show a similar pattern of behavior which becomes even stronger as the participants' task becomes more difficult, concluding that labeling advice as being derived from machine learning causes a meaningful shift in human behavior. Results from Liel and Zalmanson (2020) corroborate these findings, showing that participants significantly conformed to recommendations when they were labeled as algorithmic and reported high confidence in the system's estimates even when the advice was clearly incorrect during a simple image classification task. This finding suggests potential *automation complacency* in the participants interactions with the system.

Parasuraman and Manzey 2010 discuss *automation complacency* in detail, defining it as the phenomenon of poorer detection of system malfunctions under automation compared to manual control. When conducting research using digital methods this could manifest in lack of system monitoring and lower likelihood of detecting errors in system output. Alexander et al. (2018) measured neurophysiologic responses of participants while deciding whether to trust an imperfect 'helper algorithm'. Results showed that information representing how many peers had chosen to adopt the algorithm was more influential in participants' decisions than providing more detailed information about the accuracy of the system. Participants who were given neither social nor accuracy information showed lower cognitive engagement throughout the task. This finding reveals potential *automation complacency* as it suggests that participants did not monitor the algorithm when there was no information provided despite this being the riskiest context.

3. Positionality of the System

In this section we consider the positionality of LLMs. We see the views encoded in their design and training data as denoting their position. The hype, framing and presentation of chat based LLMs make it all the more difficult to remember that "our digital helpers are full of 'theory' and 'judgement'" (Rieder and Röhle, 2012, p.70). As such, this contribution is an important part of evaluating the use of LLMs in DSSH. The factors related to the system discussed in this section are operationalized as system level factors and are shown in red in Figure 1. The make up of the training data of LLMs can be evaluated using methods of source criticism and the mechanisms at work within the LLM can be viewed through the lens of tool criticism. Understanding the principles behind systems can enable users to critically engage with tools on a theoretical level, even if they are not technically literate enough to understand the specifics of the code used to implement the tool (van Es et al., 2018). We thus first provide a high level explanation of the principles at play in the system design of LLMs.

3.1. System Design

The goal of creating generative LLMs was not to provide an all-purpose system that can accurately answer questions or provide advice. The goal was to build a computational model that could produce human-like text. The fact that LLMs' text feels relevant and coherent doesn't make them trustworthy (Shah and Bender, 2022), it makes them convincing mimics (Bender et al., 2021). LLM output is designed to seem appropriate not to be accurate (Townsen Hicks et al., 2024), this is particularly an issue considering that many people use models to generate output that they are uncertain of, for example using an LLM to write code to use a package that they are unfamiliar with. The LLM will produce code that looks correct, and the user is not familiar enough with the package to easily spot potential errors.

Townsen Hicks et al. (2024) provide a clear overview of the goals and mechanisms of generative LLMs using the term *bullshit* to describe the truth agnostic but probable seeming nature of system output. Townsen Hicks et al. define *bullshit* as "[a]ny utterance produced where a speaker has indifference towards the truth of the utterance" (Townsen Hicks et al., 2024, p.38). An example of human *bullshit* might be a student who did not complete the pre-reading for a class and is not willing to admit so when called on to speak during a discussion. The student may say something that seems probable given the previous contributions to the discussion. They choose their statement with no regard for whether it is true, but merely based on their experience from similar situations. This analogy conceptualizes the general mechanism at work when LLMs generate text.

The basic concept behind how LLMs are trained is through next word prediction, the task of providing a probable word to fill a slot based on the previous text. In order to generate probabilities and find likely words the LLM constructs a large statistical model based on a vast amount of text which is provided during training. In our analogy, the training data is our ill-prepared student's previous experience of academic discussions. If at their previous institution all academic discussions were conducted in an aggressive adversarial style and their new institution prefers calm respectful interactions, then their statistical model of interactions in this context will not allow them to choose a statement that fits the discussion at hand, instead prompting them to potentially offend their interlocutors. The content of the training data constrains the possible behaviors of the LLM, as such the decisions of what to include in training data have a very strong influence on the type of model that is created.

For models with chat capabilities, fine-tuning is performed to align the responses of the model to what users expect for a given prompt. The main methodology used for this is instruction fine-tuning followed by preference alignment using reinforcement learning through human feedback (RLHF). This is necessary as the language modeling objective is different from the objective of following a user's instructions in a helpful and safe manner (Ouyang et al., 2022). Instruction fine-tuning is the step of training the model in the structure of questions and answers. During this stage the LLM is exposed to a large number of example questions and replies. Preference alignment is then performed in order to train the model to produce output which aligns more closely with what users want. The process of RLHF starts with human labelers who are asked to rank several system outputs for the same prompt. A reward model is then trained on these rankings to predict which output human labelers prefer. The reward model is then used as a function of the LLM to incentivize model behavior that aligns more closely with the human labelers' preference (Ouyang et al., 2022).

3.2. Structural Bias

We use the term *structural bias* to refer to biases introduced into the system via the design process, for example through the choice of task formulation, metrics, and training data (Liu, 2023; Hovy and Prabhumoye, 2021). The impact of all of these choices is often amplified through a lack of transparency (Liu, 2023). Design choices are driven by the views and values of the research community that is developing these models. As such, it is important to consider the culture of machine learning and natural language processing research. This is another external factor influencing the interactions between users and systems and is shown in orange in Figure 1.

Results from Birhane et al.'s (2022) qualitative analysis of 100 highly cited machine learning papers show that papers most frequently justify and assess themselves based on performance, generalization, building on past work, quantitative evidence, efficiency, and novelty. Birhane et al. highlight how these values are often viewed as purely technical, without consideration for how the dominance or operationalization of these values can quickly become political when they are pursued at the expense of other more ethically informed considerations. In a second analysis (Birhane et al., 2022) also found that papers with corporate ties increased by 34% in a ten year span. The breadth of companies represented in these ties also showed a shift, with the presence of a small number of very large tech firms increasing nearly fourfold. When considering university affiliations, Birhane et al. found that 80% were from within the top 50 universities by QS World University Rankings. The dominance of a small number of elite universities and big tech firms in publications is worrying as it may lead to a homogenization of values and a centralization of power within a small subset of the field.

One aspect of *structural bias* is task formulation. This encompasses aspects of how the system's use is conceptualized and operationalized through its problem definition, training objectives, and interface design. The main training objective during LLM pre-training is next word prediction (see 3.1). This objective allows the system to build a strong model of distributional semantic properties but also contributes to building biased representations. Hovy and Prabhumoye (2021) discuss how this training objective in combination with large relatively unfiltered training data takes a value neutral stance on whether the most likely next word predicted by modeling the training data represents a view or value that we wish the model to perpetuate. Another potential issue relating to task formulation is that LLMs are expected to always generate an output, even when there is uncertainty in the model or when training data is unable to provide adequate information (Hovy and Prabhumoye, 2021). Prioritizing certain use cases can also cause *structural bias* in the system. If an LLM is designed to cater to specific demographics or industries, then it may as a byproduct reinforce the biases of these groups (Ferrara, 2023). The choice to release many LLMs with chat interface is another form of *structural bias*, as it shapes the ways in which users interact with the system and potentially what they expect from it (see 2).

LLMs are often evaluated by their ability to perform downstream tasks such as classification. The choice of metrics used to measure and evaluate the behavior of the system can introduce *structural* *bias*. Traditionally the field has used metrics such as recall, precision, and F1 to evaluate systems, but in order to build a full picture of the behavior of the model metrics should take into account performance across all diverse groups represented in the data (Liu, 2023). Previous work has found that by focusing on the robustness of model behavior, researchers can gain more insight on performance than through performance metrics alone, while also providing safe guards against releasing models which systematically under perform for some groups (Hovy and Prabhumoye, 2021).

Concerning bias introduced by training data (see 3.1), previous work generally refers to the problem of unbalanced data in the training set as selection bias (Hovy and Prabhumoye, 2021). The consequences of unbalanced selection are more broadly explored in 3.3. The process of fine-tuning LLMs through RLHF may introduce *structural bias* through the selection of the labelers themselves, the examples they are annotating, and the annotation schema they follow (Søgaard et al., 2014).

The general bias towards English language is well known. English is a relative outlier in its linguistic specificity, and yet its dominance has meant that approaches that work well for English have become the default (Hovy and Prabhumoye, 2021). Hovy and Prabhumoye state that it's improbable that n-gram based approaches would have become a focus in the field if the predominant language was morphologically complex. The underlying n-gram concept is present in LLMs and influences the way in which they represent input for all languages.

Liu (2023) posits that a lack of transparency in LLMs also contributes to the *structural bias* of the models. Using methods that are not easily interpretable is a choice, and so is presenting models to users as a black box. Section 2 covered in detail how the lack of accessible information on a system's working can lead to incomplete *mental models* and inappropriate system use.

3.3. Learned Bias

We use the term *learned bias* to refer to biases introduced to the system during the training process. Bias can be introduced in several ways in this process: through the contents of the training data, through fine-tuning procedures, and finally through content moderation filters (Hartmann et al., 2023).

Training data for LLMs is incomplete, imbalanced and inaccurate as it mirrors human biases in its collection and processing (Liu, 2023). Many of the latest releases of proprietary LLMs do not share details of the data that they are trained on (Lee et al., 2023), with Achiam et al. (2023) stating that this is due to competition and safety concerns. Despite this, we do have more information about earlier iterations of OpenAI models. Brown et al. (2020) report that GPT-3, was trained on multiple datasets: Common Crawl (unfiltered), WebText2, Books1, Books2, and Wikipedia. This is a huge volume of data, but the majority of it is internet based text. The content is skewed towards the most salient beliefs and cultures in online discourse (Kuntz and Silva, 2023; Arora et al., 2023). The bias of this content is affected by both the demographics of the authors (Liu, 2023) and the groups that are the topics of these discussions (Lee et al., 2023).

Concerning RLHF, Ouyang et al. (2022) stress the importance of considering the methodology involved in preference alignment as these choices determine who we align to. They argue that for OpenAI models three entities impact the model's alignment: the labelers through their preferences, the researchers through their instructions and demonstrations to the labelers, and the OpenAI customers who submit prompts to the OpenAI API Playground, which is used to select training data. The results of fine-tuning for better alignment showed improvements in truthfulness and toxicity over GPT-3, but not bias. Xiao et al. (2025) argue that RLHF as a method suffers from inherent algorithmic bias which in extreme cases could lead to preference collapse in which minority preferences are ignored. McIntosh et al. (2024) second this view, concluding that RLHF is often unable to represent a diverse set of human values, aligning the model towards a select group: those who are in control of the LLMs. These insights show that even if we can ensure diversity in the views of labelers, researchers and customers, the mathematical operations that perform the alignment must also ensure that minority views are maintained and not flattened.

Bias can also be reduced, introduced or reinforced by content moderation filters. Policy decisions about what kinds of content should be moderated and how are made by the teams behind the models. The decision of which norms should be encoded in models is complex (Ferrara, 2023). The current concentration of power in AI as the domain of a few high profile companies and institutions (Birhane et al., 2022) means that these decisions are generally being made by certain dominant demographic groups without the input of others.

When bias is present in a model it can affect the representation of groups in multiple ways as the examples below illustrate. Several studies investigated gender based stereotyping in LLMs. Lucy and Bamman (2021) tested story generation with GPT-3, finding that the model exhibited many gender stereotypes even when prompts did not contain any explicit gender cues. Following this, Kotek et al. (2023) found that when assigning careers to pronouns, LLMs are 3-6 times more likely to choose a gender stereotyped occupation. This bias is more pronounced than in human perceptions. Model output can thus aggravate the existing difference between perception and ground truth.

Atwell et al. (2025) show consistently lower agreement between model and human judgments on projectivity of clausal complements when the subject of the clause is female. When prompted with '*X debated that a particular thing happened. Did that thing happen?*', the model correctly answered 'maybe or maybe not' when 'X' was 'someone' or 'a man', but incorrectly answered 'no' when 'X' was 'a woman'. This behavior did not align with human judgments for the same prompts. This may be a consequence of under-representation of texts written by women in the model training data (Atwell et al., 2025), as Kuntz and Silva 2023 estimate that only around 26.5% of the GPT-3's training data was composed by women.

Using GPT-2, Sheng et al. (2019) found negative associations of 'black', 'man', and 'gay' demographics with contexts relating to respect, as well as negative associations of 'black', 'woman', and 'gay' demographics related to occupation. Cheng et al. (2023) investigated biases in GPT-3.5 and GPT-4 by prompting the models to generate personas with different demographic characteristics. Results showed "higher rates of racial stereotypes than human-written portrayals" (Cheng et al., 2023, p.1504). When contrasting the terms which most characterized the content of descriptions for non-white non-male personas in comparison to white male personas, patterns of othering and exoticizing were found in the model output.

Nguyen et al. (2025) performed qualitative analysis of LLM generated narratives where one character was framed as American and another as originating from a country in the 'global south'. Their research found persistent colonial stereotypes. The characters originating from the 'global south' were depicted as intellectually inferior and undesirable compared to American characters who were depicted as culturally and hierarchically superior. Characters described as originating from Mexico and China were more likely to be shown in servitude. Narayanan Venkit et al. (2023) found prejudices against certain countries. Specifically, countries with lower representation online tended to have lower sentiment scores, with the model perhaps mimicking the view of these countries presented by internet users of different nationalities as opposed to relying on actual representations.

Results from Lee et al. (2023) "find that LLMs portray African, Asian, and Hispanic Americans as more homogeneous than White Americans" (Lee et al., 2023, p.1). This may be due to a richer representation of white individuals in the training data of LLMs, giving the model access to more varied examples for this group. This may be compounded by potentially stereotypical representations of other groups in the training data, leading to a smaller and narrower pool of representations for these groups.

Work from Li et al. (2024) showed that geopolicital biases in LLMs can be uncovered by prompting the model to discuss disputed territories in different languages. With the model expressing the beliefs about ownership of land which most closely align with the cultural connotations of the language it was prompted in. Nguyen et al. further uncover that LLMs systematically omit to mention some under-represented countries from the 'global south' in their output, with African nations being affected most severely by this behavior.

LLMs have been shown to encode biases against specific religious groups (Abid et al., 2021). Abid et al. (2021) found strong and consistent associations between Muslims and violence in the representations of GPT-3. The term 'Muslim' was analogized to 'terrorist' in 23% of cases.

As well as producing stereotyped associations, the biases encoded in LLMs can lead to other less obvious behaviors. For example, lower performance on certain topics may lead to consistently sub-par performance for some groups of users. This produces bias in the accessibility of systems and is not visible for the affected parties as it is only apparent in comparison to performance for others.

Prior research has also aimed to characterize and align the values of LLMs. We use the term 'values' to describe interconnected systems of biases and beliefs encoded in models. Münker (2025) show significant differences between AI and human moral intuitions, finding that models homogenize moral diversity. The results highlight systematically better representation of Western vs non-Western cultural contexts, as models struggled to represent the belief structures of under-represented groups in the training data. This finding suggests that current models should not be trusted to generate accurate culturally diverse synthetic populations. Work from Arora et al. (2023) corroborate this finding, detecting some differences in value representations of different cultures from model output, but concluding that these trends only weakly align with actual data from these cultures as recorded by value surveys.

Results from Durmus et al. (2023) show that the opinions and dominant values encoded in LLMs most closely resemble those of the USA, Europe and South American countries. When models were prompted to highlight perspectives from under-represented countries responses give opinions more similar to the target group but also show harmful stereotypes. Work has also shown that ChatGPT generally takes a left libertarian political standpoint (Hartmann et al., 2023; Santurkar et al., 2023). Research from Johnson et al. (2022) tests the reaction of GPT-3 to several documents in order to expose the cultural values included in the model. Results show that the model aligns most closely to American value systems, associating firearms discussions with loss of rights, putting feminism at odds with equality, showing pro-life stances on abortion, as well as conflicts with different perspectives on immigration and secularism. These findings lead the authors to conclude that "the 'ghost in the machine'[...] just may have an American accent" (Johnson et al., 2022, p.8).

The idea of using personas to steer the value representations present in LLM output to better represent a target group has received some enthusiasm. Sommerauer et al. (2025) investigate personas in LLM prompts by measuring linguistic abstraction as a marker of stereotyping. All LLM generated texts in the study showed a level of abstraction associated with stereotyped biased descriptions, and use of persona prompting did not meaningfully change the levels of abstraction in responses. These findings highlight the prevalence of the generalizing descriptive behavior of LLMs and raise concerns that using personas could be damaging in that the method seemingly evokes the voice of an under-represented or marginalized group, while continuing to produce stereotypically abstract descriptions. Results from Münker (2025) corroborate these findings, showing responses were often statistically indistinguishable. This suggests that the differences found in responses during persona prompting are at a surface level, with the underlying values of the model remaining unchanged across responses (Münker, 2025).

4. Affects on DSSH Research

Biased systems can still be useful as long as their limitations are taken into account (Ferrara, 2023) and LLMs can undoubtedly provide valuable new methods for DSSH research. Research has shown that LLMs perform very well across multiple task types (Hagos et al., 2024), with clear jumps in state of the art performance since the implementations of the first transformer based models such as BERT.

As discussed in 1 the upward arrows in Figure 1 show that as users of automated systems we are able to influence system external factors by changing our *mental models*, beliefs about system capabilities, and behavior during system interactions. An effective way to influence these factors is through strong and clearly communicated evaluation of the behavior of LLMs in research contexts. This way, the community can disambiguate the genuine capabilities and usefulness of LLMs for the field from the general hype around these models. In light of the bias discussed throughout this paper, we advocate for evaluation methodologies which consider the robustness of models' outputs across all diverse groups represented in data (see 3.2).

The accessibility of chat based LLMs is a key strength, as they may be useful for initial or exploratory research by SSH researchers without strong coding skills. Tools for downstream tasks that LLMs are often used for have been around for many decades, with research considering different implementations. We argue that in order to lessen issues with both bias and interpretability, collaboration with more technically trained colleagues may be necessary after these first exploratory steps, and that using alternative tools and methods for these tasks is sometimes more appropriate.

Considering alternatives to large general models is also an effective way to influence the interactions shown in Figure 1, as moving the focus of the field away from a single model type has the potential to impact both the values of the technical research community and the presentations of systems in publications. Large general models do not always provide the best performance on SSH data. Previous DSSH research has shown that smaller domain adapted or scratch trained models using BERT style architectures can outperform prompting and finetuning approaches with larger general models even with relatively small amounts of labeled data (Verkijk et al., 2025; Bosley et al., 2023).

It is potentially impossible to remove all biases from pretrained models (Waseem et al., 2021). We posit that research exposing biases is integral, and that reflection on how specific biases interact with the data and research questions of a research project is a necessary step in the study design process. In Sections 1, 2 and 3 we laid out our argument for the use of reflections on positionality, source criticism and tool criticism in the reflexive use of LLMs, a contribution to the field at large that DSSH is uniquely positioned to make (Rieder and Röhle, 2012; van Es et al., 2018).

5. Conclusion

This paper hopes to highlight that through the reflexive approaches of source criticism, tool criticism, and consideration of positionality DSSH researchers have the necessary skills and opportunity to be on the front lines of situated and methodologically sound research using LLMs. We present a model of factors and interactions affecting the use of LLMs in DSSH research to illustrate the interrelatedness of the issues at hand and argue that changes on the user level prompted by reflective practices and clear reporting of system evaluations provide researchers the influence needed to affect the larger system of factors at play in research with LLMs.

6. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Veronika Alexander, Collin Blinder, and Paul J. Zak. 2018. Why trust an algorithm? performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89:279–288.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese1. 2019. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI and Society*.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Katherine Atwell, Mandy Simons, and Malihe Alikhani. 2025. Measuring bias and agreement in large language model presupposition judgments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2096–2107, Vienna, Austria. Association for Computational Linguistics.

Charlotte Backerra. 2024. Source criticism for cultural history. *Rethinking History*, 28(2):194–216.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review*.

David Beer. 2017. The social power of algorithms. *Information, Communication & Society*, 20(1):1–13.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Mitchell Bosley, Musashi Jacobs-Harukawa, Hauke Licht, and Alexander Hoyle. 2023. Do we still need bert in the age of gpt? comparing the benefits of domain-adaptation and in-context-learning approaches to using llms for political science research. In *2023 Annual Meeting of the Midwest Political Science Association (MPSA)*.

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.

Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. 2023. Perceived trustworthiness of natural language generators. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23, New York, NY, USA. Association for Computing Machinery.

Matt Carlson. 2019. News algorithms, photojournalism and the assumption of mechanical objectivity in journalism. *Digital Journalism*, 7(8):1117–1133.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Rachel Doherty. 2024. Deified ai: The relationship between gods and artificial intelligence. *Mid-Atlantic Humanities Review*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Aaron Schecter Eric Bogert and Richard T. Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35(4):1092–1120.

Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE transactions on artificial intelligence*, 5(12):5873–5893.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "because ai is 100% right and safe": User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Mustafa Karataş and Keisha M. Cutright. 2023. Thinking about god increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences*, 120(33):e2218961120.

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on ai — an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352.

Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2018. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34(2):368–385.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, page 12–24. ACM.

Jessica B Kuntz and Elise C Silva. 2023. Who authors the internet. *Analyzing Gender Diversity in ChatGPT-3 Training Data. Pitt Cyber: University of Pittsburgh*.

Messi Lee, Jacob Montgomery, and Calvin Lai. 2023. The effect of group status on the variability of group representations in LLM-generated text. In *Socially Responsible Language Modelling Research*.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. 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 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Yotam Liel and Lior Zalmanson. 2020. What if an ai told you that 2 + 2 is 5? conformity to algorithmic recommendations. In *International Conference on Information Systems 2020*.

Zhaoming Liu. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. The inadequacy of reinforcement learning from human feedback—radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1561–1574.

Simon Münker. 2025. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. *arXiv preprint arXiv:2507.10073*.

Arvind Narayanan and Sayash Kapoor. 2024. *AI Snake Oil : What Artificial Intelligence Can Do, What It Cant, and How to Tell the Difference*. Princeton University Press.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Ilana Nguyen, Harini Suresh, and Evan Shieh. 2025. Representational harms in llm-generated narratives against nationalities located in the global south. In *HEAL Workshop, CHI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410. PMID: 21077562.

Bernhard Rieder and Theo Röhle. 2012. *Understanding Digital Humanities*, chapter Digital Methods: Five Challeges. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Thorsten Ries, Karina van Dalen-Oskam, and Fabian Offert. 2024. Reproducibility and explainability in digital humanities. *International Journal of Digital Humanities*, 6(1):1–7.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Johannes Schneider. 2025. Mental model shifts in human-llm interactions. *Journal of Intelligent Information Systems*.

Stephen Selka. 2022. Positionality: Identity, standpoint and the limits (and possibilities) of fieldwork. *Fieldwork in Religion*, 17(1):92–100.

Chirag Shah and Emily M. Bender. 2022. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 221–232, New York, NY, USA. Association for Computing Machinery.

Noel Sharkey and Amanda Sharkey. 2007. Artificial intelligence and natural magic. *Artifical Intelligence Review*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. 2025. Simulating identity, propagating bias: Abstraction and stereotypes in llm-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Nicolas Spatola and Karolina Urbanska. 2019. God-like robots: the semantic overlap between representation of divine and artificial entities. *AI and Society*.

S. Shyam Sundar. 2008. The main model : A heuristic approach to understanding technology effects on credibility. In *Digital Media,Youth, and Credibility*.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*.

Karin van Es. 2023. Unpacking tool criticism as practice, in practice. *Digital Humanities Quarterly*, 17(2). Publisher Copyright: © 2023, Alliance of Digital Humanities Organisations. All rights reserved.

Karin van Es, Maranke Wieringa, and Mirko Tobias Schäfer. 2018. Tool criticism: From digital methods to digital methodology. In *Proceedings of the 2nd International Conference on Web Studies*, WS.2 2018, page 24–27, New York, NY, USA. Association for Computing Machinery.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637, Vienna, Austria. Association for Computational Linguistics.

Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. 2023. People's perceptions toward bias and related concepts in large language models: a systematic review. *arXiv preprint arXiv:2309.14504*.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Sean Williams and James Huckle. 2024. Easy problems that llms get wrong. *arXiv preprint arXiv:2405.19616*.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2025. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164.

Small Can Be Beautiful in LLMs for SSH: a Case for Bulgarian

Nikolay Paev, Kiril Simov, Petya Osenova, Teodor Valchev, Stefan Marinov
Artificial Intelligence and Language Technology
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Bulgaria
{nikolay.paev, stefan.marinov}@iict.bas.bg
{kivs, petya}@bultreebank.org
teodorvulchev@gmail.com

Abstract

In the paper we present a set of small LLM-based models for solving basic NLP tasks for Bulgarian — POS tagging, Lemmatization, Dependency parsing, Named Entity Recognition, Named Entity Linking, Event Annotation, among others. In order to create fine-tuned models for these tasks, we first pre-train models using architectures like BERT, Modern-BERT, and T5 with different sizes, over Bulgarian data only. For each of the tasks we report our approach towards the fine-tuning, the results from the experiments and also the evaluation. Then we define a way to visualize the results over HTML documents which contain the analyzed texts. Our rationale is as follows: most, if not all SSH research scenarios, need a reliable processing chains that can be customized with respect to the specific needs. These scenarios would also need proper visualization for human observation. We aim to provide such a basic LLM-based toolkit.

Keywords: Bulgarian pre-trained LLMs, LLM-based NLP models for Bulgarian, Visualization

1. Introduction

Although it is widely accepted that the power of the Large Language Models[1] (LLMs) is in their ability to solve more complex tasks such as *Question Answering*, *Summarization*, **Information Retrieval**, **Chatbots**, and many more, we aim at creating a set of LLM-based models for the basic NLP tasks in Bulgarian. These are *Tokenization*, *Part-of-speech tagging*, *Lemmatization*, *Parsing (constituent or dependency syntax)*, *Named Entities Recognition*, *Named Entities Linking*, *Word Sense Disambiguation*, *Event recognition*, *Co-reference Resolution*, *Textual Entailment*, *Sentiment Analysis*, and others. In this paper, we present models for most of the above tasks, except for the last three ones, on which we are working at the moment.

In our opinion, addressing even the more basic tasks[2] can be very useful to support research activities in areas like Social Sciences, Humanities, Linguistics. In addition, such tasks might help to solve more complex problems in these areas. The aforementioned tasks (as we demonstrate in this paper) could be solved by relatively small LLMs such as BERT, T5, and similar. In this way, many more experiments can be planned and conducted, thus opening opportunities to test different approaches and achieve better results. Here we present the state-of-the-art for Bulgarian for each of the following tasks: *Tokenization*, *Part-of-speech tagging*, *Lemmatization*, *Parsing (constituent or dependency syntax)*, *Named Entities Recognition*, *Named Entities Linking*, *Word Sense Disambiguation*, *Event recognition*.

Each of the basic tasks requires to be fine-tuned on a set of appropriate data. For Bulgarian we rely on datasets that are freely available such as **BulTreeBank**, **Bulgarian Event Corpus**, **The Bulgarian part of the Balto-Slavic Corpus**. In addition, we annotated some extensions of the aforementioned or new datasets to support and improve the models.

We are aware that LLMs with billions of parameters can also be used to solve these tasks. But

[1] We are aware of the fact that there is a tendency for assuming that LLMs have more than ten billion parameters. In this paper we adhere to a broader definition and we consider an LLM each language model based on a Transformer architecture. Thus, we work with transformer models with less than two billion parameters. Here we refrain from using the term *small language models*, because a new trend emerged in the development of the field, where this term is used for another type of transformer models — reduced versions of large language models with many more than 10B parameters. They are results from applying methods like *https://github.com/jamwithai/production-agentic-rag-course*, *Pruning*, and *Quantization* the more than 10B parameters models.

[2] Another classification of tasks in NLP is intrinsic vs. extrinsic tasks which are usually used in the evaluation of different NLP applications. The nowadays LLM approaches to NLP assume that intrinsic tasks are no longer necessary, because the users need output from the extrinsic category (Question Answering, Dialog Systems, Information Extraction, Summarization, .etc)

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 187–197
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

they are not ready off-the-shelf to do this as many of the users expect. We performed some experiments with ChatGPT 4.0 services to implement these tasks using a prompting approach. For tasks like POS, Lemmatization, Universal Dependencies (UD) Parsing, the results were worse than ours.[3] For the WSD task, both approaches perform similarly. We did not conduct experiments based on fine-tuning. We imagine that such fine-tuned models will be comparable. However, generally this means that the creation of manually annotated data is still a central task for such models.

The main contributions of this work include: (1) Pre-training of LLMs with a small number of parameters used for the fine-tuning of the NLP tasks. The good results demonstrate that such models are sufficiently effective for the tasks. (2) We implemented a set of fine-tuned models for Bulgarian NLP tasks, which improved the results of the models previously known to us. Some of the models are the first of their kind to be trained on only-Bulgarian data. (3) We created a scheme that enables the integration of all annotations and their visualization in a user interface.

The structure of the paper is as follows: in the next section we present some related works. In Sect. 3 we introduce the main language resources that are used in the fine-tuning of the models. In Sect. 4 all the models that we pre-trained and fine-tuned for the different tasks are presented. Sect. 5 describes how we incorporate the annotations produced by the different models within an HTML editor in order for the users to have the possibility to edit the documents and to have access to the annotations. Such possibilities will be very useful for interested colleagues in the area of Humanities. The last section concludes the paper.

Some citations in the paper are temporarily excluded in order to meet the anonymity requirement.

[3]For example, we tried asking ChatGPT about the morpho-syntactic features of the marked phrase in the sentence:
(**BG**) Struvalo mu se sramno da razkrie, che `e roden` brat na brodyaga s lice na prestypnik.
(**EN**) He felt ashamed to reveal that he `was the biological` brother of this vagrant with the face of a criminal.
The phrase "`was the biological`" in Bulgarian is homonymic to "`was born`". But in the example the real use of the form "roden/ADJ" (biological) is an adjective. However ChatGPT categorized it as a participle which would be correct in the case of "was born". We tried with several different prompts, but the result was the same. Our fine-tuned classification model predicts the correct use. Similar for other examples and tasks.

2. Related Work

For each of the tasks, we present here some related work. The LLM approaches to dependency parsing rely on some observations over the learning of syntactic information by the models. For example Zhou et al. (2023) show that prepositional phrase attachment poses the biggest challenge to understanding syntax by LLMs. The case study on the training dynamics of LLMs revealed that most syntactic knowledge is learned during the initial stages of training. In some cases, syntactic knowledge is encoded directly in the LLM. For example, Shen et al. (2021) propose a new syntax-aware language model — Syntactic Ordered Memory (SOM). The model explicitly models the structure with an incremental parser and maintains the conditional probability setting of a standard language model (left-to-right). In our case, instead of an incremental approach, we use predefined partial syntactic information. With respect to dependency parsing Özates et al. (2020) use special rules to introduce dependency relations between certain word forms in sentences. Each rule identifies some arcs within the dependency tree. In our implementation of dependency parsing, we follow the approach of McDonald et al. (2006) about a graph-based dependency parsing performed in two steps: (1) determination of dependency arcs in the syntactic tree — the immediate domination relation over the tokens in the sentence — for each token to find its immediate parent token (adding special token for the root of the sentence); and (2) labeling the selected arcs with the appropriate dependency relations.

Word sense disambiguation is the task of determining the sense of a polysemic word in a specific context. In our work, we started with a set-up very close to the one described in (Huang et al., 2019). In their work, they construct *context-gloss* pairs. These pairs are constructed by combining the word form for which we want to assign the correct sense in a context (the text in which the word form appears) and the glosses from wordnet. Each of these pairs are classified as correct or not, depending on the gold annotation. In this way, the training examples are constructed. There are some variations in the approach to constructing training examples depending on whether or not the selected word is marked. In some cases, additional knowledge is added to the examples, depending on the context of the senses within WordNet — see (Song et al., 2021b).

The event extraction task is very often defined as consisting of two subtasks: (1) *Event Detection* (ED), and (2) *Event Argument Extraction* (EAE) — see (Simon et al., 2024), (Lai, 2022), and citations within them. In the first task, the system is expected to identify the span and the type of events. Usually, the event detection starts with a trigger recognition since triggers anchor the events. The second task identifies the arguments (participants) of the event in the text and relates them to their roles. When the two tasks are solved separately by different models, the result is called a *pipeline* approach. In cases where the model solves both tasks together, the result is called a *joint* approach. We implemented a joint generative event extraction system that produces a formal textual representation of the extracted event information. More specifically, we follow the `Text2Event` approach to EE (Lu et al., 2021). `Text2Event` defines an end-to-end generative model that transforms an input of tokens into a linearized event structure.

The paper defines the linearized representation as an S-expression that for each event occurrence contains an expression pair containing the type *the type of the event* and *the corresponding text span* followed by a list of *role* and *span* pairs representing the arguments of the event mentioned in the text. The transformed dataset (for example, ACE, mentioned above) has then been used to train a `T5` encoder-decoder language model (Raffel et al., 2019b). In order to restrict the output to the required event representation, the authors explore the *constrained decoding* that provides a mechanism for exploiting the knowledge of an event schema to form the output.

For the visualization of the annotations, we rely on our own experience with the following systems: the GATE Teamware — (Bontcheva et al., 2013), the INCEpTION platform — (Klie et al., 2018), SpaCy: Industrial-Strength Natural Language Processing[4]. All of them have functionality for creation of rules for automatic text processing including regular expression rules and programming languages — Java, Python, for processing the predefined document data models. Neves and Ševa (2019) provide a comprehensive review of manual annotation tools. They defined a set of evaluation criteria for what makes an annotation tool useful.

NB: In the last version, we will include more related works.

3. Datasets for Bulgarian

In this section, we present the main datasets used in our work as a basis for fine-tuning of the different NLP models.

- *BulTreeBank*: the original constituent variant (Simov et al., 2002) contained 256 000 tokens. We extended it to the morphological level with 2 445 407 tokens.
- *The Bulgarian part of the Balto-Slavic Corpus*: For the shared task (Piskorski et al., 2021) the corpus has been annotated with Named Entities (NEs) adhering to pre-defined guidelines, and also the NEs were linked cross-lingually through the lemmas. On this available corpus, we mapped the NEs to Wikipedia URLs and processed the texts on the morphological level. In addition, we lemmatized all the tokens. The total amount of tokens is 400 000.
- *The Bulgarian Event Corpus (BEC)*: BEC (Simov et al., 2025) comprises 227 documents with 291 196 tokens of texts that are in the area of history, biographies, ethnography, etc. The corpus has been annotated on two levels: NEs and events with triggers and roles. In addition, the co-reference chains were marked as well.

4. LLM-based NLP Models for Bulgarian

Here we describe the construction of the Bulgarian NLP models based on LLMs. Our understanding is that such models have to be relatively small in order for them to be easily re-trained when necessary. The resulting models are also much easier to deploy, as they have modest hardware requirements. We follow the well established paradigm of pre-training language models on a large collection of text and later fine-tuning them for specific tasks on smaller sized supervised datasets.

4.1. Pre-trained models

In this section, we introduce our pre-training setup and the produced models. For the pre-training, we use unsupervised corpora of 29B Bulgarian tokens. We sourced them from 3 openly available datasets: CulturaX (Nguyen et al., 2024), Macocu (Bañón et al., 2023), and HPLT (de Gibert et al., 2024), as well as from some other smaller datasets that we gathered manually, such as News sites (processed manually (5 BW)), Wikipedia, PHD theses, research papers, and other. All data is publicly available. The dataset was deduplicated on document level with approximated Jaccard Similarity using MinHashLSH from datasketch package.[5] We used a threshold of 0.7. We have pre-trained various types of models — *encoder only, encoder-decoder, decoder only*. They are suitable for different NLP tasks. In the work reported here, we present only the first two types of models. The pre-trained models are publicly available on HuggingFace[6].

[4] `https://spacy.io/`

[5] `https://ekzhu.com/datasketch/`

[6] `https://huggingface.co/AIaLT-IICT`.

4.1.1. Encoder models

We consider the BERT (Devlin et al., 2018) architecture the most popular encoder architecture. We developed BERT models of different sizes. All use a vocabulary of 50 176 tokens processed with the WordPiece tokenization tool — Song et al. (2021a). Recently, a modern and more efficient version of the BERT architecture was proposed in Warner et al. (2024). Following the ModernBERT architecture, we also pre-trained a number of models with various sizes. The basic characteristics of the models are presented in Table 4.1.1. It is well known that such types of model are widely used for classification tasks as well as for producing text embeddings. In our work, we use them for tasks like POS Tagging and Token classification. We also use them for Universal Dependencies oriented parsing.

Model	casing	dim	layers	pars
bert-base	both	768	12	124M
bert-large	both	1024	24	355M
bert-XL	no	1024	48	657M
mbert-base	no	768	22	149M
mbert-large	no	1024	28	395M

Table 1: BERT and ModernBERT based pre-trained encoder models. The first two models are available in both — cased and uncased variants — while the others are uncased. The columns denote: name of the model, casing type, token embedding size (**dim**), number of the transformer layers (**layers**), and number of parameters (**pars**).

The training procedure used the Masked Language modeling objective (as proposed in the BERT paper) with 20% noise tokens. We trained for 3 epochs with a learning rate of 1×10^{-4}, batch size of 256 chunks, and context-length of 512 tokens. For applications requiring longer context we extended the model context-length by a subsequent training on longer sequences (from 4096 to 8192 tokens).

4.1.2. Encoder-Decoder models

As mentioned above, encoder models excel especially in classification tasks. Unfortunately, not all tasks can be modeled in this way. An example for this is the lemmatization task: Each word form must be mapped to its lemma, with the possible lemmas being too many to be efficiently modeled as classes. Thus, a generative model capable of making a good representation of the input text seems to be more suitable. Thus, we considered the architecture of T5 (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer) (Raffel et al., 2019a), which combines an encoder and a decoder. We trained different sized T5 models, presented in Table 2.

The training procedure used a span denoising objective as proposed in the original paper with a noise density of 25% and a mean noise span of 3 tokens. We used the same hyperparameters for the training as in the encoder models. We also used the SentencePiece BPE tokenizer (Kudo and Richardson, 2018) since it proved to be better suited for text generation.

Furthermore, some sequence-to-sequence applications benefit from lower level tokenization (for example, spellchecking). We trained T5 models on character level tokenization. The training objective was again span denoising with a mean noise span of 7 characters, since 7 letters is the average length of the words in the corpus. Since the character level tokenization segments the training corpus in more tokens, we used a version of the pre-training corpus filtered with a more aggressive deduplication to 10B words.

Arch	tokenization	dim	layers	pars
T5	subword	1024	12-12	403M
T5	character	1024	16-16	470M
T5	subword	1536	16-16	1.1B

Table 2: T5-based pre-trained models. The first and third models used uncased subword level tokenization, while the second model employed a character level tokenization. The first column describes the types of the architecture, the second column shows the tokenization level, the third column (**dim**) presents the token embedding size, the fourth (**layers**) — the number of the transformer layers for each model, and the last column (**pars**) contains the number of parameters for each model.

4.2. Task specific fine-tuning

After pre-training, the models can be tuned for various specific tasks. This subsection describes our efforts with respect to some of them.

4.2.1. Encoder based classification

Many tasks can be modeled as classification — assigning a label from a fixed number of classes to each token in the text or the whole text. We fine-tuned our pre-trained encoder models on different tasks, as well as the publicly available at HuggingFace *bert-web-bg*[7] and *bert-web-bg-cased*[8] which is one of the few Bulgarian Transformer models available. Below are more details about the tasks. Table 3 gives test result reports, comparing the pre-trained models.

[7] https://huggingface.co/usmiva/bert-web-bg

[8] https://huggingface.co/usmiva/bert-web-bg-cased

Pre-trained Model	Tok+SS (F1)	NER (F1)	XPOS (Acc)	UAS (Acc)	Text Err (F1)	WSD (F1)
bert-web-bg	0.9669	0.9198	0.9787	0.9033	0.8511	0.7756
bert-web-bg-cased	0.9849	0.9005	0.9810	0.9192	0.8866	0.7507
bert-base	0.9877	0.9289	0.9865	0.9564	0.9214	0.7970
bert-base-cased	0.9927	**0.9497**	0.9858	0.9488	0.9131	0.7949
bert-large	0.9895	0.9378	**0.9878**	0.9412	0.9225	0.8155
bert-large-cased	**0.9940**	**0.9497**	0.9877	**0.9571**	0.9245	0.8136
modernbert-base	0.9902	0.9174	0.9864	0.9523	**0.9462**	0.8119
modernbert-large	0.9895	0.9206	0.9875	0.9444	0.9423	**0.8244**

Table 3: Test results across encoder fine-tuning tasks. Test results on the classification tasks. Rows show results for the following tasks: Tokenization and Sentence Segmentation, Named Entity Recognition, Part of speech and morphological tagging (**XPOS**), Unlabeled syntax head prediction (Unlabeled attachment score - **UAS**), Text error tagging and Word Sense Disambiguation. Macro F1 is reported for Tokenization+Sentence Segmentation, NER, Text Error tagging, and WSD. Accuracy is reported for the remaining tasks.

Tokenization and Sentence segmentation is a task that classifies the beginning of the words and the sentences. In Bulgarian sentence and word boundaries are ambiguous, and thus a model is needed. [9] Often tokenization and sentence splitting is the first objective to be completed in the NLP pipe, since the other annotations usually work on sentence level. We model the task by assigning to each token a class from B-SEN, B-WOR, and I-WOR. These correspond to the beginning of a first word of a sentence, the beginning of a non-first word of sentence, and the continuation of a word. Any supervised tokenized dataset with consequent sentences can be used as training data. Here, the *Bulgarian Event Corpus* has been considered. We divided the corpus into 766 train, 96 validation, and 96 test chunks of 512 tokens. We train for 20 epochs with a linearly decaying learning rate starting from 1×10^{-4} and batch size of 8*32. We pick the best model by validation loss (usually it is around 17 epoch with negligible increase of the val loss after).

Named Entity Recognition is the classical task of assigning labels to word spans that are classified as names. We use *The Bulgarian part of the Balto-Slavic Corpus*. It contains 11 classes with 5 NER categories: person, location, organization, event, and pronoun. We use the provided train and test splits. We partition a validation split from the train data. That amounts to 6 032 train, 671 validation and 2 180 test sentences. We train for 5 epochs with a linearly decaying learning rate starting from 2×10^{-5} and batch size of 8*16. We pick the best model by validation loss (usually it is around 17 epoch with negligible increase of the val loss after).

Part of speech tagging is the task of assigning a label corresponding to the grammatical features of the word, grouped in POS tags. For training, we use our extended version of the *BulTreeBank* corpus. We split the data into 134 331 train, 7 071 validation, and 15 712 test sentences. We train for 3 epochs with linearly decaying learning rate from 5×10^{-5} and batch size of 8*16 sentences. Our models score around 0.99 accuracy, which to our knowledge is the SOTA for the Bulgarian on this task.

Text error tagging is the task of finding spelling, grammatical, and punctuation errors in a given text. We developed data for this task by noising a text that we consider correct with various rules corresponding to common mistakes, including casing errors. In this way, we automatically created training dataset of 11 394 612 tokens in 188 908 text chunks. We model the task as binary token classification and train for 1 epoch with linearly decaying learning rate from 2×10^{-05} and batch size of 8*16 text chunks. During training of the uncased models, we ignore the casing error labels. For testing, we developed a small set of error correction exercises from publicly available Bulgarian exams consisting of 3 101 tokens in 46 texts. On this set, both cased and uncased models are tested. The Macro F1 test scores are presented in Table 3. Manual evaluation showed that the uncased models spot difficult punctuation errors better than the cased models but are naturally unable to assess casing errors. Thus, we got the best results by combining the error predictions from the best uncased model and the best cased model, trained only on casing errors.

4.2.2. UD syntax parsing

Syntax parsing refers to the tree analysis of a given sentence. As mentioned above, we modeled the task in two steps: (1) Annotating the head (parent node) of every word in the sentence; (2) Classifying the syntax relation between each token-head pair. The annotation of the head is done by adding an extra layer on top of the encoder that calculates scores between each pair of tokens similarly to the scoring part in the attention heads. Then after

9

calculating the scores, a maximum spanning tree algorithm (McDonald, 2006) is applied to produce the syntax tree. The classification of the syntactic relation was performed by concatenating the vectors of the tokens and pairs predicted by the first step and linearly projecting them to the set of classes. We used *BulTreeBank*'s 8 907 training, 1 115 validation, and 1 116 testing sentences. We train for 5 epochs with linearly decaying learning rate from 5×10^{-5} and batch size of 8*48. The accuracy of the head prediction is referred to as UAS - Unlabeled Attachment Score, while the accuracy of both head prediction and class labeling is referred to as LAS - Labeled Attachment Score. Our best result for the UAS is 0.9571 and the LAS is 0.9330. We report the UAS scores across the pre-trained models in Table 3.

4.2.3. Named Entity Linking

Named Entity Linking is the task of disambiguating names in a text to a knowledge base. In our case, we labeled the names in the text with their Wikipedia article URL. For training, validation, and testing, we used datasets with Wikipedia URL annotation of named entities: *The Bulgarian part of the Balto-Slavic Corpus*, the *BulTreeBank*, and sentences from *The Bulgarian Event Corpus (BEC)*. The resulting set consisted of 53 829 contexts and 7 065 unique named entity urls. We modeled the task as a semantic search. We enclose the entity that needs to be linked with special tokens and treat the whole context as a query. The corresponding Wikipedia article acts as a document. We fine-tuned an encoder for producing embeddings using the Sentence Transformers library (Reimers and Gurevych, 2019). The model was optimized to make the embeddings of the context and the Wikipedia article closer than the random (noise contrasting objective). We split the data into 43 063 training, 5 384 validation, and 5 382 test pairs of entity context and wikipedia article. We train the sentence transformer for 3 epochs with a peak learning rate of 1×10^{-5} and a batch size of 8*16 pairs.
After training, the named entity linking is performed by finding the Wikipedia article with the closest embedding. We perform experiments on the entities with contexts from the test set by disambiguating them to the whole Bulgarian Wikipedia consisting of 151 708 entities. We report Recall at positions: 1, 3, 5, and 10 - 0.83, 0.93, 0.95, and 0.96. The system may benefit from having a cross-encoder model to perform re-ranking, but we leave this for future work.

4.2.4. Word Sense Disambiguation

Word Sense Disambiguation is the task of linking semantically ambiguous words in a text to a thesaurus like WordNet. We modeled the task in two steps: (1) We use a lexicon and a lemmatizer model to find the candidate senses of the word; (2) We use a cross-encoder model to classify the similarity between the context with the word enclosed in special tokens and the definition of the candidate sense from the Bulgarian BTB-WordNet (Simov and Osenova, 2023).
To train the model, we used the same WordNet to create positive and negative data pairs. The BTB-WordNet has example usage for every synset. More specifically, we considered all words that are ambiguous (linked to more than one sense) and created example-definition pairs. The positive pairs were created directly by pairing the sense definition and one of its examples, while the negative ones were taken from the examples of other senses for the same word that is not semantically related. Thus, we created positive and negative pairs. We split the set into 85 254 train, 3 698 validation, and 4 006 test pairs, while making sure that there are no pairs with the same sense in the different partitions, to avoid leakage. We also enhanced the definitions with other words that are linked to the same sense, which improved the accuracy. The models are trained for binary classification in the classes: similar and different. We train for 3 epochs with linearly decaying learning rate from 2×10^{-5} and batch size of 8*32. We report the F1 score over the test set pairs in Table 3.

It resulted in a F1 score of of 82.44% in finding the correct sense of ambiguous words.

4.2.5. Encoder-Decoder conditional generation

As mentioned above, not every task can be modeled as a token classification. For such tasks, more general free form text generation is preferable. Thus, we fine-tuned our T5 models.

Lemmatization is the task of converting a word to its base form (lemma). We modeled it by passing a sentence with words separated by special tokens and training a model to generate the corresponding lemmas in the correct positions. In this way, the whole sentence was lemmatized in a single generation. We fine-tuned subword level T5 model on a train split of our extended version of the *BulTreeBank* corpus and achieved a test accuracy score of 0.9946.

We fine-tuned T5 model for **Spelling correction** of a text with spelling and punctuation errors. We used the same dataset as the text error tagging task described in Subsection 4.2.1. The model achieved a BLEU score of 0.9936 on a test set.

Translation from old spelling is another sequence-to-sequence task. In Bulgaria in the 19th and 20th centuries, there were different spelling systems in comparison to the only one that is used

Model	Tokenization	BLEU (word level)
Lexicon substitution (baseline)		0.8769
T5	Character	**0.9532**
T5	Subword	0.9201
EuroLLM 1.7B	Subword	0.9421

Table 4: BLEU Scores of the tested fine-tuned models for the translation from old spelling task.

in modern Bulgarian. The differences are confined to variations in letter usage. For training, validation and testing data, we used parallel sentences paired from different versions of novels and short stories published in the period of 1910-1944 for which we found also electronic versions published after 1945. We used some heuristics in the alignment, with manual checks on the result for nearly half of the texts. We fine-tuned the models on the dataset for 10 epochs with a peak learning rate of 1×10^{-4}. The best checkpoint according to the validation set is selected. We found that the T5 model that tokenized the text at the character level is better suited than the one that used subword tokenization. The best model achieved a BLEU score of 0.9532 on a test set. We fine-tuned EuroLLM-1.7B (Martins et al., 2025) on the same task and observed that the results are close to the T5 character level results. (Table 4) We hypothesize that the reason is that the tokenizer of EuroLLM, being multilingual, tokenizes old Bulgarian words in smaller subwords, maybe even characters, similarly to the T5 character level.

4.2.6. Event Extraction

Event Extraction is the task of finding and extracting structured information from a given text in the form of events and participants in them. We modeled the task as a sequence-to-sequence generation that takes a sentence as input and produces an output in a JSON format which consists of a list of the events present in the sentence. Each event is represented as an object and has a type, a span - the position in the sentence, and a list of roles - each with a type and a span. We used the *Bulgarian Event Corpus* dataset, splitting it into 13 233 training, 1 471 validation, and 1 634 test sentences. We evaluate the model on the test sentences and report an F1 score of 0.77 in retrieval of event spans and an accuracy of 0.86 in prediction of the type of true positive events.

5. Visualization of the annotations

After accumulating all the above mentioned models for the various tasks, we began developing a user interface to observe the different annotations either in isolation or in combination over documents of interest. For that reason, in this section we present a basic user interface implemented as an HTML editor. This allows the user to observe the documents together with their formatting – different headings, sections, footnotes, bold, italics, etc. Additionally, the HTML mark-up has been extended with information about the available annotation. Here, we consider two scenarios of usage, depending on the user needs and the available corpora. If users want to annotate their own corpora, then it would be useful for them to have the possibility to annotate the corpora with selected types of annotation. However, if the corpora have been shared by many users with different needs, then we imagine that all types of annotation have to be performed. Apart from the NLP tasks discussed above in Sect. 4 we also consider some obvious extensions to the set of NLP tasks to be supported by the set of models. Such a task is for example co-reference resolution.

Having in mind the user requirements, we decided that the best way to encode the annotation information is to represent it on a token level. In this way, the interaction of the HTML mark-up as well as the editing operation with the annotation is maximally reduced. Thus, our goal here is the projection of the corresponding annotation to token level annotation. As is well known, this is possible for many types of annotation. For example, the Universal Dependency (UD) annotations are represented in such a way. In UD we have three main types of annotation: (1) “category” annotation, where the information is represented as a single value, which we interpreted as a category, independently from the fact that these values can be lemmas, tags (representing a bundle of values, positionally), or a list of feature:value pairs; (2) “tree” annotation, where the annotation is represented by arcs between the tokens which form tree structures over the tokens in the text. In this way, UD encodes the dependency structures; (3) “dep-graph” annotation, where the annotation is represented as multiple arcs from a given token to other tokens in the text. UD encodes in a similar way the Enhanced dependency graph, where a list of head-deprel pairs is attached to the tokens. Each head-deprel consists of two elements: the first element is the position of the token in the current sentence, and the second element is the dependency relation. In our work, we would like to extend this annotation to allow the encoding of in-coming and out-leaving arcs, bidirectional arcs, and having different labels.

In order to present more complex graphs than the enhanced dependency graph, we extended the notion of head-deprel to *graph-edge-rel*. Each graph-edge-rel consists of the following elements: *target node* — the position of the token to which the edge

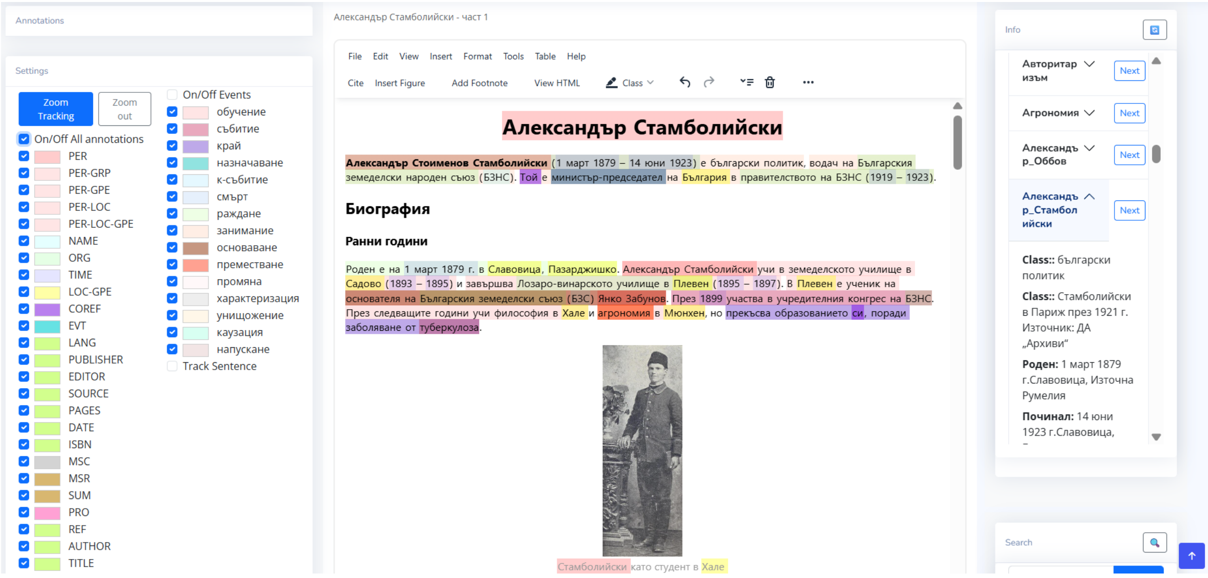


Figure 1: This screenshot depicts the editor main screen containing an annotated document. The annotation scheme has three levels of annotation: Named entity; Named entity linking; Event annotation. Named entity and Named entity linking levels are of type “category”. The Event is represented as a graph between tokens that determines the event in the text. The left part of the screen presents the different categories of Named entities or events. The right part of the screen presents information related to the Named entity links.

is connected; *edge type* — one of the values: *in* (in-coming arc), *out* (out-leaving arc), and *bi* — a bidirectional arc; *label* — the label from a set of labels assigned to the edge. Thus, we want to have an annotation scheme with optional annotations. For example, in the case of the UD CoNLL format we could require having cases in which only XPOS tags are presented.

In order to meet these requirements, for each document, we define an annotation scheme consisting of a list of declarations of the annotation scheme elements. Similarly to CoNLL format, the first element of the annotation scheme is the indexing of the tokens within the text. The token number is global over the whole text, allowing annotations over more than one sentence. Each element declaration in a scheme definition has the form: *type of annotation* — one of the values: “category”, “tree”, “dep-graph”, and “graph”, as defined above; Secondary type of the annotation — one of the values: “single”, “list-of-value”; Set of values — what the values are. The values could be some of the types: “number”, “string”, “tag”

Fig. 1 presents an example of annotated document on three layers of annotation: Named entity layer, Named entity linking layer, and Event layer.

6. Conclusion and Future Work

In this paper, we presented a number of LLMs, fine-tuned to basic NLP tasks for Bulgarian. In our practice, most of them proved to be useful in supporting research in the area of Humanities. To provide the annotations to the researchers, we implemented an approach for encoding the annotations in HTML documents.

We think that all of these tasks are useful for researchers within SS&H. Here, we describe some of the possible scenarios. (1) POS and lemmatization services provide important information for searching in large corpora.[10] Both tasks, mentioned above, are necessary for the preparation of corpora loaded into NoSketch Engine. For most of the users, this information is not necessary. However for linguists, lexicographers, language teachers, and others, the information from the annotation is necessary. (2) WSD service is important for linguists, lexicographers, but also for humanities specialists who are interested in the annotation of their corpora with specific terminological lexicons. (3) NER and NEL services allow for the diverse representation of one or many documents (see the user interface above). (4) UD Parsing and Event services interact for extensions of event extraction within the current event inventory, but also for extensions to new types of events. All of these services are easier to maintain, to extend to new areas, etc. They also significantly contribute to the explainability, and are much cheaper than LLMs with billions

[10] See the massive list of corpora uploaded in NoSketch Engine from CLASSLA: Knowledge centre for South Slavic languages — https://www.clarin.si/ske.

of parameters.

Needless to say, LLMs with billions or even trillions of parameters can also be used. But they are not ready off-the-shelf as many of the users expect. We performed some experiments with ChatGPT services to implement these tasks using the prompting approach. For tasks like POS, Lemmatization, UD Parcing, the results were worse than ours. For the WSD task, both approaches perform similarly. We did not conduct experiments based on fine-tuning of LLMs with billions of parameters. We imagine that such fine-tuned models will be comparable.

In our future work, we plan to extend the set of models to new tasks. We will work on the creation of RAG systems in SS&H in which we consider to use encoder models for vectorization of the document database, and small decoder models for the generation of the answers depending on preselected snippets. Thus, our strategy is to keep the models small but powerful enough for the tasks in hand as much as possible, and to gradually extend the models with respect to architecture and size towards related new tasks.

7. Bibliographical References

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. 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 3509–3514. ACL.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Viet Dac Lai. 2022. Event extraction: A survey.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Comput. Sci.*, 255(C):53–62.

Ryan McDonald. 2006. *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. Ph.D. thesis.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, page 216–220, USA. Association for Computational Linguistics.

Mariana Neves and Jurica Ševa. 2019. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163.

Saziye Betül Özates, Arzucan Özgür, Tunga Güngör, and Balkiz Öztürk. 2020. A hybrid approach to dependency parsing: Combining rules and morphology with deep learning. *CoRR*, abs/2002.10116.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019a. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019b. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Yikang Shen, Shawn Tan, Alessandro Sordoni, Siva Reddy, and Aaron Courville. 2021. Explicitly modeling syntax in language models with incremental parsing and a dynamic oracle. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1660–1672, Online. Association for Computational Linguistics.

Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2024. Generative approaches to event extraction: Survey and outlook. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 73–86, Miami, Florida, USA. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021a. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021b. Improved Word Sense Disambiguation with Enhanced Sense Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4311–4320. ACL.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions.

8. Language Resource References

Marta Bañón, Mălina Chichirău, Miquel Esplà-Gomis, Mikel Forcada, Aarón Galiano-Jiménez, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vit Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 505–506, Tampere, Finland. European Association for Machine Translation.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Kiril Simov and Petya Osenova. 2023. Recent Developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 220–227, University of the Basque Country, Donostia -

San Sebastian, Basque Country. Global Wordnet Association.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Kiril Simov, Nikolay Paev, Petya Osenova, and Stefan Marinov. 2025. Bulgarian event extraction with llms. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1163–1171, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Multimodal LLM-Based Nutrition Label for Analyzing Social Media Feed Exposure

Tim Gollub, Armin Heidari, Cem Ertürkan, Benno Stein
Bauhaus-Universität Weimar, Germany
{tim.gollub, armin.heidari, cem.ertuerkan, benno.stein}@uni-weimar.de

Abstract

Algorithmically curated social media feeds shape political exposure, commercial influence, and cultural consumption, yet they remain difficult to study systematically due to limited data access and opaque recommendation mechanisms. We present a research-oriented framework that operationalizes feed-level exposure analysis using a browser extension combined with a server-side multimodal large language model (LLM). The system logs visible posts and their view time, performs zero-shot multimodal classification, and aggregates results into a customizable nutrition label summarizing exposure across analytical categories. It further supports retrieval-grounded conversational querying, dataset export and sharing, and human validation of LLM classifications. Designed as a methodological instrument for Social Sciences and Humanities, the framework enables both observational analysis and experimental research on transparency interventions, while critically examining epistemic, methodological, and ethical implications of LLM-based exposure analysis.

Keywords: Large Language Models; Multimodal Analysis; Social Media Research; Algorithmic Transparency; Exposure Measurement; Digital Humanities; Computational Social Science; Information Nutrition Label

1. Introduction

Algorithmically curated feeds increasingly mediate how individuals encounter political information, commercial messaging, and cultural content. For researchers in the Social Sciences and Humanities (SSH), understanding feed-level exposure is essential for studying algorithmic amplification, public discourse formation, and systemic risks.

However, empirical research on personalized feeds faces structural barriers. Platform APIs rarely provide access to individualized exposure streams, and recommender systems remain opaque. Regulatory developments such as the European Union's Digital Services Act (DSA) highlight this tension: Article 27 mandates transparency regarding recommender systems (European Union, 2022a), while Article 40 establishes data access provisions for vetted researchers studying systemic risks (European Union, 2022b). Yet practical methodological tools for operationalizing exposure-level analysis remain scarce.

We present a research-oriented framework that combines a browser extension with a multimodal LLM to log and analyze feed exposure. The system records visible posts and their view time, performs zero-shot classification across user-defined analytical categories, aggregates results into an exposure-oriented "nutrition label" (see Figure 1), and enables retrieval-grounded conversational querying. It further supports dataset export/import and human validation of automated classifications.

While inspired by the metaphor of consumer-facing nutrition labels, we primarily position the system as a methodological instrument for SSH research and as a testbed for studying transparency interventions.

2. Related Work

Our work builds on prior research tools for social media monitoring, most notably Zeeschuimer, a browser extension that allows users to collect data from social media feeds for research purposes (Peeters, 2025). While Zeeschuimer supports feed logging, it does not record post visibility duration and does not analyze or summarize content.

Inspired by food nutrition labels, several researchers have proposed "information nutrition labels" to communicate properties such as credibility, bias, or sourcing (Fuhr et al., 2017; Gollub et al., 2018; Kevin et al., 2018; Willinsky and Pimentel, 2024). These approaches primarily focus on item-level or publisher-level information. In contrast, our work emphasizes personalized, feed-level aggregation weighted by view time.

Most closely related is the recommender system label proposed by Belli and Wisniak (2023), which aims to reveal parameters influencing post amplification. Whereas their proposal focuses on transparency for content creators, our framework supports content consumers and researchers by aggregating exposure across customizable analytical categories.

More broadly, our work connects to emerging SSH uses of LLMs for qualitative coding, discourse analysis, and multimodal interpretation, while explicitly addressing reproducibility and epistemic concerns.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 198–204
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

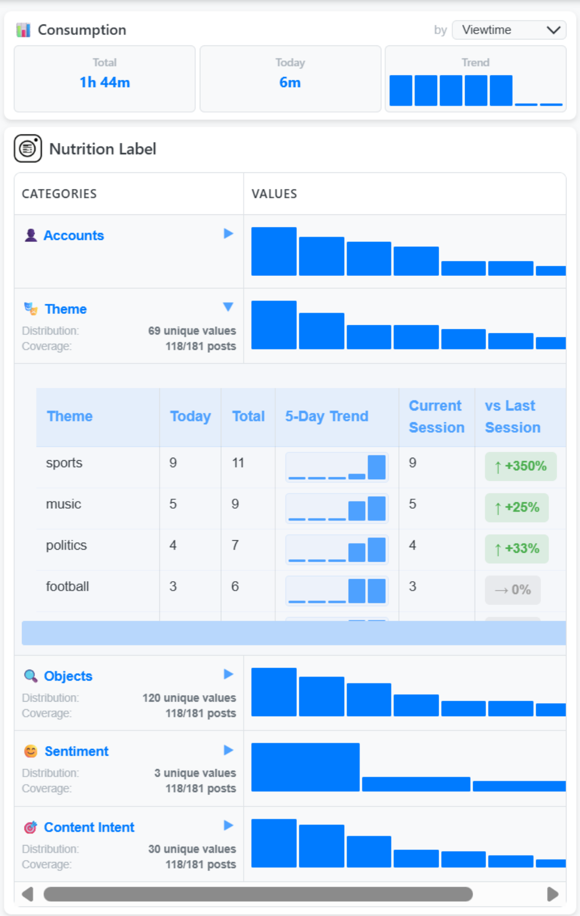


Figure 1: Feed-level nutrition label. Aggregate statistics are computed per category and can be displayed by post count or view time (see select box at the top right) to approximate exposure. Selecting a category value enables drill-down to the corresponding posts in the local database.

3. Research Motivation

Studying exposure requires moving beyond isolated content analysis toward temporally weighted feed composition. Researchers may ask which themes dominate exposure over time, how political and commercial content are interwoven, or what emotional tones characterize curated streams. Such questions require operationalizing exposure not merely as post frequency but as time-weighted visibility, since feed consumption typically involves rapid scrolling through numerous posts with only occasional sustained attention to particular items. Capturing view time therefore provides a more meaningful approximation of exposure intensity than simple post counts alone.

Beyond observational analysis, the framework enables intervention research. The nutrition label and conversational interface can serve as transparency treatments, allowing researchers to investigate how aggregated exposure information influences scrolling behavior, engagement patterns, or perceptions of algorithmic bias.

4. System Architecture

The system consists of two components: (1) a client-side browser extension responsible for feed monitoring, storage, retrieval, and user interaction; and (2) a server-side multimodal LLM that performs semantic analysis and reasoning tasks. The architecture is designed to keep exposure datasets locally under researcher control while delegating computationally intensive multimodal inference to dedicated hardware.

In the current implementation, we deploy the open-weight multimodal model `Qwen2.5-VL-7B-Instruct` (Alibaba Cloud, 2024). The backend processes requests transiently and does not persist user data. This setup enables researchers to run the system on their own infrastructure without relying on commercial APIs.

4.1. Client-Side Data Capture

The browser extension currently operates within the Instagram web interface, with support for additional platforms such as TikTok currently under development. A content script detects which post is visible in the viewport using intersection observers and timestamp tracking. For each post, the system initially records the following captured fields:

- Post identifier
- Caption text
- Media reference (image or video URL)
- Metadata (e.g. timestamp, interaction counts such as likes and comments)
- User interaction signals (e.g., whether the user liked, saved, or commented on the post)
- View time (milliseconds visible)

All captured data are stored locally using IndexedDB. Thousands of posts can be stored without performance degradation, with local browser storage constituting the primary scalability constraint. The database supports keyword search, field-specific filtering, Boolean logic (AND, OR, NOT), and numerical comparisons ($<$, $>$, $=$) for metadata fields.

The database can be explored in an extra tab provided by the browser extension. Besides a classical table view, a gallery and feed view are available for an image/video focused exploratory analysis (Figure 2).

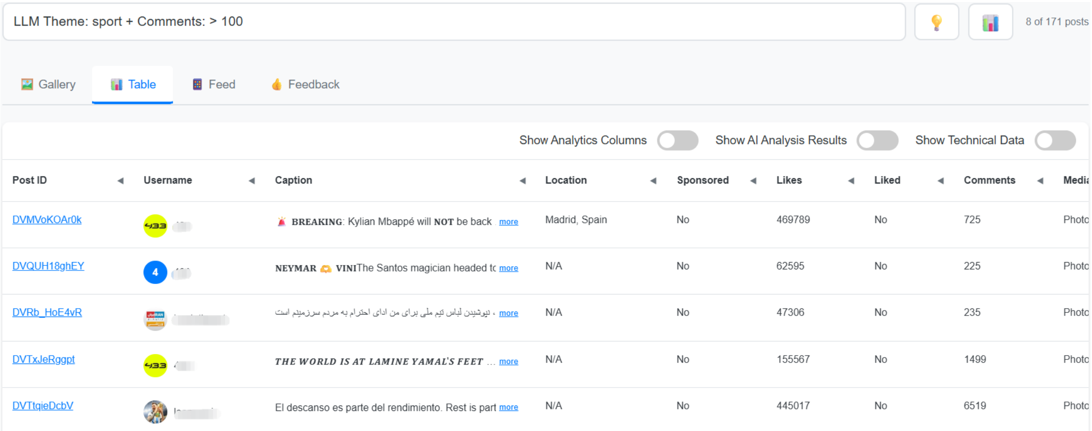


Figure 2: Database tab (table view). The extension stores posts locally and provides keyword search and filter controls (including nutrition label categories and post metadata) to retrieve subsets of the recorded feed for analysis.

4.2. Multimodal Description Generation

After initial storage, the system generates a textual description for the image or video (based on a video frame) of each post using the multimodal LLM. These descriptions are stored alongside the captured data as derived representations. Generating descriptions by default ensures that each post has a unified textual representation, facilitating consistent downstream classification and conversational reasoning.

4.3. Zero-Shot Category Classification

Classification is performed via structured zero-shot prompting. For each active nutrition label category, the prompt includes:

- Category name (required)
- Optional description
- Optional set of candidate values
- Whether the value set is open or closed

The LLM is guided to return valid JSON output conforming to a predefined schema. This design allows researchers to define new analytical categories dynamically without retraining the model. The resulting category assignments are appended to the corresponding post records in the local database.

Category definitions are managed through the extension settings interface, where researchers can specify names, optional descriptions, and candidate values (Figure 3).

For transparency at the item level, each post includes an inline overlay displaying the predicted category assignments (Figure 4).

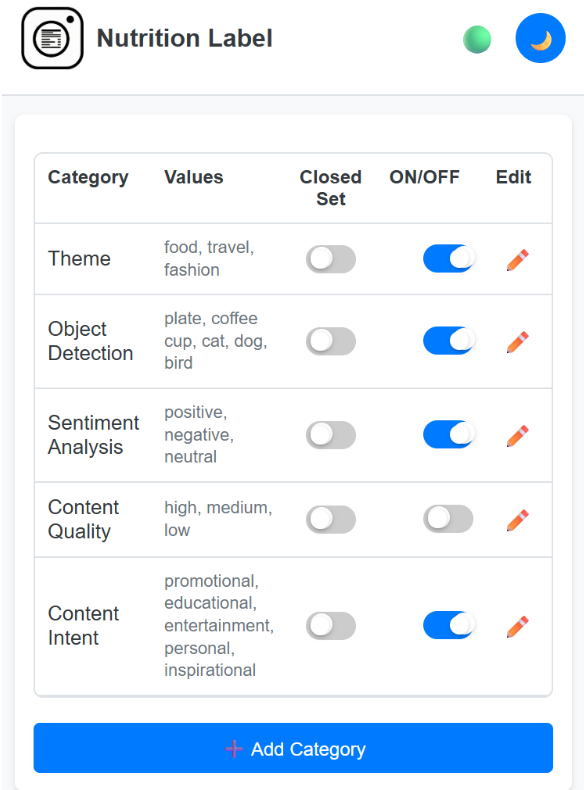


Figure 3: Extension settings. Researchers define nutrition label categories by specifying a name and optionally a description and candidate values. Categories are applied via zero-shot prompting without retraining.

4.4. Feed-Level Aggregation

Category assignments are aggregated across the recorded feed to generate a nutrition label summarizing exposure patterns. Aggregation can be computed either by post count or by cumulative

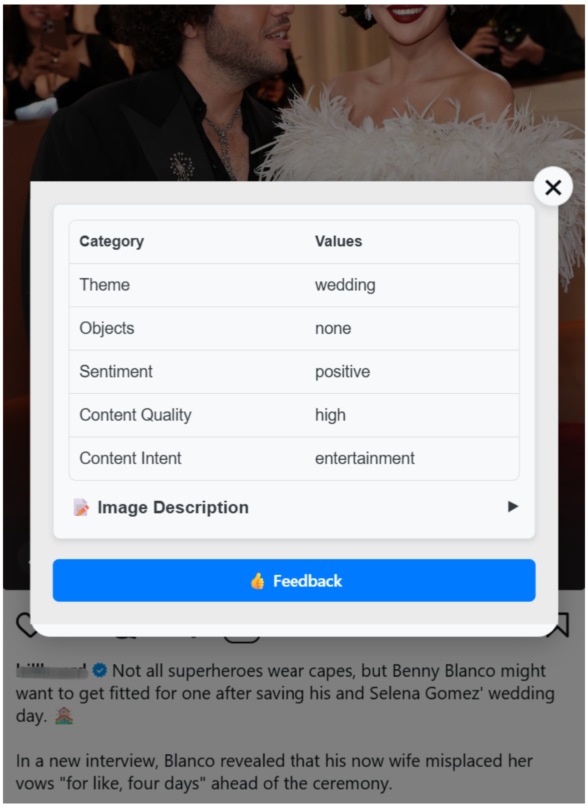


Figure 4: Post-level nutrition overlay. The LLM's category assignments for the selected post are displayed inline within the social media interface.

view time, thereby approximating exposure intensity under rapid scrolling conditions.

The aggregated results are presented in a feed-level nutrition label that supports drill-down interaction (see again Figure 1). Selecting a category value filters the local dataset and opens the corresponding subset in the database view.

4.5. Retrieval-Grounded Conversational Querying

The "Talk to your Feed" interface implements a two-step reasoning pipeline. Given a natural language question, the LLM first generates a structured query expressed in the system's custom search syntax. The extension executes this query locally against the IndexedDB dataset. Retrieved posts, including caption text, generated image/video descriptions, metadata, interaction signals, and category assignments, are then sent back to the LLM. The LLM produces an answer grounded explicitly in this retrieved content.

This separation between query generation and answer generation constrains reasoning to the recorded exposure dataset and reduces ungrounded outputs. The conversational interface exposes this retrieval-grounded reasoning process directly to researchers (Figure 5).

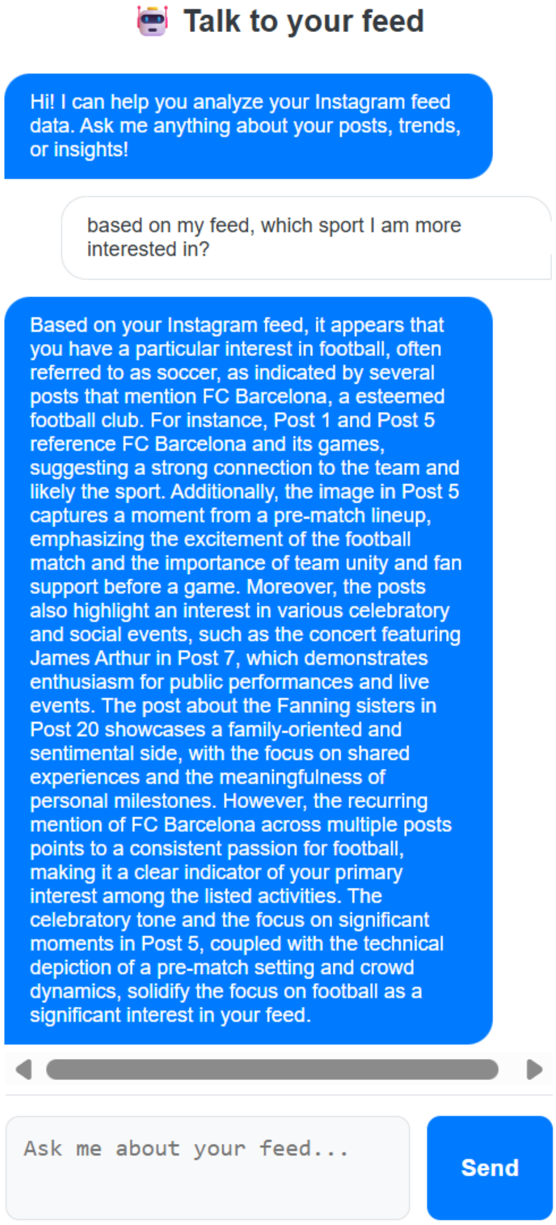


Figure 5: Retrieval-grounded conversational interface. The LLM first generates a structured query over the local database. Retrieved posts are then used to generate a grounded response.

4.6. Human-in-the-Loop Validation

To assess classification reliability, the system provides a review interface allowing researchers to navigate posts sequentially using keyboard controls and evaluate predicted category values. Validation feedback (correct/incorrect) is stored as additional fields associated with each post record.

Annotations can be exported to compute classification accuracy and inter-annotator agreement, enabling mixed-method research designs that combine automated coding with human validation (Figure 6).

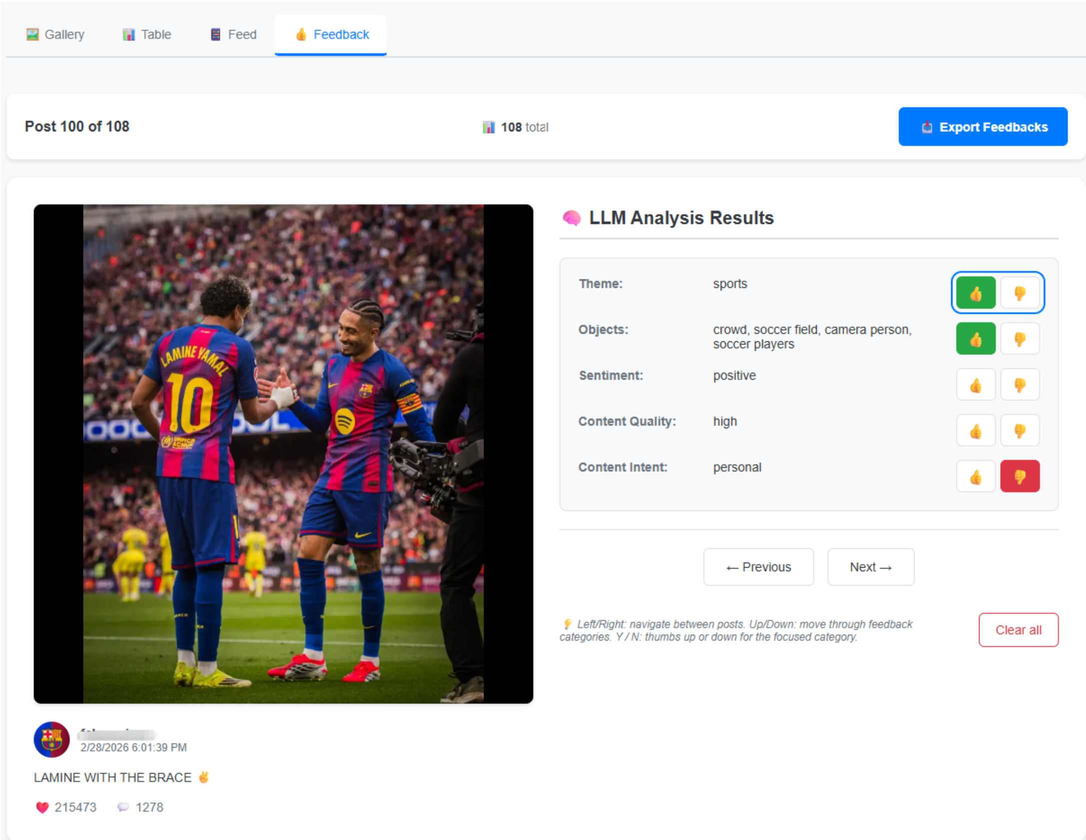


Figure 6: Review interface for validating LLM classifications. Researchers assess predicted category assignments and export annotations to compute accuracy and inter-annotator agreement.

4.7. Data Portability and Reproducibility

Exposure datasets can be exported and imported using a custom JSON format. Exported data include captured fields (metadata, interaction signals, view time), derived representations (image/video descriptions and category assignments), and human validation annotations. This functionality enables collaborative workflows in which multiple researchers analyze identical exposure datasets independently.

By combining local storage, open-weight LLM deployment, structured prompting, retrieval-grounded reasoning, and exportable datasets, the architecture supports flexible yet methodologically transparent exposure analysis.

5. Methodological, Epistemic, and Ethical Considerations

Using multimodal LLMs as analytical instruments for feed-level exposure analysis introduces methodological, epistemic, and ethical challenges that require careful consideration.

Zero-Shot Operationalization. Zero-shot classification enables researchers to define analytical categories dynamically without retraining. While this flexibility lowers barriers for exploratory research, it introduces prompt sensitivity and model-version dependency. Category definitions should therefore be understood as operational constructs rather than objective ground truth labels. Reproducibility depends on documenting prompts, model versions, and hardware configurations.

Multimodal Interpretation and Derived Representations. The system generates textual descriptions of images and video frames to create a unified representation for reasoning and retrieval. While this facilitates consistent analysis, multimodal LLMs may hallucinate details or overgeneralize visual cues. Researchers should treat generated descriptions and classifications as interpretive artifacts rather than direct observations.

Operationalizing Exposure Through View Time. Weighting aggregation by view time approximates attention allocation in scrolling environments. However, visibility duration does not fully capture cognitive engagement, affective response, or background exposure. The metric should therefore be interpreted as a proxy for temporal prominence rather than a direct measure of psychological impact.

Normativity and Bias. Certain analytical categories, such as "Content Quality" or "Intent", embed normative assumptions. LLM outputs may reflect biases present in training data, including cultural or linguistic biases. Human validation and inter-annotator agreement mechanisms are therefore essential components of responsible use.

Privacy and Data Governance. All exposure data are stored locally within the browser using IndexedDB. The backend LLM processes requests transiently and does not persist images or text. Images and video frames are transmitted for inference but not retained. The architecture is designed for deployment on researcher-controlled hardware rather than centralized service provision. Nonetheless, systematic feed logging raises broader ethical questions regarding surveillance normalization, contextual integrity, and informed consent in studies involving social media content.

Taken together, these considerations underscore that the system should be understood as a methodological instrument whose outputs require interpretive caution, documentation, and, where appropriate, human oversight.

6. Evaluation

We are currently implementing a two-part evaluation protocol addressing (1) the reliability of zero-shot LLM classifications and (2) the behavioral and perceptual effects of the nutrition label as a transparency intervention.

6.1. Classification Reliability and Agreement

To evaluate classification performance, we export recorded exposure datasets in the system's custom JSON format and distribute them to independent reviewers. Reviewers import the dataset into their local instance of the extension and use the built-in validation interface to annotate whether predicted category assignments are correct.

The exported annotation files are returned and aggregated using a dedicated analysis script. For each category, we compute:

- Inter-annotator agreement (e.g., Cohen's or Fleiss' κ)
- Majority-based ground truth labels
- LLM performance metrics with respect to majority assessments (e.g., accuracy, precision, recall)

This protocol enables systematic evaluation of zero-shot category definitions and provides insight into which types of categories (e.g., descriptive vs. normative) yield higher agreement and classification reliability.

6.2. User Perception and Intervention Effects

Beyond classification reliability, we plan to investigate the effects of the nutrition label as a transparency intervention. Participants will install the extension and use it with their own feed. The study will explore:

- How exposure summaries influence scrolling and engagement behavior
- Whether awareness of aggregated exposure alters perceptions of feed composition
- How accurately users can predict the distribution of their feed across categories prior to viewing the nutrition label

In particular, we are interested in comparing users' predicted category distributions with measured distributions derived from recorded exposure data. This comparison may reveal systematic misperceptions regarding feed composition and algorithmic influence.

Together, these evaluation components address both methodological validity of LLM-based classification and the societal implications of transparency interventions.

7. Discussion and Outlook

We presented a research-oriented framework for operationalizing feed-level exposure analysis using multimodal large language models. By combining view-time logging, structured zero-shot classification, retrieval-grounded conversational querying, and human validation mechanisms, the system provides an integrated environment for studying algorithmically curated media streams.

The framework contributes methodologically in three ways. First, it introduces time-weighted exposure as an operational construct for analyzing scrolling-based media consumption. Second, it demonstrates how open-weight multimodal LLMs can be embedded into reproducible research workflows through structured prompting, exportable datasets, and agreement-based evaluation. Third, it positions transparency interfaces not only as user-facing tools but as experimental instruments for studying the behavioral and perceptual effects of algorithmic disclosure.

The planned evaluation protocol will assess both classification reliability and the epistemic status of zero-shot category definitions across descriptive and normative constructs. By combining inter-annotator agreement with majority-based performance metrics, we aim to clarify under which conditions LLM-based feed analysis yields stable and interpretable results.

Beyond methodological validation, the system enables empirical investigation of transparency interventions. Comparing users' predicted feed composition with measured exposure distributions may reveal systematic misperceptions about algorithmic influence and content balance. Such findings could inform debates in platform governance, digital literacy, and algorithmic accountability.

Future work includes extending support to additional platforms, incorporating richer video analysis, exploring multilingual category definitions, and formalizing reproducibility standards for LLM-based content analysis pipelines. As multimodal models continue to evolve, maintaining transparency in prompt design, model versioning, and hardware configuration will remain essential for ensuring methodological rigor.

By bridging exposure logging, multimodal reasoning, and human validation, the proposed framework contributes to emerging methodological toolkits for studying algorithmically curated information environments in the Social Sciences and Humanities.

References

Alibaba Cloud. 2024. Qwen2.5-vl-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct.

Luca Belli and Marlena Wisniak. 2023. What's in an algorithm? empowering users through nutrition labels for social media recommender systems. *Knight First Amendment Institute at Columbia University*.

European Union. 2022a. Article 27: Recommender system transparency, digital services act (eu) 2022/2065. Obliges online platforms to explain the main parameters of recommender systems and provide users with meaningful options to influence content recommendations.

European Union. 2022b. Article 40: Data access and scrutiny, digital services act (eu) 2022/2065. Providers of Very Large Online Platforms must provide access to data for vetted researchers to study systemic risks.

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2017. An information nutritional label for online documents. *SIGIR Forum*, 51(3):46–66.

Tim Gollub, Martin Potthast, and Benno Stein. 2018. Shaping the Information Nutrition Label. In *2nd International Workshop on Recent Trends in News Information Retrieval (NewsIR 2018) at ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 9–11.

Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Şahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker. 2018. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 28–33, Brussels, Belgium. Association for Computational Linguistics.

Stijn Peeters. 2025. Zeeschuimer.

John Willinsky and Daniel Pimentel. 2024. The publication facts label: A public and professional guide for research articles. *Learned Publishing*, 37(2):139–146.

Charting the European LLM Benchmarking Landscape: A New Taxonomy and Registry

Špela Vintar*†, Taja Kuzman Pungeršek*, Mojca Brglez*†, Nikola Ljubešić*•‡
*Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
{spela.vintar, taja.kuzman, mojca.brglez, nikola.ljubesic}@ijs.si
†Faculty of Arts, University of Ljubljana, Aškerčeva 2, Ljubljana, Slovenia
• Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
‡ Institute of Contemporary History, Privoz 11, Ljubljana, Slovenia

Abstract

While new benchmarks for large language models (LLMs) are being developed continuously to catch up with the growing capabilities of new models and AI in general, using and evaluating LLMs in non-English languages remains a poorly-charted landscape. We give a concise overview of recent developments in LLM benchmarking, and then propose a new taxonomy for the categorization of benchmarks that is tailored to multilingual or non-English use scenarios. We further propose a registry of benchmarks implementing the new categorization and documenting benchmarks with a rich set of metadescriptors. While still at a pilot stage, such a registry can lead to a more coordinated development of benchmarks for European languages. We conclude with a review of current trends and advocate for a higher language and culture sensitivity of evaluation methods.

Keywords: large language models, benchmarking, taxonomy, benchmark registry, cultural competence, quality recommendations

1. Introduction

The rapid advancement of large language models (LLMs) has brought unprecedented capabilities in natural language understanding and generation, reasoning, coding, and more. With the global race in raising the bar, new commercial models are being released on a weekly basis. They are also being employed in increasingly complex pipelines and are given more and more agency to carry out tasks, perform strategic planning, and independently interact with other applications. While open-source models generally score lower on most leaderboards, they too grow larger and smarter.

In the brief history of LLMs, many evaluation frameworks have been set up – both human and automatic – to assess their evolving performance across different linguistic and non-linguistic tasks of growing complexity. However, the overwhelming majority of evaluation datasets are developed primarily for English, creating a significant evaluation gap for other languages and language varieties.

To evaluate the performance of LLMs in non-English contexts, a widely used approach so far has been to translate existing English benchmarks using machine translation, with or without human revision. This might seem reasonable: several major international benchmarks or benchmark collections (e.g., SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), Hellaswag (Zellers et al., 2019) exist together with their parallel (translated) versions, which allows for a direct comparison of LLMs across a wide range of tasks and languages. However, Global-MMLU (Singh et al., 2024) revealed that success in MMLU depends heavily on learning Western-centric concepts, with 28% of all questions requiring culturally sensitive knowledge. Moreover, for questions requiring geographic knowledge, an astounding 84.9% focus on either North American or European regions. Such cultural biases are not uncommon in other widely used benchmarks, and their machine-translated versions overlook language- and culture-specific phenomena, exhibit skewed performance, or fail to address issues which may be critical for users of an LLM in a particular language.

On the other spectrum of multilingual evaluation, there are several cases of language- and culture-specific benchmarks, such as BenCzechMark (Fajcik et al., 2025), PLCC (Dadas et al., 2025) or ITALIC (Seveso et al., 2025), which have been developed specifically for a particular language or community of speakers. Such benchmarks provide a deeper insight into a model's performance for that language, but typically do not allow us to assess the model's multilingual capacities.

Despite the proliferation of evaluation platforms, research projects, and benchmarking initiatives across the multilingual landscape, the field lacks a comprehensive overview that synthesizes current practices, identifies critical gaps, and provides clear guidance for developing more inclusive and effective evaluation methodologies for LLMs in non-English contexts. To address this gap, this paper

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 205–217
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

makes the following contributions:

1. We propose a new **benchmark taxonomy** that is designed to better capture the linguistic and cultural diversity in non-English settings.
2. We propose a **structured registry of European benchmarks** and present its pilot implementation.
3. We outline a set of **methodological recommendations** for the development of future benchmarks, and give an overview of recent trends.

The remainder of this paper is structured as follows. In Section 2, we review recent developments in LLM benchmarking, with particular attention to multilingual evaluation and to approaches that address linguistic and cultural competence. In Section 3, we propose a new categorization of benchmarks, and in Section 4, we propose quality standards and metadescriptors, which in combination serve as a foundation for charting ongoing and future LLM evaluation activities in Europe and beyond. We conclude with a review of recent trends and reflect on best practices.

2. Recent Developments in LLM Benchmarking

Several comprehensive overviews of LLM benchmarking have recently been published, including Chang et al. (2023) and Ni et al. (2025). Both surveys clearly show trends in both the development of datasets and the evolution of evaluation metrics. However, they lack a focus on non-English and multilingual scenarios, which is the main motivation for our own overview.

2.1. Major Benchmarks

By major or global, we refer to benchmarks most frequently used in current evaluation platforms and leaderboards. These benchmarks are without exception in English. The ones that evaluate generic language understanding and commonsense reasoning have their origins in the 2018–2022 period, when the challenges still roughly corresponded to natural language processing (NLP) research areas. Some of these benchmarks have seen multiple revisions, extensions and updates, and can be seen as "parent" datasets on which many adaptations, translations, or local versions are based.

Some of the most prominent for language understanding and reasoning include MMLU (Hendrycks et al., 2021) and its derivatives MMLU-Pro (Wang et al., 2024), MMLU-Prox (Xuan et al., 2025) and Global MMLU (Singh et al., 2024); the SuperGLUE benchmark collection comprising BoolQ (yes/no questions, Clark et al., 2019), CommitmentBank (textual entailment, De Marneffe et al., 2019), COPA (Choice of Plausible Alternatives for causal reasoning, Roemmele et al., 2011), MultiRC (multi-sentence reading comprehension, Khashabi et al., 2018), ReCoRD (reading comprehension with commonsense reasoning, Zhang et al., 2018), RTE (Recognizing Textual Entailment, Giampiccolo et al., 2007), WiC (Words in Context, Pilehvar and Camacho-Collados, 2019), and WSC (Winograd Schema Challenge, Levesque et al., 2012); ARC (Clark et al., 2018) with multiple-grade science questions; Hellaswag (Zellers et al., 2019) and its recent derivative GoldenSwag (Chizhov et al., 2025). An attempt to create a more challenging benchmark collection is BIG-bench (Beyond the Imitation Game Benchmark, Srivastava et al., 2023), a massive collaborative benchmark consisting of 204 tasks contributed by more than 450 authors across 132 institutions, designed to probe large language models on tasks believed to be beyond their current capabilities. Finally, SUPERB (Speech processing Universal PERformance Benchmark, Yang et al., 2021) is a unified speech-focused benchmark for evaluating self-supervised and general-purpose speech representations across a wide spectrum of speech processing tasks.

2.2. Multilingual Benchmarks

Most state-of-the art models have multilingual capabilities, but the precise amounts of non-English data used in their pre-training are usually obscure. It is therefore hard to say to what extent the language competence of a model in a particular language is in correlation with the amount of language-specific data it has seen. In addition to this, models differ in their representations of intermediate layers, which may result in cultural conflicts between latent internal and target output language (Zhong et al., 2024).

Since many authors observe a marked decline in performance for low-resource languages, benchmarks are now being developed both as parallel evaluation sets based on existing "parent" datasets to allow for direct comparison across languages, and as language-specific benchmarks, usually aimed at assessing LLM performance in a particular linguistic community and/or culture (see Section 2.4 for the latter).

Although the datasets in the first category are parallel, they may differ considerably in the methods used for their creation. Some were translated using machine translation engines or LLMs, for example, EU20-MMLU, EU20-HellaSwag, EU20-ARC, EU20-TruthfulQA, and EU20-GSM8K (Thellmann et al., 2024); or MMLU-Prox (Xuan et al., 2025). Other multilingual benchmarks were created with a special focus on cultural sensitivity by dividing the original subsets into culturally sensitive and culturally agnostic ones (Global MMLU, Singh et al., 2024), or by using professional translators or multiple rounds of revision to raise the quality of the dataset, e.g., BenchMax (Huang et al., 2025), Flores-101 and FLORES-200 (Goyal et al., 2022) and Belebele (Bandarkar et al., 2024).

For speech, ML-SUPERB (Multilingual Speech processing Universal PERformance Benchmark, Shi et al., 2023) extends the English SUPERB speech benchmark to 143 languages, evaluating self-supervised speech representations on automatic speech recognition and language identification. FLEURS (Conneau et al., 2022) is a speech-based extension of the FLORES multilingual benchmark, with focus on language identification, automatic speech recognition, and retrieval evaluation. DIALECTBENCH (Faisal et al., 2024) is the first large-scale benchmark for language variety understanding, aggregating 10 text-level tasks for 281 varieties.

2.3. Dynamic Benchmarks

Recent developments emphasize dynamic and contamination-resistant evaluation. The period 2022–2025 has witnessed fundamental shifts toward more sophisticated evaluation approaches. One such attempt is LiveBench (White et al., 2024), the first benchmark designed to resist training data contamination through frequently updated questions from recent sources, automatic scoring, and monthly updates. This dynamic approach remains challenging, with the top models achieving accuracy below 80%.

2.4. Language- and Culture-Specific Benchmarks

Many European languages have established their own country- or region-specific evaluation frameworks, and in most cases, these combine traditional datasets translated from English with more culture-aware native benchmarks. Examples include HuLu[1] for Hungarian (Ligeti-Nagy et al., 2024), which covers a number of well-known tasks such as plausible alternatives (HuCoPa), textual entailment (HuRTE), and linguistic acceptability (HuCoLa), of which the latter was originally constructed using sentences from selected Hungarian linguistics books. BenCzechMark for Czech (Fajcik et al., 2025) is a complex benchmark collection comprising 50 tasks, of which 14 were newly created, and only 10% of the collective instances were machine-translated. Another recent benchmark for Czech, Ukrainian, and Slovak called CUS-QA (Libovický et al., 2025) focuses specifically on cultural competence and crafts questions, both textual and visual, from Wikipedia articles that exist in only one of the languages.

For Iberian languages, a comprehensive and extensible framework has been established under IberBench (González et al., 2025), spanning 22 task categories and addressing both generic and industry-relevant tasks. In parallel and under a similar name, IberoBench (Baucells et al., 2025) offers 62 tasks, of which several were created from scratch from native data, and others were included only if they satisfied rather strict quality criteria.

For Slovenian, the SloBENCH[2] evaluation framework offers natural language inference (SI-NLI), machine translation, speech recognition, Slovene SuperGLUE, and two pragmatics benchmarks: SloPragMega and SloPragEval. To create the latter, full localization of the originally English dataset was performed by adapting cultural references and occasionally completely rewriting examples to better match the linguistic and cultural context. A similar approach has been taken to translate and adapt the COPA benchmark (Roemmele et al., 2011) to four standard languages and three dialects of the South Slavic language group, resulting in the DIALECT-COPA (Ljubešić et al., 2024) benchmark collection. Today's best-performing proprietary models still score only halfway between random and optimal on dialectal data (Chifu et al., 2024).

A fully native benchmark is ITALIC for Italian (Seveso et al., 2025), which comprises 10,000 instances from 12 domains and was built entirely from exam materials offered by various public institutions or government bodies.

An interesting combination of multilinguality and culture-specificity is BLEnD (Myung et al., 2024), a benchmark probing cultural competences of LLMs in 13 languages. The questions were manually crafted for each language using a set of templates, targeting cultural competence in 6 domains: food, sports, family, education, holidays, and work life.

An exceptionally active approach to language- and culture-specific benchmarking can be observed for Polish[3], with a range of generic and domain-specific evaluations including multi-turn conversation (MT-Bench), emotional intelligence (EQ-Bench), comprehensive text understanding (CPTUB[4]), a medical domain benchmark, linguistic and cultural competency (PLCC, Dadas et al., 2025), educational (LLMs Behind the School Desk), a cultural vision benchmark, and legal QA. Most of these benchmarks were developed anew, by care-

[1] https://hulu.nytud.hu

[2] https://slobench.cjvt.si

[3] https://huggingface.co/spaces/speakleash/polish-llm-benchmarks

[4] https://huggingface.co/spaces/speakleash/cptu_bench

fully selecting tasks and examples, verifying them by experts, and collecting human annotations. Similarly, the PLCC consists of 600 manually crafted questions and is divided into six categories: history, geography, culture and tradition, art and entertainment, grammar, and vocabulary. The leaderboard results[5] indicate that even the largest models still reach mediocre performance in Polish grammar and vocabulary, thus justifying the need for a detailed assessment of linguistic competence for other European languages as well. A final example of a culturally specific benchmark is the Polish Cultural Vision Benchmark[6], a collection of images with text descriptions to evaluate the cultural competence of multimodal models. The dataset is part of a citizen science project aimed at collecting 1 million culturally specific images[7] and recruiting user donations under the slogan of "technopatriotism". While similar platforms have been established before to collect text data, this is a positive example of a contemporary and at the same time participatory benchmark.

As a final note on language- and culture-specific benchmarks, while the efforts and datasets listed above are certainly a step in the right direction, the European LLM community would benefit from a common methodology for creating such evaluation sets, starting with a clear overview of the tasks/aspects involved in linguistic and cultural competence and of the benchmarks that address them.

3. Categorization of Benchmarks

3.1. Existing Taxonomies

As new benchmarks are continuously presented to evaluate the emerging capabilities of LLMs, many attempts have been made to organize them in a structured and logical way.

The **AI Verify Foundation** has established one of the most systematic approaches to LLM benchmark categorization globally. In their October 2023 publication "Cataloguing LLM Evaluations" (AI Verify Foundation, 2023), LLM benchmarks are organized into 5 top categories (further divided into subcategories). These are *General Capabilities* (natural language understanding, natural language generation, reasoning, knowledge and factuality, tool use effectiveness, multilingualism, and context length handling); *Domain Specific Capabilities* (specialized industry performance across various domains); *Safety and Trustworthiness*; *Extreme Risks*; and *Undesirable Use Cases*.

The catalogue represents a comprehensive and valuable contribution to the field, and has many positive features: The taxonomy is based on LLM capabilities, occasionally also referred to as tasks, which seems intuitively most pragmatic, as this is usually the way we think about (and evaluate) human performance too. Complex benchmarks can appear in several categories simultaneously (e.g., BigBench as a massive collaborative benchmark appears in almost all taxonomy categories), and the recommendations for future LLM evaluations are a solid starting point to reinforce minimum quality standards for fair and trustworthy LLM assessment.

However, the catalogue also has some drawbacks which render it unsuitable for our purposes. Firstly, it has not been updated since 2023 and hence does not include many benchmarks that have since become mainstream, nor does it address recent developments in LLMs and AI in general. Secondly, although it includes Multilinguality as a separate category, it falls short in capturing some aspects of LLM performance which may be critical for the evaluation of European models; i.e., models specifically developed to be used in region-, language-, culture-, or domain-specific contexts. Thirdly, and this is less of a drawback than simply an observation, the taxonomy and the quality recommendations are primarily focused on the safety and trustworthiness of LLMs, in the context of AI governance and alignment research. While these are indeed crucial priorities, especially for the so-called "frontier models", the European landscape of LLM development and evaluation is – at least for now – gyrating around a different set of goals, such as how to reach state-of-the-art levels of understanding and generation in non-English languages, or how to de-bias English-centric models.

Other approaches to taxonomization include HELM (Holistic Evaluation of Language Models), developed at Stanford University (Liang et al., 2023). The authors introduce the concept of scenarios (what we want to evaluate) and metrics (which performance aspects are measured, and how), then propose a taxonomy of scenarios and desiderata. Today, the framework[8] includes a number of leaderboards with support for multimodality and model-graded evaluation. While the scenarios proposed in HELM and the framework itself leave room for continuous extension, they do not, in fact, offer a hierarchical structure with sufficient focus on multilinguality and issues related to the use of LLMs in non-English contexts.

Similarly, Chang et al. (2023) provide an overview of existing LLM evaluations, which they examine from three aspects: what, where, and how to eval-

[5] https://huggingface.co/spaces/sdadas/plcc

[6] https://huggingface.co/spaces/speakleash/Polish_Cultural_Vision_Benchmark

[7] https://obywatel.bielik.ai

[8] https://crfm.stanford.edu/helm/

uate. They divide the evaluation tasks into eight top-level non-exclusive categories, namely *Natural language processing*; *Robustness, ethics, biases and trustworthiness*; *Social sciences*; *Natural science and engineering*; *Medical applications*; *Agent applications*; and *Other applications*.

Huber and Niklaus (2025) present a cognitive-based view on benchmarking by mapping the well-known Bloom's taxonomy of cognitive abilities to LLM capabilities across six hierarchical cognitive levels: *Remember, Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*, revealing significant gaps in the coverage of higher-order thinking skills.

Another comprehensive attempt at taxonomizing benchmarks is by Guo et al. (2023) who introduce a three-pillar framework that categorizes LLM evaluation into three major groups: *Knowledge and capability evaluation*, *Alignment evaluation*, and *Safety evaluation*.

3.2. New Taxonomy Proposal

In this paper, we present a new taxonomy for the categorization of LLM benchmarks and its pilot implementation in an online database. The categorization and the proposed system of metadescriptors allow us to better compare LLMs across languages, easily find benchmarks for a specific language, use case, modality, or domain, to set common priorities, and to work towards common goals.

Our taxonomy is loosely based on AI Verify Foundation's catalogue, with the following important modifications:

- We merge all **language-related tasks and scenarios** under a single top-level category called Language capabilities.
- We further **dissolve the traditional NLP divide between natural language understanding and natural language generation** into a single subcategory. The fact is that state-of-the-art LLMs more often than not combine these two capabilities, and even straightforward tasks such as question answering or text summarization involve both.
- We **expand the category for general linguistic competence** with further subcategories for style, conversation, and pragmatics, and allow for other, more fine-grained aspects of measuring the grammaticality or coherence of generated outputs.
- We **expand the category of specific linguistic competence** to include creativity, atypical communication, the use of domain-specific language, etc.
- We also **expand and redefine the category of multilinguality** to include code-switching, multilingual interaction, and dialectal flexibility.
- We introduce **cultural competence** as a separate category.
- We introduce **speech** as a separate category to include benchmarks specifically aimed at performing tasks unique to spoken language as input or output.[9]
- We add **agency** as a form of long-term, consistent, or strategic reasoning.

Figure 1 shows the four top-level taxonomy categories with subcategories. The last category, Alignment and dangerous capabilites, currently has no subcategories for the simple reason that we did not find many European benchmarks addressing this field. The full taxonomy, along with fine-grained third-level categories, can be inspected at our demo website[10]

The proposed taxonomy serves as a *hierarchy of labels* to organize benchmarks. As many (or indeed most) belong to more than a single category, they can be assigned non-exclusive categories, and can be further characterized with metadescriptors that we present in the following section.

4. Quality Standards and Metadescriptors for Benchmarks

As demonstrated in our demo benchmark registry (which, at the time of writing, contains 91 benchmarks), the LLM development community might benefit from a set of common metadescriptors which are used to label evaluation sets. Below, we present some of the features we consider important in terms of benchmark categorization and quality assurance.

4.1. Provenance

While it is clear that the development of original and sufficiently complex evaluation datasets is highly time-consuming and costly, the drawbacks of automatically translated and culturally maladapted benchmarks have been clearly pointed out (Singh et al., 2024; Xuan et al., 2025). We should thus strive towards clearer – even if more complex –

[9]We propose Modality as one of the metadescriptors, allowing for any benchmark to be implemented in any of the modes. The Speech category refers to evaluations targeted at speech-related activities.

[10]The full taxonomic hierarchy with category descriptions and a demo benchmark registry: `https://mojcabrglez.github.io/EU-LLM-Bench/`

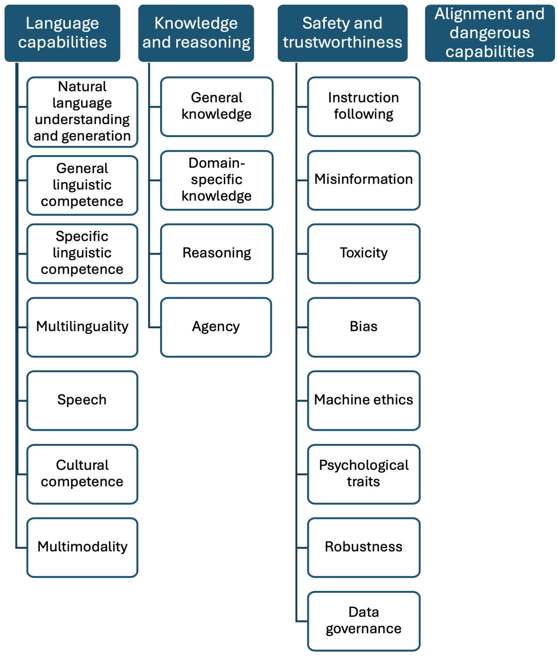


Figure 1: Top-level categories with subcategories.

descriptors which indicate how a dataset or benchmark was created. We propose the following descriptors:

- **Original** Applies to datasets which have been originally created in the language they appear in, *by any method other than translation* (e.g., collecting original exam questions, employing experts to provide domain-specific tasks, adapting authentic texts in a particular language to create tasks in that language).

- **Machine translation** Applies to datasets created *by any automated translation service*, including those created by LLMs, and workflows with machine revision. The tools and workflows must be specified.

- **Machine translation and Human revision** Applies to datasets where the result of *automated translation was revised by a human* professional or non-professional translator or reviewer. Since many non-English benchmarks are created by machine translating the – usually – English original, followed by human revision of only a small portion of the dataset, the recommendation would be to use all labels that apply.

- **Human translation** Used only for datasets which have been *fully translated and revised by humans*. Only a few European non-English datasets satisfy this criterion.

- **Full localization** Used for datasets which have not only been translated professionally, but for which a *full linguistic and cultural adaptation* was performed. This might mean the replacement of culturally-specific or untranslatable tasks with new ones, or removing parts of the dataset deemed culturally unsuitable.

- **Other** If several of the above scenarios were used, the dataset should be labeled with all that apply. Other methods and scenarios used to create the dataset should be specified here.

4.2. Accessibility

The tension between open benchmarking and data contamination poses a significant challenge for AI evaluation. While public datasets enable reproducible research and fair comparisons, they risk contamination when models train on test data, inflating performance scores and undermining benchmark validity. Private evaluation sets offer a potential solution by keeping data hidden from training processes, ensuring cleaner performance measurements.

- **Public** Applies to fully open datasets shared with labels through common platforms.

- **Public without labels** Applies to datasets where labels are not distributed to prevent direct training.

- **Private (academic/research access)** Where authors encourage reproducibility but wish to prevent contamination.

- **Private (closed/proprietary)** Where datasets are typically not shared as they are used for internal or industry-specific evaluation.

- **Other** This may include dynamic benchmarks where tasks are continually updated (such as White et al., 2024).

4.3. Language Coverage

This category indicates the prominence, reach, or scale of a benchmark in terms of its presence on major leaderboards, coverage, global spread, but also purpose. We are aware that the boundaries between the proposed buckets may be fuzzy.

- **Major global benchmark** See section 2.1 for examples.

- **Multilingual benchmark** This category can be used for benchmarks derived from the above, for instance, by developing a multilingual variant of a well-known benchmark for a set of new languages. Examples include XCOPA (Ponti et al., 2020), MMLU-Prox (Xuan et al., 2025) or xHumanEval (Raihan et al., 2025).

- **Language-group or region-specific benchmark** This category is used for language-specific benchmarks as well as benchmarks that cover multiple languages from a similar language group or target a certain geographical region. Examples include IberoBENCH (Baucells et al., 2025) and DIALECT-COPA (Ljubešić et al., 2024).

4.4. Evaluation Type

An important factor for present and future benchmarks is the divide between **closed-ended** types of tasks, most prominently multiple-choice questions, but also other types of tasks where the solution is included in the task, and **open-ended** tasks, typically generation of text, speech, image, or multimodal output. Few benchmarks to date address the latter, despite the fact that generative LLMs are now mainstream and the vast majority of application scenarios exploit generative abilities.

4.5. Evaluation Metrics and Frameworks

The performance of an LLM can be evaluated in several ways, depending on the type of task. For multiple choice questions, text classification tasks, or cloze tests, where the correct answer is determinate, **accuracy** or **F1** (Powers, 2020) can be used. However, to evaluate longer, more complex responses resulting from generative tasks, many other methods were proposed. In reference-based evaluation, the LLM response is compared to a reference using various distance measures (e.g., **BERTScore** (Zhang et al., 2020), **Rouge-1** (Lin, 2004), **METEOR**(Banerjee and Lavie, 2005)), while in reference-less contexts, the quality of the response is directly assessed (e.g., by an **LLM-as-a-judge**). As we have seen, recently developed benchmarks employ more complex evaluation methodologies, and a common alternative to algorithmic benchmarks is human preference voting (in so-called chatbot arenas, e.g., https://lmarena.ai/).

Another important element is the **existence of a human baseline**, and its quality. Important factors to consider are participant selection and training; task design and instructions (to ensure fair comparison between humans and LLMs); and control for attention, bias, or fatigue. Human annotators or participants can also provide relevant insight into the overall quality of the benchmark.

If the benchmark or dataset is integrated into an evaluation framework, this should be indicated together with a link or other reference to the evaluation site.

4.6. Other Metadata

We propose collecting rich metadata for each benchmark that allows researchers to quickly understand its content, characteristics, and provenance (see demo registry[11] for examples):

- **Description**: A short summary of the dataset's content and purpose.
- **Benchmark family**: The broader benchmark initiative or collection to which the dataset belongs. For example, the COPA benchmark family would include the English COPA (Roemmele et al., 2011) and its many parallel variants, such as the COPA datasets in Hungarian (Ligeti-Nagy et al., 2024), Croatian (Ljubešić and Lauc, 2021), South Slavic dialects (Ljubešić et al., 2024), etc.
- **Number of test instances**: The total number of instances in the test set, enabling quick comparison of scale.
- **Language**: The language(s) in which the dataset is provided.
- **Language type**: Specification of whether the language of the dataset is standard, non-standard, or a dialect.
- **Modality**: The input modality of the dataset, such as text, speech, sign language, or audio-visual signal.
- **Authors**: The creators or curators of the dataset.
- **Paper link**: Reference to the main publication describing the dataset.
- **Access info**: Information on how to obtain the dataset, e.g., a link to a website or repository from where the dataset can be acquired, or contact information of the dataset owner if not public.
- **Last revised**: The date of the most recent update or revision of the dataset.
- **More information**: Additional notes, links, or resources relevant to the dataset.

5. Trends and Future Directions

Several challenges of LLM evaluation have been pointed out by a number of studies (e.g., Laskar et al., 2024 or AI Verify Foundation, 2023: p. 16–22), most notably reproducibility, reliability (including contamination, obscure evaluation methods,

[11]The demo registry of LLM benchmarks is available at https://mojcabrglez.github.io/EU-LLM-Bench/.

and unfair comparisons), and robustness. There are numerous parallel activities in progress to set the course of the European LLM evaluation landscape, agree on common principles, and establish a dialogue between different stakeholders. While a full review of the above-mentioned challenges is beyond the scope of this paper, we list some trends which apply in particular for the European benchmarking landscape and evaluations in non-English settings.

5.1. Translated vs. Native

As outlined in the sections above, it has meanwhile been widely recognized that translated benchmarks suffer from translationese, English-centrism, and bias. The trend towards native benchmarks is increasing (see Section 2.4), and language-specific nuances can be evaluated more accurately. Native benchmarks have been or are being constructed for many EU languages, including Polish, Czech, Hungarian, Italian, Spanish, and others. One area where native benchmarks are certainly better than translated ones is figurative language entailing humour, sarcasm, metaphor, idioms, and word play.

5.2. Language- and Cultural Competence for In-house Models

With AI expanding to most business and public sectors, and with the trend to deploy in-house models instead of relying on major commercial ones, there is an urgent need to evaluate such models in terms of their language and culture capabilities in the local setting. For many use cases, evaluation via public evaluation platforms may not be practical or feasible, so alternatives should be provided.

5.3. Speech and Other Modalities

While the SUPERB benchmark (Yang et al., 2021) marked the beginning of a new era in the evaluation of speech processing, and projects like Mozilla's Common Voice[12] provide downloadable datasets for many languages, the landscape of non-English speech benchmarks is still sparsely populated. The creation of evaluation sets for all spoken varieties, including dialect, slang, child speech, and learners' speech, should be a priority, as well as comprehensive coverage of all speech-related phenomena. Even less data is available for sign languages, although projects like SignON (Shterionov et al., 2022) have provided valuable resources.

[12] https://commonvoice.mozilla.org/en

5.4. Vision and Multimodal Language Models

With the rise of models that can process images and video input, as well as generate visual or multimodal output, it is important to consider the cultural dimension of visual communication. It may seem that images are language-independent (as some indeed may be), but one should keep in mind that a large portion of the world we see through our eyes is actually embedded into a particular geographic, cultural, and historical context. Datasets for the training and evaluation of VLMs should incorporate the cultural dimension and should be locally or regionally specific.

5.5. Context-specific Evaluations

There is a lack of nuanced, context-specific evaluations that address the multi-faceted nature of real-world LLM deployments. This includes domain-specificity, but also other elements of attuning evaluation to the users it serves. For example, legislative frameworks can differ wildly across languages and cultures, so any AI-driven public service needs to be tailored to the relevant legal context it operates in.

5.6. Emerging Capabilities

As new capabilities emerge, such as agentic AI, long-term memory and reasoning, and physical interaction, so must new evaluation sets. So far, these developments are in progress mostly for large commercial solutions, but may shortly also impact the European research community. Especially the EuroHPC AI Factories[13] initiative aims to boost innovation for EU businesses and public entities alike.

We believe that the rapidly evolving benchmarking landscape for European languages can benefit from a registry, which we present as a demo implementation. The categorization and documentation of benchmarks according to the principles discussed above may facilitate collaboration and coordination of efforts, while at the same time contribute to the overall quality and transparency of LLM evaluation practices. The presented solution is extensible and flexible in that a benchmark may be assigned multiple categories and features, and new categories and metadescriptors may be added at any time.

The most important question remains: how can such a registry be maintained and regularly updated? The first version of our demo implementation was presented at the 2025 Clarin.eu conference, and the LLMs4SSH community was invited

[13] https://www.eurohpc-ju.europa.eu/ai-factories_en

to contribute. Only a few researchers responded, hence the registry in its current form is far from representative. It is perhaps optimistic to expect from researchers that they would invest effort into extra documentation activities, however, if such activities were sufficiently supported and prioritized by an organisation like Clarin.eu, the benchmark registry could conceivably evolve into something similar as the Resource Families[14], or become a part thereof.

6. Conclusion

We have presented recent trends in LLM benchmarking for European languages and proposed a new taxonomy for their categorization, intended to be implemented alongside a range of metadescriptors in the context of a registry of LLM benchmarks. Our taxonomization strategy focuses on the linguistic, cultural, factual, and reasoning capabilities of models and also incorporates emerging abilities. The proposed considerations follow the widespread belief amongst European researchers and developers that the traditional Western-centric, likely contaminated and linguistically inappropriate datasets no longer satisfy our needs, and that targeted efforts should be invested in filling the evaluation gaps for all European languages.

We also present a demo implementation of the above proposal, currently including 91 benchmarks categorized according to the new taxonomy and offering basic statistics about language, domain and/or LLM capabilities covered.

The initiative presented in this paper is the result of a series of discussions and reflections within the framework of several international research communities, collecting and integrating feedback from a number of researchers, developers, and benchmark creators. With the rapid advancement of the field, we envisage continuous extensions and revisions both of the taxonomy and the associated set of metadescriptors and recommendations.

7. Acknowledgements

This work was supported by the projects “Large Language Models for Digital Humanities” (Grant GC-0002), “Early parent-child communication in Slovenian: Corpus-based Insights” (Grant J6-70222), the research programme “Language Resources and Technologies for Slovene” (Grant P6-0411), “Slovene Language - Basic, Contrastive, and Applied Studies” (P6-0215), the "Jožef Stefan" Infrastructure Programme (Grant I0-0005) and the Research Infrastructure DARIAH-SI (Grant I0-E007), all funded by the ARIS Slovenian Research and Innovation Agency.

[14] https://www.clarin.eu/resource-families

8. Bibliographical References

AI Verify Foundation. 2023. Catalogue of LLM Evaluations.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. 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.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Pavel Chizhov, Mattia Nee, Pierre-Carl Langlais, and Ivan P Yamshchikov. 2025. What the HellaSwag? On the Validity of Common-Sense Reasoning Benchmarks. *arXiv preprint arXiv:2504.07825*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, and Rafał Poświata. 2025. Evaluating Polish linguistic and cultural competency in large language models. *arXiv preprint arXiv:2503.00995*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, et al. 2025. BenCzechMark : A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism. *Transactions of the Association for Computational Linguistics*, 13:1068–1095.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

José Ángel González, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2025. IberBench: LLM Evaluation on Iberian Languages. *arXiv preprint arXiv:2504.16921*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2310.19736*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. *arXiv preprint arXiv:2502.07346*.

Thomas Huber and Christina Niklaus. 2025. LLMs Meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, pages 552–561. Institute of Electrical and Electronics Engineers Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Jindřich Libovickỳ, Jindřich Helcl, Andrei Manea, and Gianluca Vico. 2025. CUS-QA: Local-Knowledge-Oriented Open-Ended Question Answering Dataset. *arXiv preprint arXiv:2507.22752*.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. HuLU: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties,* *and Dialects (VarDial 2024)*, pages 89–98, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, et al. 2025. A Survey on Large Language Model Benchmarks. *arXiv preprint arXiv:2508.15361*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mHumanEval - a multilingual benchmark to evaluate large language models for code generation. 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 11432–11461, Albuquerque, New Mexico. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proceedings of the 2011 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. 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 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei-Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. Ml-superb: Multilingual speech universal performance benchmark. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2023*, pages 884–888.

Dimitar Shterionov, Mirella De Sisto, Vincent Vandeghinste, Aoife Brady, Mathieu De Coster, Lorraine Leeson, Josep Blat, Frankie Picron, Marcello Paolo Scipioni, Aditya Parikh, et al. 2022. Sign language translation: Ongoing development, challenges and innovations in the signon project. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 325–326.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. *arXiv preprint arXiv:2412.03304*.

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, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards Multilingual LLM Evaluation for European Languages. *arXiv preprint arXiv:2410.08928*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. 2024. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. *arXiv preprint arXiv:2406.19314*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. *arXiv preprint arXiv:2503.10497*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Wei-Cheng Huang, Wei-Cheng Tseng, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li Wang, Zhaoheng Ni Ma, Benjamin Chen, Chih-Liang Chang, Kevin Lin, Wen-Chin Huang, Andy T. Wu, Po-Chun Hsu, Chun-Lin Chen, Han Lu Cheng, Yu Tsao, Hung-Yi Hsieh, and Hung-Yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Proceedings of Interspeech*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

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*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond English-centric LLMs: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

Cross-Lingual Abstractive Keyphrase Generation for Historical Newspapers

Simon Clematide[1], **Jenifer Meyer**[1], **Juri Opitz**[1], **Maud Ehrmann**[2], **Kaspar Beelen**[3]
[1]Department of Computational Linguistics, University of Zurich (UZH), Switzerland,
[2]Digital Humanities Laboratories, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]School of Advanced Study, University of London, United Kingdom
<firstname>.<lastname>@{uzh.ch,epfl.ch,sas.ac.uk}

Abstract

We investigate large language models (LLMs) for cross-lingual abstractive keyphrase generation from historical newspapers. The task consists of producing a small set of English keyphrases for articles written in German, French, and Luxembourgish, combining translation, abstraction, and normalization. We conduct a human-centered pilot study comparing model outputs using human selections, LLM-as-judge assessments, and inter-annotator agreement analysis, followed by a medium-scale application to multilingual data from the Impresso corpus. Results show that LLM-generated keyphrases can support semantic enrichment and exploratory analysis of historical collections, while highlighting the subjective and methodologically challenging nature of keyphrase evaluation.

1. Introduction

Digitized historical newspapers provide unprecedented access to past public discourse, but their effective exploration remains challenging due to optical character recognition (OCR) noise, multilinguality, diachronic language variation, and sparse metadata (Bingham, 2010). Semantic enrichment techniques are therefore essential to support retrieval, browsing, and large-scale exploratory analysis (Düring et al., 2021).

Keyphrase-based indexing provides a compact abstraction of document content and follows a long archival tradition of concept-based indexing and thesaurus-driven cataloguing. Automatic keyphrase generation has therefore become essential for large digitized collections. However, most existing methods are designed for contemporary, monolingual, and relatively clean text, limiting their applicability to noisy historical data and low-resource languages (Chiron et al., 2017; van Strien et al., 2020; Ehrmann et al., 2023b).

Large newspaper digitization initiatives further highlight the need for semantic enrichment (Ridge et al., 2019; Neudecker, 2022; Ehrmann et al., 2023a). While full-text search enables lexical lookup, exploratory research often requires higher-level thematic access across time periods, languages, and publication contexts (Gaillard, 2022; Bunout et al., 2023).

Conceptual keyphrases provide a compact abstraction layer that can be aggregated, compared, and visualized across collections. In multilingual settings, English keyphrases additionally function as a pivot representation enabling cross-lingual comparison. They can therefore complement search, faceted navigation, and other semantic access methods, and may support exploratory interfaces by suggesting concepts users did not initially query (Whitelaw, 2015; Düring et al., 2024).

Recent instruction-following LLMs offer a promising alternative (Ouyang et al., 2022). These models can jointly perform translation, abstraction, and normalization, making them suitable for generating conceptual keyphrases across languages and time periods. We evaluate their use for cross-lingual conceptual keyphrasing of historical newspapers and assess their suitability for semantic enrichment of multilingual collections.

Task Definition We study cross-lingual abstractive keyphrase generation for historical newspapers, primarily drawn from Swiss and Luxembourgish collections. Given an article written in a source language (German, French, or Luxembourgish), the goal is to generate a small set (typically three to five) of *conceptual keyphrases in English* that summarize its main topics.

This task differs from classical extractive keyphrase extraction in three ways. First, the keyphrases need not occur verbatim in the source text but may be semantically implied. Second, they are produced in a target language (English), introducing an explicit cross-lingual abstraction step. Third, the objective is conceptual coverage and topical diversity rather than surface-form fidelity. We use the term *abstractive* to denote keyphrases that need not occur verbatim in the source article. Instead, the model is asked to infer higher-level conceptual descriptors, translate them into English, and normalize across OCR noise, historical spelling, and multilingual variation. In this sense, the task combines summarization, translation, and semantic abstraction rather than surface-form extraction alone.

Named entities (persons, locations, organizations, events) are deliberately excluded from the

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 218–223
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

> You are an archivist for historical newspaper articles who indexes historical newspaper articles with conceptual keyphrases in English. Given a JSON object containing metadata and a historical newspaper article, please index with an adequate number (between 3 and 5) of most relevant keywords in English in JSON format. Do not create keywords consisting of names for persons, locations, or events. In addition, add one summary sentence in English.

Figure 1: System prompt for keyphrase abstraction.

keyphrase inventory. This design choice reflects the availability of a dedicated named entity recognition and linking pipeline within the broader semantic enrichment workflow. By separating conceptual keyphrasing from entity-centric annotation, the approach avoids redundant effort and allows keyphrases to focus on higher-level thematic dimensions.

The keyphrases function as lightweight semantic descriptors for indexing, clustering, and exploratory analysis rather than as gold-standard annotations.

2. Related Work

Prior work on keyphrase extraction and generation typically assumes contemporary monolingual data. Historical texts pose additional challenges, including OCR noise and linguistic variation, and reliable gold standards are difficult to obtain (Piotrowski, 2012; McGillivray and Tóth, 2020).

Recent studies address multilingual keyphrase generation, for example through retrieval-augmented models (Gao et al., 2022) or multilingual datasets such as EUROPA (Salaün et al., 2024). However, cross-lingual abstractive keyphrase generation for noisy historical texts remains underexplored.

Recent work has begun to examine large language models for zero-shot keyphrase extraction and generation more systematically. Mohan et al. (2025) investigate instruction-tuned LLMs for zero-shot keyphrase generation and show that increasingly specialized instructions do not consistently improve results, whereas multi-sample generation with aggregation can yield clear gains. In a complementary setting, Kang and Shin (2025) study zero-shot keyphrase extraction and show that prompt design matters substantially: simple prompts can already be competitive, task-relevant role prompting often helps, and combining direct extraction with candidate-based selection can further improve performance. Together, these studies suggest that LLM-based keyphrase methods are highly sensitive to prompting and decoding choices, while also confirming their promise as flexible zero-shot alternatives to more traditional extraction and generation pipelines.

Given these challenges, we evaluate systems using an exploratory, human-centered framework that prioritizes adequacy, consistency, and downstream usefulness.

	DE1	DE2	FR1	FR2	LB1	LB2
Claude-3.5-Sonnet	3	5	4	5	4	4
DeepSeek-V3	3	5	5	5	5	5
GPT-3.5 Turbo	5	5	5	4	5	5
GPT-4o mini	3	5	3	5	5	5
# Unique KP	12	12	14	18	15	16

Table 1: Keyphrase counts by model and article

	DeepSeek-V3	GPT-3.5 Turbo	GPT-4o mini
Claude 3.5 Sonnet	10.4%	5.9%	4.1%
DeepSeek-V3	—	9.6%	12.5%
GPT-3.5 Turbo	—	—	8.3%

Table 2: Pairwise overlap (%) between models.

3. Pilot Study: Human and Model Judgments

To implement this evaluation strategy, we conducted a pilot study comparing human and model judgments.

Data and Models We selected a pilot sample (N=6) of historical newspaper articles from the *Impresso* corpus,[1] with two articles each in German, French, and Luxembourgish, covering different periods. Several LLMs (Claude 3.5 Sonnet, DeepSeek-V3, GPT-3.5 Turbo, GPT-4o mini) were prompted, using provider-default decoding settings, to generate English keyphrases using the instruction shown in Figure 1. The user prompt provided the article text together with two metadata fields, namely the newspaper title and the publication date.

Inter-Model Agreement Keyphrase counts are similar across models (Table 3), reflecting adherence to the prompt constraint. However, pairwise overlap scores remain low (Table 2), suggesting substantial variation in model interpretations of document content. Applying a stemmer before computing overlap does not substantially alter the results.

Low pairwise overlap should not be interpreted straightforwardly as system failure. For conceptual keyphrasing, especially in a cross-lingual historical setting, multiple non-overlapping but semantically plausible outputs may adequately represent the same article. At the same time, this variability limits the usefulness of strict lexical overlap as a primary evaluation criterion and cautions against strong claims of comparative superiority.

[1] `https://impresso-project.ch`

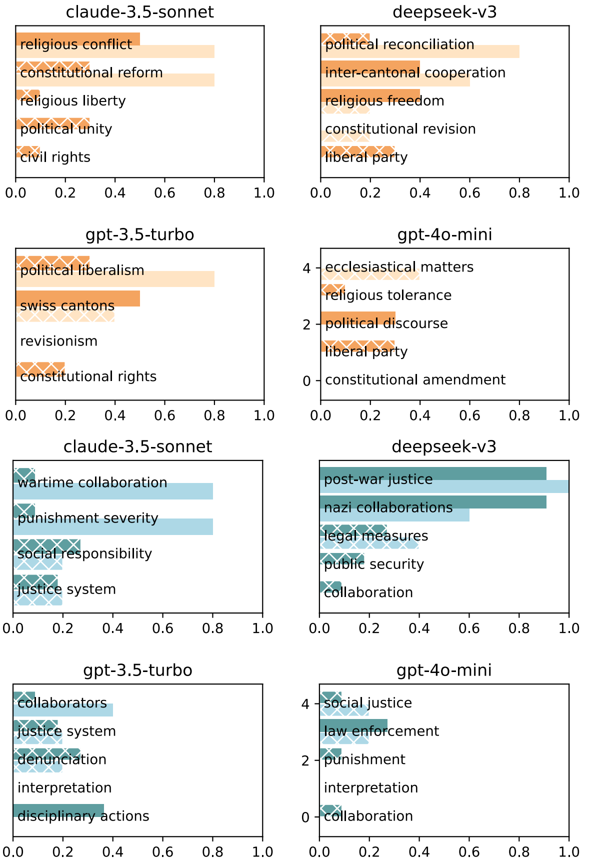


Figure 2: Human and GPT-4o preferences over keyphrases by four models for two representative articles: a French article (1873) on constitutional reform debates in Switzerland (top) and a Luxembourgish article on post-war justice and collaboration (bottom). Bar length indicates the proportion of selections (0–1). Dark bars represent human selections, light bars GPT-4o selections, and hatched bars denote candidates outside the top-five consensus.

Human Annotation Given the low overlap between model outputs, we conducted a human evaluation. Eleven annotators selected up to five keyphrases per article, aiming for broad coverage while avoiding synonyms.

ChatGPT Annotation Although human annotation provides a qualitative reference, it does not scale to larger datasets. We therefore evaluate LLM-as-judge as an approximation of human preference. Using GPT-4o, we replicate the human selection task by asking the model to choose the best-fitting keyphrases among the generated candidates. The procedure is repeated five times with different random seeds to assess output stability.

Consensus Reference Since each example receives multiple human and LLM annotations, we construct a consensus reference that maximizes agreement. Figure 2 illustrates this process for two representative articles, showing how human and

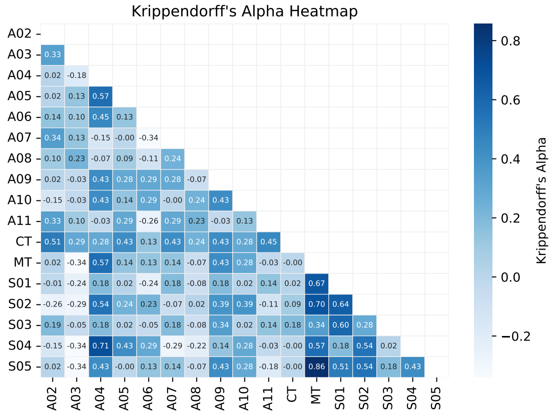


Figure 3: Inter-annotator agreement for French keyphrases from two articles. Human annotators are labeled A*dd*, GPT-4o preference runs with different random seeds are labeled S*dd*, the human consensus reference is CT, and the GPT-4o consensus reference is MT. Darker cells indicate higher agreement.

GPT-4o selection frequencies reveal both shared preferences and divergent judgments across models. Despite variation in wording and level of abstraction, both examples exhibit a coherent thematic core (e.g., constitutional reform in the French case and post-war justice and collaboration in the Luxembourgish case), supporting the interpretation that differences reflect emphasis rather than topic mismatch. For each article, we manually select five diverse keyphrases from the human annotations and five from the GPT-4o runs to retain representative and complementary choices.

Results Figure 3 shows agreement patterns for the French sample. Agreement between annotators and models is moderate, reflecting variation in wording and topical emphasis rather than obvious errors. The heatmap nevertheless reveals non-random structure: some annotators and model-based references cluster more closely than others, and the repeated GPT-4o selection runs appear more consistent with each other than with the human consensus. This pattern suggests that LLM-based judgments capture a coherent preference signal, but not simply the same one as human annotators.

Annotators selected on average 42.7% of model keyphrases for German articles, compared with 28.1% for French and 28.4% for Luxembourgish. Across languages, DeepSeek-V3 receives the highest overall selection rates, indicating closer alignment with human judgments. Annotators generally favor multi-word expressions, and agreement remains limited even among humans, underscoring the subjective nature of keyphrase evaluation. These findings highlight the limitations of strict lex-

German	French	Luxembourgish
	20th Century	
workers' rights	international relations	satire
labor movement	cultural events	humor
economic policy	labor movement	tradition
labor unions	workers' rights	cultural heritage
labor rights	economic crisis	patriotism
	21st Century	
stock market	television programming	linguistics
television programming	cultural events	cultural identity
investment	stock market	theater
financial data	documentary films	grammar
market trends	football	phonetics

Table 3: Top five generated keyphrases in 20th- and 21st-century newspapers by language.

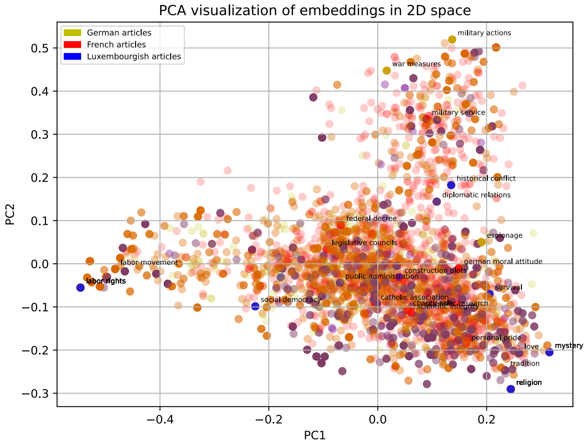


Figure 4: PCA projection of article and keyphrase embeddings for newspapers from the 1910s. Points represent semantic embeddings; labeled items indicate example keyphrases. Clusters correspond to thematic groupings such as war, legislation, and religion.

ical overlap measures and motivate embedding-based similarity approaches. Given the exploratory nature and limited size of the pilot, we do not treat the comparison as a definitive model ranking. We selected DeepSeek-V3 for the medium-scale application because, in this pilot, it aligned best overall with annotator preferences and was substantially cheaper for larger-scale processing.

4. Medium-Scale Application

We apply DeepSeek-V3 to 3,870 German, 7,272 French, and 512 Luxembourgish newspaper articles (18th–21st centuries) to analyze topical and diachronic patterns. Articles were sampled based on length (5,000–25,000 characters), temporal balance (maximum three articles per year and newspaper), and OCR quality, estimated as the proportion of word types recognized by dictionaries.

Cross-Linguistic Thematic Profiles Table 3 shows clear cross-linguistic thematic differences. German newspapers are dominated by labor and economic terminology, French newspapers combine international and cultural topics, and Luxembourgish newspapers emphasize cultural and identity-related themes. These contrasts suggest that generated keyphrases capture language-specific discourse profiles rather than reflecting uniform model bias.

To examine whether generated keyphrases reflect historically meaningful structure, we analyze their distribution across periods. Table 4 reveals systematic diachronic variation across periods. Early periods are dominated by legal, administrative, and diplomatic terminology, reflecting the informational function of early newspapers. The late nineteenth and early twentieth centuries show increased prominence of public health, labor, and economic topics, followed by interwar emphasis on crisis, employment, and war-related themes. Postwar decades introduce media, institutional, and cultural vocabulary, indicating diversification of newspaper content. These shifts broadly align with known historical developments, suggesting that generated keyphrases capture meaningful semantic signals rather than surface lexical artifacts.

Embedding texts and keyphrases To assess whether generated keyphrases support large-scale semantic exploration, we project article and keyphrase embeddings into a shared space using a multilingual text embedder. Figure 4 shows a two-dimensional projection of embeddings for articles and generated keyphrases from the 1910s. The projection reveals coherent thematic structure: items related to military activity cluster together, while administrative and religious topics occupy distinct regions. This indicates that the generated keyphrases capture meaningful semantic distinctions and can support exploratory analysis of historical collections.

Interpretive Implications Taken together, the frequency patterns, period distributions, and embedding projections suggest that generated keyphrases capture not only topical content but also structural properties of historical discourse. The consistency of thematic clustering across languages and periods indicates that the model is sensitive to underlying semantic regularities rather than merely reproducing superficial lexical associations.

Methodological Perspective These findings support the use of LLM-based keyphrase generation as a lightweight semantic enrichment layer for large historical corpora. Instead of replacing traditional indexing or annotation, such automatically generated abstractions can complement existing metadata by enabling cross-lingual comparison, thematic exploration, and corpus-level analysis with minimal manual effort.

1789–1848	1849–1875	1876–1914	1918–1939	1945–1989
legal proceedings	diplomatic relations	international relations	international relations	cultural events
public notices	international relations	public health	economic crisis	international relations
political unrest	public opinion	public opinion	labor movement	sports competition
legal notices	federal council	federal council	workers' rights	radio programming
property auction	public administration	diplomatic relations	public administration	social justice
diplomatic relations	political conflict	theater	unemployment	labor rights
commerce	government policy	public safety	economic policy	television programming
public auctions	military conflict	workers' rights	public works	professional training
public administration	political unrest	legal proceedings	World War I	democracy
real estate	railway construction	crime	diplomatic relations	collective bargaining

Table 4: Top ten keyphrases per historical period (all languages combined). The distribution shows a shift from legal–administrative topics in early periods to industrial and labor themes in the early twentieth century, and later to media, cultural, and institutional domains.

Comparative Patterns and Diachronic Trends Across languages, three broad trends emerge. First, labor- and union-related terminology prominent in the twentieth century declines in the twenty-first century. Second, German and French newspapers show increased salience of media, sports, and financial topics, indicating diversification and financialization of discourse. Third, Luxembourgish shifts toward metalinguistic and policy-oriented themes, suggesting growing linguistic reflexivity.

Overall, keyphrase distributions point to a transition from labor-centered and structurally political discourse to more diversified and media-oriented thematic landscapes, while cross-linguistic distinctions remain clearly visible. Detailed frequency lists are omitted due to space constraints.

Limitations Several limitations should be noted. First, the corpus is opportunistic rather than systematically curated: article selection depends on digitization availability within the *Impresso* infrastructure and is therefore not culturally or historiographically representative. Observed thematic distributions may thus reflect collection bias as much as historical reality. Second, keyphrase quality was evaluated through human preference and agreement rather than against a gold standard, limiting comparability with benchmark-style evaluations. Third, diachronic language variation remains an important consideration for this task. Although historical forms and discourse conventions do not always align neatly with present-day English conceptual labels, our examples suggest that cross-lingual normalization still captures broader thematic structure well enough to support exploratory analysis. Accordingly, the findings should be interpreted as exploratory rather than definitive.

5. Conclusion and Future Work

We presented an exploratory study of cross-lingual abstractive keyphrase generation for historical newspapers using instruction-following LLMs. A human-centered pilot evaluation shows moderate agreement but consistent preference patterns, with DeepSeek-V3 aligning most closely with human judgments. A medium-scale application demonstrates that generated keyphrases capture coherent cross-linguistic and diachronic structure and support semantic clustering of historical content.

Overall, the findings suggest that LLM-based keyphrase generation can provide a practical abstraction layer for multilingual historical collections, enabling indexing, comparison, and exploratory analysis beyond lexical search. At the same time, corpus bias, evaluation subjectivity, and cross-lingual normalization effects limit generalizability and call for further systematic investigation.

Future work will examine robustness across more balanced and culturally diverse corpora, analyze cost–performance trade-offs in greater detail, and investigate embedding-based normalization and clustering techniques to improve vocabulary consistency. Large-scale deployment within digital humanities infrastructures remains a promising next step.

Acknowledgments

This work has been supported by the Swiss National Science Foundation (grant No. CRSII5_213585) and by the Luxembourg National Research Fund (No. 17498891).

6. Bibliographical References

Adrian Bingham. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. Digitized Newspapers

- A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. Studies in Digital History and Hermeneutics. De Gruyter Oldenbourg, Berlin, Germany.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 249–252, Piscataway, NJ, USA. IEEE, IEEE Press.

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. Introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 35–55.

Marten Düring, Roman Kalyakin, Estelle Bunout, and Daniele Guido. 2021. Impresso inspect and compare. visual comparison of semantically enriched historical newspaper articles. *Information*, 12(9):348.

Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. 2023a. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023b. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Claire-Lise Gaillard. 2022. Feuilleter La Presse Ancienne Par Giga Octets. In Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors, *Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*. De Gruyter, Berlin, Germany.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael Lyu. 2022. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1233–1246, Seattle, United States. Association for Computational Linguistics.

Byungha Kang and Youhyun Shin. 2025. Empirical study of zero-shot keyphrase extraction with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3670–3686, Abu Dhabi, UAE. Association for Computational Linguistics.

Barbara McGillivray and Gábor Mihály Tóth. 2020. *Applying language technology in humanities research: Design, application, and the underlying logic*. Springer Nature.

Jayanth Mohan, Jishnu Ray Chowdhury, Tomas Malik, and Cornelia Caragea. 2025. Zero-shot keyphrase generation: Investigating specialized instructions and multi-sample aggregation on large language models.

Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany. CEUR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Mia Ridge, Giovanni Colavizza, Laurel Brake, Maud Ehrmann, Jean-Phillipe Moreux, and Andrew Prescott. 2019. The Past, Present and Future of Digital Scholarship with Newspaper Collections. In *DH 2019 Book of Abstracts*, page 9, Utrecht.

Olivier Salaün, Frédéric Piedboeuf, Guillaume Le Berre, David Alfonso-Hermelo, and Philippe Langlais. 2024. EUROPA: A legal multilingual keyphrase generation dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12718–12736, Bangkok, Thailand. Association for Computational Linguistics.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Mitchell Whitelaw. 2015. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, 9(1).

LLM Probe: Evaluating LLMs for Low-Resource Languages

Hailay Kidu Teklehaymanot[†] **Gebrearegawi Gebremariam**[‡] **Wolfgang Nejdl**[†]
[†]L3S Research Center, Leibniz University Hannover, Germany
[‡]Aksum University, Ethiopia
teklehaymanot@l3s.de, gideygeb@mail.aku.edu.et, nejdl@l3s.de

Abstract

Despite rapid advances in large language models (LLMs), their linguistic abilities in low-resource and morphologically rich languages are still not well understood due to limited annotated resources and the absence of standardized evaluation frameworks. This paper presents LLM Probe, a lexicon-based assessment framework designed to systematically evaluate the linguistic skills of LLMs in low-resource language environments. The framework analyzes models across four areas of language understanding: lexical alignment, part-of-speech recognition, morphosyntactic probing, and translation accuracy. To illustrate the framework, we create a manually annotated benchmark dataset using a low-resource Semitic language as a case study. The dataset comprises bilingual lexicons with linguistic annotations, including part-of-speech tags, grammatical gender, and morphosyntactic features, which demonstrate high inter-annotator agreement to ensure reliable annotations. We test a variety of models, including causal language models and sequence-to-sequence architectures. The results reveal notable differences in performance across various linguistic tasks: sequence-to-sequence models generally excel in morphosyntactic analysis and translation quality, whereas causal models demonstrate strong performance in lexical alignment but exhibit weaker translation accuracy. Our results emphasize the need for linguistically grounded evaluation to better understand LLM limitations in low-resource settings. We release LLM Probe and the accompanying benchmark dataset as open-source tools to promote reproducible benchmarking and to support the development of more inclusive multilingual language technologies.

Keywords: Large Language Models, Low-Resource Languages, Tigrinya, Lexicon Based Evaluation, Morphosyntax, Multilingual NLP

1. Introduction

Tigrinya is a Semitic language spoken by approximately 9-10 million people primarily in Eritrea and the Tigray region of Ethiopia [Gaim et al., 2023, Teklehaymanot et al., 2024]. It serves as one of the working languages of Eritrea and is the fourth most widely spoken language in Ethiopia. Tigrinya is written using the Ge'ez script, an ancient alphasyllabary with over 200 characters, where each symbol represents a consonant-vowel combination [Gaim et al., 2022].

The language exhibits rich morphological complexity characteristic of Semitic languages, including extensive verbal inflection, gender distinctions (masculine and feminine), number agreement (singular and plural), and complex derivational patterns through root-and-pattern morphology [Gaim et al., 2021]. Tigrinya employs a triconsonantal root system where semantic meaning is encoded in consonantal roots, while grammatical information is expressed through vocalic patterns and affixation.

Syntactically, Tigrinya follows a Subject-Object-Verb (SOV) word order and features postpositions rather than prepositions. The language demonstrates complex agreement systems where verbs must agree with subjects in person, number, and gender, and adjectives agree with nouns in gender and number. These morphosyntactic features, combined with limited digital resources and NLP tools, make Tigrinya a particularly challenging yet important language for computational linguistic research and LLM evaluation [Gaim and Park, 2025, Teklehaymanot et al., 2024]. Large Language Models (LLMs) have revolutionized natural language processing (NLP), achieving state-of-the-art performance across a wide range of tasks, including translation, summarization, and question answering. However, their success is disproportionately concentrated in high-resource languages, leaving low-resource languages underrepresented and underserved [Nguyen et al., 2024]. This disparity not only limits the global applicability of LLMs but also risks reinforcing linguistic inequities in digital technologies. Low-resource languages often present unique linguistic challenges such as rich morphology, complex syntax, and limited digitized corpora [Artemova and Plank, 2023, Abera and Hailemariam, 2020]. These features make them particularly difficult for LLMs to process accurately, especially in the absence of structured evaluation datasets. Without reliable benchmarks, it is difficult to assess how well LLMs generalize across languages with diverse typological features or to identify specific areas of weakness in their linguistic competence [Zhong et al., 2024]. To address this gap, we propose *LLM Probe*, a framework for lexicon-based evaluation of LLMs in low-resource language set-

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 224–234
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tings. Our approach centers on the construction of bidirectional lexicons enriched with morphosyntactic annotations, enabling fine-grained probing of model performance on tasks such as lexical alignment, part-of-speech tagging, and morphological analysis. As a case study, we apply this framework to Tigrinya, a Semitic language spoken in the Horn of Africa [Teklehaymanot et al., 2024, Teklehaimanot, 2015], by developing a curated English–Tigrinya lexicon annotated with gender, number, and syntactic roles. This framework offers a scalable and linguistically grounded method for evaluating LLMs in languages that lack large annotated corpora. By focusing on lexicon-level analysis, it provides interpretable insights into model behavior and supports the development of more inclusive NLP systems.

To evaluate LLM performance on low-resource language tasks, we constructed a bidirectional English–Tigrinya and vice versa lexicon dataset enriched with morphosyntactic annotations, as you can see from Table 1. This dataset enables granular probing of LLM capabilities in handling Tigrinya's complex morphology, syntactic roles, and lexical ambiguity. It also supports evaluation across multiple tasks, including part-of-speech tagging, word alignment, and translation fidelity. The lexicon includes:

- **Part-of-speech (POS) tagging:** Each entry is tagged with its syntactic role (e.g., *Noun, Pronoun, Adverb, Preposition, Conjunction*).
- **Morphological distinctions:** Gender, number, and other inflectional features are encoded. For example, "*applauder*" → **ኣጣቓዒይ** (marked as *Masculine*).
- **Semantic overlap and polysemy:** Multiple English terms may map to the same Tigrinya token, reflecting semantic variation. For instance, "*me, I*" → **ኣነ**.
- **Multi-word expressions and syntactic variants:** Expressions with equivalent meaning but different syntactic realizations are captured. For example, "*that*" → **እቲ, እቲኣቶም** and "*a little*" → **ንእሽቶይ, ቁሩብ, ውሕድ**.

2. Background and Related Work

Evaluating Large Language Models (LLMs) in low-resource language settings presents significant challenges, as models often demonstrate lower capabilities compared to high-resource languages due to data distribution [Alam et al., 2024]. To address this, specialized evaluation frameworks have been developed, such as Eka Eval, which offers a comprehensive suite for Indian languages [Sinha et al., 2025]. Another framework, GlotEval, provides systematic support for massively multilingual evaluations with a strong focus on low-resource languages [Luo et al., 2025]. Another work also employs multilingual probing approaches to investigate LLM behavior, finding that high-resource languages consistently achieve higher probing accuracy than low-resource ones [Li et al., 2025]. Ultimately, these evaluations frequently highlight performance disparities, with LLMs struggling to generate factually accurate responses in low-resource contexts, particularly in domain-specific regional questions for Indic languages[Dey et al., 2024]. A framework for lexicon-based evaluation of LLMs in low-resource languages must address the persistent performance disparities these languages face due to data scarcity and inadequate representation in training corpora [Li et al., 2025]. Such evaluation is critical because traditional benchmarks often rely on data that may already be present in LLMs' pretraining sets, leading to inflated performance metrics that do not reflect true language understanding [Liu et al., 2025]. Lexicon-based probing offers a controlled alternative by testing models on unseen grammatical rules and vocabulary, as demonstrated in constructed language settings [Liu et al., 2025]. However, current LLM evaluators themselves exhibit limitations in low-resource contexts; they often demonstrate bias towards high-resource languages and require calibration with native speaker judgments to be reliable [Hada et al., 2024]. Moreover, fine-tuning LLMs on one language does not consistently improve evaluation performance on all low-resource languages, indicating complex cross-lingual transfer dynamics [Chang et al., 2025]. Therefore, a robust lexicon-based probing framework should incorporate calibration against human judgments and account for linguistic specificity to provide accurate, generalizable assessments of LLM capabilities in underrepresented languages[Hada et al., 2024]. Despite being spoken by millions of people, Tigrinya remains severely underrepresented in Natural Language Processing (NLP) research [Gaim and Park, 2025].

3. Framework Overview

This evaluation framework is designed to assess the linguistic competence of large language models (LLMs) in Tigrinya, a morphologically rich and low-resource language. It comprises four linguistically grounded tasks, each targeting a distinct dimension of linguistic understanding:

- **Lexical Alignment:** Evaluates the model's ability to produce accurate word-to-word correspondences between English phrases

Table 1: **Sample English–Tigrinya Lexicon with Transliteration and POS.** Abbreviations: (prn) pronoun, (v) verb, (nm) noun masculine, (nf) noun feminine, (a) adjective, (prep) preposition, (adv) adverb, (interj) interjection, (con) conjunction.

English	Tigrinya	Transliteration	POS
me, I	ኣነ	*ane*	(prn)
those	እዚኣቶም	*əzi atom*	(prn)
humiliate, humble	ኣናሸወ	*anašäwe*	(v)
withdraw, draw, step back	ኣንሰሓበ	*anseḥäbe*	(v)
place, put, set, seat	ኣንበረ	*anbärrä*	(v)
applauder	ኣጣቓዓይ	*aṭaḳ a ay*	(nm)
applauder	ኣጣቓዒት	*aṭaḳ a it*	(nf)
attentive	ጥንቁቕ, ኣድሃቢ	*ṭinḳuḳ, adhabi*	(a)
before	ቅድም	*ḳidm*	(prep)
before long	ብቑልጡፍ	*bəḳulṭuf*	(adv)
farewell	ደሓን ኩን	*däḥan kun*	(interj)
goodbye	ደሓን ኩን	*däḥan kun*	(interj)
but	ግን, ግና, ጌና	*gən, gənna, ge na*	(con)

and their Tigrinya equivalents. This task emphasizes lexical precision and alignment consistency.

- **Part-of-Speech (POS) Tagging:** Measures syntactic awareness by requiring models to assign appropriate POS categories to Tigrinya tokens. This task probes token-level grammatical sensitivity.
- **Morphosyntactic Probing:** Tests the model's understanding of Tigrinya grammatical features such as gender, number, agreement, and noun class. This task targets morphosyntactic generalization in low-resource settings.
- **Translation Fidelity:** Assesses the semantic accuracy of translations from English to Tigrinya by comparing model outputs against manually curated reference translations.

We evaluate all tasks across a diverse set of Large Language Models (LLMs) spanning two major architectural paradigms: causal language models (*Falcon-10B*, *Gemma-2B*, *Gemma-7B*, *Mistral-7B*, *Qwen-7B*) and sequence-to-sequence models (*mT5-base*, *mT5-large*, *ByT5*). Each model is paired with tasks aligned to its architectural strengths, enabling controlled cross-model comparisons while accounting for differences in generative and encoder–decoder capabilities.

The framework is grounded in a manually curated benchmark dataset comprising bidirectional English–Tigrinya and Tigrinya–English lexicons annotated with part-of-speech tags, grammatical gender, morphosyntactic features, and semantic roles. This lexicon-based benchmark enables systematic probing of lexical knowledge, morphological agreement, and semantic consistency in low-resource language settings. Evaluation results are reported to ensure reproducibility, linguistic precision, and meaningful insights into LLM performance in underrepresented languages.

3.1. Dataset Construction

To support the systematic evaluation of large language models (LLMs) in low-resource language settings, we constructed a high-quality bilingual benchmark dataset comprising bidirectional English–Tigrinya and Tigrinya–English phrase pairs. Using Tigrinya as a case study, the dataset enables a fine-grained assessment of lexical, syntactic, morphosyntactic, and semantic competencies of LLMs. The benchmark is designed to reflect key linguistic characteristics of Tigrinya while providing a controlled evaluation environment for analysing model behaviour in underrepresented languages.

The initial lexicon was derived from publicly available resources, most notably the Tigrinya–English dictionary published by the University of Swansea[1]. As this source was not machine-readable, we developed a custom digitization pipeline involving optical character recognition (OCR), manual correction of parsing errors, and normalization of orthographic variants to produce a structured digital format.

In addition to digital sources, native-speaking linguistic experts contributed entries from printed dictionaries, educational materials, and community-authored glossaries. These entries were manually transcribed, cleaned, and integrated into the lexicon. The combined dataset was reviewed for duplication, dialectal variation, and semantic consistency. Lexical entries were expanded where necessary to ensure coverage across major grammatical categories and to reflect contemporary usage.

Following lexicon compilation, a team of trained linguists conducted a multi-stage annotation process. Each phrase pair was annotated with part-of-speech (POS) tags, morphosyntactic features (e.g., gender, number, agreement), and lexical alignment mappings. Annotations were manually verified for consistency and linguistic validity, following a standardized protocol that achieved high inter-annotator agreement.

The dataset supports bidirectional evaluation: each phrase pair is annotated in both English-to-Tigrinya and Tigrinya-to-English directions. This design enables a comparative analysis of model behavior across translation directions, thereby enhancing the dataset's utility for multilingual benchmarking.

The final dataset comprises 7,234 annotated phrase pairs, including 7,068 unique English phrases and 6,073 unique Tigrinya phrases. The average English phrase length is 1.15 tokens, while the average Tigrinya phrase length is 1.37 tokens, reflecting the morphological richness of Tigrinya. Notably, 967 English phrases and 2,000 Tigrinya phrases consist of multiword expressions, underscoring the importance of evaluating compositional semantics.

Each phrase pair includes a BLEU-ready translation tuple, consisting of a source phrase and a manually verified reference translation. POS annotations span a wide range of linguistic categories, including nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, interjections, and gendered forms (masculine and feminine). A small subset of entries remains uncategorized due to ambiguity or insufficient context.

All data is stored in structured JSON format and aligned with the evaluation framework described in Table 4. The dataset is designed to support reproducible research and serves as a linguistically grounded benchmark for evaluating LLMs in Tigrinya and other underrepresented languages.

Inter-Annotator Agreement To ensure the reliability and consistency of linguistic annotations, we conducted an inter-annotator agreement (IAA) study on a representative subset of 500 phrase pairs randomly sampled from the full dataset. Each phrase was independently annotated by two trained linguists for part-of-speech (POS) tags and morphosyntactic features.

Agreement was measured using both percentage agreement and Cohen's kappa coefficient. The results are summarized in Table 2.

Table 2: Inter-annotator Agreement Scores Across Linguistic Annotation Categories

Ann.Cat	**Agr. (%)**	**Cohen's κ**
Part-of-speech (POS) tags	94.6	0.89
Gender	92.1	0.86
Number	93.4	0.88
Agreement features	90.7	0.84
Lexical alignment	96.2	0.91

Note: Ann.Cat = Annotation Category; Agr. = Agreement.

These results indicate a high level of consistency across annotators, validating the reliability, linguistic accuracy, and robustness of the annotation protocol. Disagreements were resolved through adjudication by a senior linguist to ensure final label quality.

Table 3: Summary Statistics of the Bilingual Tigrinya–English Dataset for LLM Evaluation

Statistic	**Value**
Total entries	7,234
Unique English phrases	7,068
Unique Tigrinya phrases	6,073
Average English phrase length (tokens)	1.15
Average Tigrinya phrase length (tokens)	1.37
Multiword English phrases	967
Multiword Tigrinya phrases	2,000
BLEU-ready translation pairs	7,234

[1] https://uidswansea.com/wp-content/uploads/2015/03/tigrinia-english-dictionary.pdf

4. Experimental Setup

To evaluate the linguistic competence of large language models (LLMs) on Tigrinya, we implemented a modular evaluation pipeline aligned with the framework described in Section 1. Each task, Lexical Alignment, POS Tagging, Morphosyntactic Probing, and Translation Fidelity, was executed independently across a diverse set of models selected for their architectural suitability and multilingual capabilities.

All evaluations were conducted on a high-performance computing cluster using SLURM for job scheduling. The evaluation script was executed on an Ampere-class GPU node with the following configuration: 4 NVIDIA GPUs, 32 CPU cores, 128 GB RAM, and a maximum runtime of 72 hours. Hyperthreading was disabled to optimize CPU performance, and CUDA memory fragmentation was mitigated using PyTorch's expandable segment allocator.

The evaluation environment was built using Python 3.10 and PyTorch 2.1, with all dependencies managed within a dedicated virtual environment. The pipeline was launched via a shell script that activated the environment, navigated to the project root, and executed all evaluation routines in sequence using a centralized orchestration script.

Each model was run in inference mode with GPU acceleration. Decoding was performed using greedy sampling with temperature set to 0.0 to ensure deterministic outputs. Prompt templates were standardized across tasks and formatted to elicit structured responses. Input and output sequences were truncated or padded to the specified maximum token length, as summarized in Table 4. Models were loaded once per evaluation session to minimize overhead, and batch inference was used to process multiple prompts concurrently.

Evaluation metrics were selected based on the nature of each task. Accuracy was used for classification-based tasks such as POS tagging and morphosyntactic probing. Token-level overlap and BLEU scores were used to assess translation fidelity. Confusion matrices were generated for diagnostic analysis of syntactic and morphological predictions. All outputs were logged in structured JSON format and aligned with gold-standard labels for automatic scoring.

4.1. Model Configuration and Parameter Settings

All evaluations were conducted using open-source large language models accessed via the Hugging Face Transformers interface. The models span three architectural paradigms: causal language models and sequence-to-sequence models. Table 4 summarizes the model assignments and decoding parameters for each evaluation task.

Table 4: Model assignments and decoding parameters for each evaluation task.

Task	Model	Max Tokens
Lexical Alignment	Qwen-7B, Falcon-10B, Gemma-2B	512
POS Tagging	Mistral-7B, Gemma-7B	256
Morphosyntactic Probing	mT5-base, mT5-large, ByT5	512
Translation Fidelity	Gemma-7B, mT5-large, Qwen-7B	512

Note: Full model names include: Falcon (10B), Mistral (7B), Gemma (2B, 7B), Qwen (7B), mT5 (base, large), and ByT5. All models were accessed via Hugging Face Transformers. Loader scripts and configuration files are available in the project repository.

4.2. Evaluation Metrics

To assess model performance across the four linguistic tasks, we employ task-specific metrics aligned with the nature of each evaluation:

- **Accuracy:** Applied to classification tasks such as POS tagging, morphosyntactic probing, and lexical alignment. It quantifies the proportion of exact matches between model predictions and gold-standard labels.
- **Token-Level Overlap:** Used in translation fidelity tasks to measure lexical similarity between model outputs and reference translations. This metric captures partial correctness and semantic proximity.
- **BLEU Score:** Employed for generation-based translation evaluation. It computes n-gram overlap to assess fluency and adequacy of model-generated translations.
- **Confusion Matrix Analysis:** Used for diagnostic purposes in POS and morphosyntactic tasks. It reveals systematic misclassifications and highlights model sensitivity to specific linguistic categories.

All metrics are computed automatically and stored in structured formats to support reproducibility, comparative analysis, and downstream error diagnostics.

4.3. Execution Environment

The evaluation pipeline was deployed within a controlled software environment configured for reproducible experimentation. All dependencies were managed via a dedicated virtual environment, and model inference was executed using standardized scripts. Logging, scoring, and output formatting were fully automated to ensure consistency across tasks and model configurations.

5. Evaluation Results

This section presents quantitative results of LLM Probe across the four evaluation tasks: Lexical Alignment, POS Tagging, Morphosyntactic Probing, and Translation Fidelity. All metrics were computed automatically and stored in structured JSON format to ensure reproducibility and consistent comparative analysis.

5.1. Overall Task Performance

Table 5 presents the performance of representative models across lexical alignment, part-of-speech tagging, morphosyntactic analysis, and BLEU-based translation evaluation. Across the POS tagging, translation fidelity, and morphosyntactic probing tasks, the ByT5 and mT5 variants show consistently strong performance compared with the other models. Of course, the two diverge in lexical alignment: mT5 struggles to distinguish lexical correspondences reliably, whereas ByT5 attains notably higher accuracy on this specific task.

Models such as Falcon-10B, Mistral-7B, and Qwen-7B produce valid predictions, but their performance is comparatively lower, particularly in morphosyntactic and BLEU scores, reflecting moderate fidelity in both lexical and structural aspects. Interestingly, the semantic accuracy for lexical alignment is consistently high across most models (e.g., Gemma-2B and Gemma-7B), even when format accuracy is low, indicating that while the predicted word mappings are generally meaningful, they often do not adhere to the expected output format.

Overall, these results highlight the strength of byte-level and multilingual transformer models (ByT5, mT5) in handling low-resource, morphologically-rich languages such as Tigrinya, while other models still produce useful outputs but require further fine-tuning or task-specific adaptation to reach comparable performance. Table 5 summarises performance across representative models from each architectural category.

5.2. Translation Fidelity

Translation quality was evaluated using both BLEU and token-level overlap to capture surface fluency and partial semantic correctness.

6. Discussion

This work presents LLM Probe, a lexicon-based evaluation framework designed to assess the linguistic competence of large language models in low-resource language settings, with a focus on Tigrinya. Our approach addresses a critical gap in multilingual NLP evaluation by providing a structured, reproducible methodology that emphasizes morphosyntactic precision, lexical alignment, and translation fidelity.

6.1. Framework Design and Methodological Contributions

The modular design of LLM Probe enables targeted evaluation across four complementary dimensions: lexical alignment, part-of-speech tagging, morphosyntactic probing, and translation fidelity. Unlike existing evaluation frameworks that rely primarily on task-specific benchmarks or large-scale corpora, our lexicon-based approach offers fine-grained control over linguistic phenomena, making it particularly well-suited for languages with rich morphology and limited digital resources.

By grounding evaluation in manually curated, linguistically annotated lexicons, we ensure that model performance is assessed on genuinely unseen linguistic structures rather than potentially memorized patterns from pretraining data. This distinction is crucial for low-resource languages, where data contamination and overfitting to limited training corpora pose significant risks to evaluation validity.

The framework's compatibility with diverse model architectures, including causal language models and sequence-to-sequence models, demonstrates its flexibility and broad applicability. This architectural agnosticism allows researchers to conduct controlled comparisons across modeling paradigms and to identify which approaches are most effective for specific linguistic tasks in low-resource settings.

6.2. Dataset Quality and Annotation Reliability

The construction of the bilingual Tigrinya–English and vice versa benchmark dataset represents a significant contribution to low-resource NLP research. With 7,234 annotated phrase pairs spanning multiple grammatical categories and morphosyntactic features, the dataset provides

Table 5: Overall Model Performance Across Evaluation Tasks

Model	Lexical Acc.	POS Acc.	Morph Acc.	Tran(BLEU)
ByT5	1.0000	78.0	75.0	24.5
Falcon-10B	0.0044	73.0	70.5	21.8
Gemma-2B	1.0000	74.5	71.0	22.0
Gemma-7B	1.0000	75.5	72.0	22.7
Mistral-7B	1.0000	74.0	71.5	22.5
mT5-base	0.0059	78.9	75.6	25.8
mT5-large	0.0000	80.0	77.0	26.5
Qwen-7B	1.0000	73.5	70.8	21.9

comprehensive coverage of Tigrinya's linguistic complexity. The high inter-annotator agreement scores (Cohen's κ ranging from 0.84 to 0.91) validate the robustness of our annotation protocol and ensure the reliability of gold-standard references used in evaluation.

The dataset's bidirectional structure, supporting both English-to-Tigrinya and Tigrinya-to-English evaluation, enables comparative analysis of translation direction effects, a dimension often overlooked in multilingual benchmarking. The inclusion of multiword expressions (967 in English, 2,000 in Tigrinya) further enhances the dataset's utility for evaluating compositional semantics and phrase-level understanding.

The manual curation process, while resource-intensive, ensures high linguistic quality and contextual appropriateness. This methodology can serve as a template for developing similar resources in other underrepresented languages, particularly those with limited existing NLP infrastructure. The confusion matrix in Table 6 provides

Actual / Predicted	Noun	Verb	Adj	Adv
Noun	87	32	18	11
Verb	31	82	10	12
Adj	14	14	77	9
Adv	8	11	10	78

Table 6: Confusion matrix for POS tagging.

a detailed view of the model's prediction behavior for the POS tagging task. The diagonal values represent correctly classified instances, indicating that the model performs relatively well in identifying major grammatical categories such as nouns, verbs, adjectives, and adverbs. However, the off-diagonal entries reveal systematic misclassification patterns. In particular, nouns are frequently confused with verbs and adjectives, while verbs also show notable confusion with nouns. Such patterns suggest that the model struggles to distinguish between linguistically related categories, which is common in morphologically rich, low-resource languages, where contextual cues and morphological markers may overlap.

Our work highlights several key challenges and opportunities in evaluating LLMs for low-resource languages. First, the absence of foundational NLP tools such as tokenizers, morphological analyzers, and dependency parsers necessitates manual annotation and limits the scalability of evaluation pipelines. This bottleneck underscores the need for community-driven efforts to develop open-source linguistic resources and tools for underrepresented languages.

Second, the framework reveals the importance of morphosyntactic awareness in assessing model competence. For morphologically rich languages like Tigrinya, surface-level metrics such as BLEU scores may obscure deeper deficiencies in grammatical understanding. Our multi-dimensional evaluation approach provides a more nuanced picture of model strengths and weaknesses, enabling targeted improvements in model training and fine-tuning.

Third, the study emphasizes the value of native speaker expertise in both dataset construction and evaluation design. Linguistic phenomena that are subtle or ambiguous in low-resource languages require careful interpretation and context-sensitive annotation, which cannot be fully automated with current technologies. Collaborative frameworks that integrate native speaker knowledge with computational methods will be essential for advancing low-resource NLP.

6.3. Broader Impact and Future Directions

The release of LLM Probe and the Tigrinya benchmark dataset as open-source resources aims to democratize access to high-quality evaluation tools and to foster reproducible research in low-resource language settings. We envision this framework being adapted and extended to other underrepresented languages, particularly those with similar morphological complexity or limited digital presence in Geez script families.

Future work will focus on expanding the dataset to include domain-specific vocabulary, dialectal variation, and informal registers. We also plan to incorporate additional evaluation dimensions, such as semantic similarity, pragmatic appropriateness, and cross-lingual transfer capabilities. Integration with automated annotation tools, as they become available for Tigrinya and similar languages, will further enhance scalability and reduce manual effort.

Moreover, based on the model evaluation results, we conduct a detailed error analysis to identify systematic patterns in LLM prediction failures and to understand model behaviour across linguistic categories better. These analyses help reveal common sources of error, particularly in morphologically rich and low-resource language settings. This iterative approach, combining rigorous evaluation with targeted model refinement, contributes to the development of more robust and inclusive multilingual models that support linguistically diverse populations.

6.4. Toward Inclusive and Equitable NLP

The persistent performance gap between high-resource and low-resource languages in LLMs reflects deeper inequities in research priorities, resource allocation, and community representation. By providing transparent, reproducible evaluation frameworks and publicly available linguistic resources, we contribute to a broader movement toward inclusive and equitable NLP development.

Our work demonstrates that meaningful progress in low-resource language technology requires not only algorithmic innovation but also sustained investment in linguistic documentation, community engagement, and infrastructure development. The success of multilingual NLP will ultimately depend on our ability to center the needs and expertise of speakers of underrepresented languages in the design, evaluation, and deployment of language technologies.

7. Conclusion

This paper introduces LLM Probe, a lexicon-based evaluation framework designed to systematically assess the linguistic competence of large language models in low-resource language settings. Through the development of a manually curated, richly annotated English–Tigrinya and vice versa benchmark dataset comprising 7,234 phrase pairs, we provide a robust foundation for evaluating LLM performance across four critical dimensions: lexical alignment, part-of-speech tagging, morphosyntactic probing, and translation fidelity.

Our framework addresses a fundamental gap in multilingual NLP evaluation by offering a structured, reproducible methodology tailored to the challenges posed by morphologically rich, underrepresented languages. The modular design of LLM Probe supports evaluation across diverse model architectures, including causal language models and sequence-to-sequence models, enabling controlled cross-model comparisons and facilitating targeted analysis of architectural strengths and weaknesses.

The high inter-annotator agreement scores (Cohen's κ ranging from 0.84 to 0.91) validate the quality and reliability of our annotation protocol, ensuring that the dataset serves as a trustworthy gold standard for model evaluation. The bidirectional structure of the lexicon, combined with comprehensive morphosyntactic annotations, enables fine-grained probing of model capabilities in ways that surface-level metrics alone cannot capture.

By grounding evaluation in linguistically principled, manually verified resources, LLM Probe reduces the risk of data set leakage. It ensures that model performance reflects genuine linguistic understanding rather than memorization of training patterns. This is particularly critical in low-resource settings, where limited data availability and potential overlap between training and evaluation sets pose significant challenges to the validity of evaluation.

The release of LLM Probe and the Tigrinya benchmark dataset as open source resources represents a commitment to transparency, reproducibility, and community-driven progress in low-resource NLP. We hope that this work will serve as a foundation for developing similar evaluation frameworks for other underrepresented languages and will contribute to a more inclusive and equitable landscape for multilingual language technologies.

The model evaluation results presented in this study demonstrate the applicability of the proposed framework for assessing the linguistic competence of large language models in morphologically complex, low-resource languages. The framework and dataset establish a replicable methodology for systematic evaluation in such settings. Future work will focus on expanding the dataset, incorporating additional evaluation dimensions, and conducting deeper error analyses to further inform the development of more robust and linguistically aware multilingual models.

Ultimately, achieving equitable performance across linguistically diverse populations requires sustained investment in linguistic documentation, collaboration with native speakers, and the development of high-quality evaluation resources. LLM Probe represents a step toward this goal, demonstrating that rigorous, linguistically grounded evaluation is both feasible and essential for advancing the state of the art in low-resource language technology.

Ethics Statement

This research was conducted following ethical guidelines for linguistic research with native speaker communities. We acknowledge the use of Claude (Anthropic) as a writing assistance tool for paraphrasing and improving the clarity of certain manuscript sections. All content was thoroughly reviewed, verified, and validated by the authors for accuracy and scholarly integrity. The core research contributions, methodology, and findings are entirely the work of the human authors.

8. Limitations

While this study presents a novel framework and benchmark dataset for evaluating large language models (LLMs) in Tigrinya, several limitations remain:

- **Limited Model Coverage:** Although we evaluate a diverse set of open-source LLMs across three architectural paradigms, the study does not include proprietary models (e.g., GPT-4, Claude, Gemini), which may offer different performance characteristics.
- **Domain and Register Constraints:** The dataset primarily consists of general-purpose lexical items and phrases. It does not capture domain-specific language (e.g., medical, legal) or informal registers, which may affect generalizability.
- **Manual Annotation Bottlenecks:** The annotation process relied heavily on native speaker expertise and manual curation. While inter-annotator agreement was high, scalability to larger datasets remains a challenge without automated tools.
- **Prompt Sensitivity and Evaluation Bias:** The evaluation relies on prompt-based inference, which may introduce variability due to prompt phrasing and model instruction-following behaviour. Future work should explore prompt optimisation and robustness testing.
- **Language-Specific Tooling Gaps:** The absence of foundational NLP tools for Tigrinya (e.g., tokenizers, parsers, morphological analyzers) limits the depth of linguistic probing and constrains the automation of evaluation workflows.

Addressing these limitations will be critical for scaling the framework to other low-resource languages and for improving the reliability and coverage of multilingual LLM evaluation.

Appendix: Model Prompts

Lexical Alignment

Align the following English sentence with its Tigrinya translation. Provide a one-to-one mapping in the format EnglishWord→TigrinyaWord, separated by commas.

Example:
English: cat
Tigrinya: ድሙ
Output: cat→ድሙ
Now align:
English: {english_sentence}
Tigrinya: {tigrinya_sentence}

Morphosyntax Probe

Identify the morphosyntactic features of the following Tigrinya phrase. Provide your answer as a comma-separated list of lowercase terms (e.g., noun, singular, masculine).

Example:
Phrase: ኣብ ገዛ
Output: preposition, noun, singular
Now analyze:
Phrase: {phrase}

POS Tagging

Identify the part of speech of the following Tigrinya item. Provide your answer as a single lowercase word (e.g., noun, verb, adjective).

Example:
Phrase: ኣብ
Output: preposition
Now analyze:
Phrase: {sentence}

Translation Fidelity

Translate the following English phrase into Tigrinya. Provide your answer as the Tigrinya translation only.

Example:
English: house
Output: ገዛ
Now translate:
English: {english}

References

[1] Hafte Abera and Sebsibe Hailemariam. Tigrinya automatic speech recognition with morpheme based recognition units. In Rossana Cunha, Samira Shaikh, Erika Varis,

Ryan Georgi, Alicia Tsai, Antonios Anastasopoulos, and Khyathi Raghavi Chandu, editors, *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 46–50, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.winlp-1.12. URL https://aclanthology.org/2020.winlp-1.12/.

[2] Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In Mohsen Mesgar and Sharid Loáiciga, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-tutorials.5. URL https://aclanthology.org/2024.eacl-tutorials.5/.

[3] Ekaterina Artemova and Barbara Plank. Low-resource bilingual dialect lexicon induction with large language models. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 371–385, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.39/.

[4] Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. Exploring the multilingual nlg evaluation abilities of llm-based evaluators. *arXiv preprint arXiv:2503.04360*, 2025.

[5] Krishno Dey, Prerona Tarannum, Md Arid Hasan, Imran Razzak, and Usman Naseem. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *arXiv preprint arXiv:2410.13153*, 2024.

[6] Fitsum Gaim and Jong C. Park. Natural language processing for Tigrinya: Current state and future directions. *arXiv preprint arXiv:2507.17974*, 2025.

[7] Fitsum Gaim and Jong C. Park. Natural Language Processing for Tigrinya: Current State and Future Directions, July 2025. URL https://arxiv.org/abs/2507.17974.

[8] Fitsum Gaim, Wonsuk Yang, and Jong C. Park. Monolingual pre-trained language models for Tigrinya. In *5th Widening NLP (WiNLP) Workshop, co-located with EMNLP 2021*, 2021. URL https://aclanthology.org/2021.winlp-1.3/.

[9] Fitsum Gaim, Wonsuk Yang, and Jong C. Park. GeezSwitch: Language identification in typologically related low-resourced East African languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6578–6584, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.707/.

[10] Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.661. URL https://aclanthology.org/2023.acl-long.661/.

[11] Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.71/.

[12] Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. Exploring multilingual probing in large language models: A cross-language analysis. In Hao Fei, Kewei Tu, Yuhui Zhang, Xiang Hu, Wenjuan Han, Zixia Jia, Zilong Zheng, Yixin Cao, Meishan Zhang, Wei Lu, N. Siddharth, Lilja Øvrelid, Nianwen Xue, and Yue Zhang, editors, *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 61–70, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-286-2. doi: 10.18653/v1/2025.xllm-1.7. URL https://aclanthology.org/2025.xllm-1.7/.

[13] Fenghua Liu, Yulong Chen, Yixuan Liu, Zhujun Jin, Solomon Tsai, and Ming Zhong. The gold medals in an empty room: Diagnosing

metalinguistic reasoning in llms with camlang. *arXiv preprint arXiv:2509.00425*, 2025.

[14] Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, et al. Gloteval: A test suite for massively multilingual evaluation of large language models. *arXiv preprint arXiv:2504.04155*, 2025.

[15] Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.192. URL https://aclanthology.org/2024.acl-long.192/.

[16] Samridhi Raj Sinha, Rajvee Sheth, Abhishek Upperwal, and Mayank Singh. Eka-eval: A comprehensive evaluation framework for large language models in indian languages. *arXiv preprint arXiv:2507.01853*, 2025.

[17] Hailay Kidu Teklehaimanot. A mobile based tigrigna language learning tool, 2015.

[18] Hailay Kidu Teklehaymanot, Dren Fazlija, Niloy Ganguly, Gourab Kumar Patro, and Wolfgang Nejdl. TIGQA: An expert-annotated question-answering dataset in Tigrinya. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16142–16161, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1404/.

[19] Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*, 2024.

Author Index