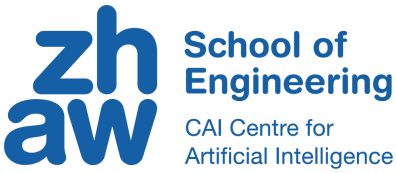
SwissText 2025

Proceedings of the 10th Edition of the Swiss Text Analytics Conference

13–15 May, 2025
Swiss Association for Natural Language Processing
Winterthur, Switzerland

Presenting Organizations



Co-Organizers of Swiss NLP Days

Gold Sponsor

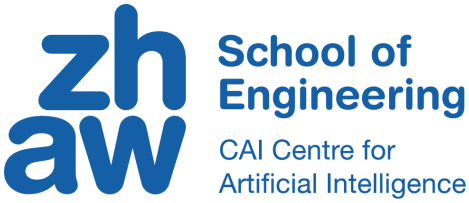
Bronze Sponsors

Partners


unine
Université de Neuchâtel


HSLU Hochschule Luzern


Universität Zürich UZH


zh aw
School of Engineering
CAI Centre for Artificial Intelligence


datalab
www.zhaw.ch/datalab


zh aw
School of Applied Linguistics
ILC Institute of Language Competence


Datalab – The ZHAW Data Science Laboratory


SGAICO


SD:>_
[swissdevjobs.ch]


WIN LINK
INFORMATION AND COMMUNICATION TECHNOLOGY


swiss made software


COMMUNITY PARTNER
NLP ZURICH


alp ict
western switzerland digital cluster

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-950737-19-2

Introduction

In 2025, the SwissText conference celebrated its 10th anniversary! To mark this milestone, we organized a special anniversary edition:

- SwissText 2025 returned to the Zurich University of Applied Sciences (ZHAW) in Winterthur, where SwissText was first held in 2016
- We invited selected Keynote Speakers from previous years to share their (updated!) insights on the current status of NLP
- We extended SwissText with two new events:
  - A "Crash Course" in NLP, held biweekly in 2025, designed to introduce Natural Language Processing to beginners
  - The "NLP Expo 2025", a business fair that welcomed more than 350 participants where companies, startups and research groups showcased their NLP & GenAI solutions

These three events – SwissText, Expo and Crash Course – were combined under the umbrella of the "Swiss NLP Days 2025", jointly organized by SwissNLP and Zurich University of Applied Sciences, with the support of numerous partners, sponsors and co-organizers.

The 10th SwissText conference took place at the Eulachpassage of the Zurich University of Applied Sciences (ZHAW) in Winterthur from 14th to 15th May. Around 100 participants attended keynotes, talks, workshops, and the poster session. Our call for papers asked for contributions in two major tracks: Scientific and Impact. In addition, we introduced a new "Corpora Track", which focused on presenting Swiss datasets. We received a total of 35 submissions, of which we selected 8 for oral presentation, and 23 as posters.

In addition, there were four workshops on the following topics:

- LiRI Corpus Platform
- Battle of NLP Ideas
- Instruction fine-tuning & QLoRA
- LLM Post-Training & DPO

These proceedings include:

- 13 full papers from the Scientific Track
- 3 papers from the Corpora Track
- 16 abstracts from the Impact Track
- Summaries and contributions from the workshops and shared tasks.

We would like to thank our keynote speakers Claudiu Musat (Google DeepMind), Marco Passarotti (Università Cattolica del Sacro Cuore), Iryna Gurevych (Technical University of Darmstadt), Eneko Agirre (University of the Basque Country), Holger Schwenk (Meta AI), and Margot Mieskes (University of Applied Sciences Darmstadt). Their insights and contributions were much appreciated.

We are also deeply grateful to our sponsors and partners who supported us. In addition, we would like to express our sincere gratitude to everyone who has contributed to this conference for their excellent work.

Your sincerely,

Mark Cieliebak, Don Tuggener, and Manuela Hürlimann

Winterthur

October 2025

General Chairs:

Prof. Dr. Mark Cieliebak - Centre for Artificial Intelligence

Program Chairs:

Dr. Don Tuggener - Centre for Artificial Intelligence
Manuela Hürlimann - Centre for Artificial Intelligence

Organizers:

Jasmin Heierli - Institute of Business Information Technology
Jonathan Gerber - Intelligent Information Systems
Jasmin Saxer - Intelligent Information Systems

Program Committee:

Fernando Benites, Pascual Cantos-Gómez, Ernesto William De Luca, Serge Heiden, Valia Kordoni, Brigitte Krenn, Alexander Mehler, Margot Mieskes, Andrei Popescu-Belis, Georg Rehm, Fabio Rinaldi, Tanja Samardžić, Yves Scherrer, Helmut Schmid, Sabine Schulte Im Walde, Manfred Stede, Ludovic Tanguy, Don Tuggener, Juergen Vogel, Manfred Vogel, Egon Werlen, Torsten Zesch

Invited Speakers:

Dr. Claudiu Musat, Google DeepMind

Prof. Dr. Marco Passarotti, Università Cattolica del Sacro Cuore, Milan, Italy

Prof. Dr. Iryna Gurevych, Technical University of Darmstadt (Germany), co-affiliated with Mohamed bin Zayed University of Artificial Intelligence in Abu Dhabi (UAE) and INSAIT in Sofia (Bulgaria)

Prof. Dr. Eneko Agirre, HiTZ center, University of the Basque Country UPV/EHU

Prof. Dr. Holger Schwenk, Meta AI

Prof. Dr. Margot Mieskes, University of Applied Sciences Darmstadt

Invited Talks

Claudiu Musat – "Natural Interactions with Foundation Models"

Marco Passarotti – "Life after ERC. What's New and What's Next in the LiLa Knowledge Base"

Iryna Gurevych – "Fact or Fiction: How to Spot and Debunk Misleading Content?"

Eneko Agirre – "LLMs and low-resource languages"

Holger Schwenk – "Beyond token-based Large Language Models"

Margot Mieskes – "R^3 - Responsible - Replicable - Research"

Table of Contents

Chapter 1

Scientific Track

20min-XD: A Comparable Corpus of Swiss News Articles

Michelle Wastl[1] **Jannis Vamvas**[1] **Selena Calleri**[2] **Rico Sennrich**[1]
[1]Department of Computational Linguistics, University of Zurich [2]20 Minuten (TX Group)
{wastl,vamvas,sennrich}@cl.uzh.ch, {selena.calleri}@20minuten.ch

Abstract

We present *20min-XD* (**20 Min**uten **cross**-lingual **d**ocument-level), a French-German, document-level comparable corpus of news articles, sourced from the Swiss online news outlet *20 Minuten*/*20 minutes*. Our dataset comprises around 15,000 article pairs spanning 2015 to 2024, automatically aligned based on semantic similarity. We detail the data collection process and alignment methodology. Furthermore, we provide a qualitative and quantitative analysis of the corpus. The resulting dataset exhibits a broad spectrum of cross-lingual similarity, ranging from near-translations to loosely related articles, making it valuable for various NLP applications and broad linguistically motivated studies. We publicly release the dataset in document- and sentence-aligned versions and code for the described experiments[1,2].

1 Introduction

Cross-lingual datasets play a crucial role in Natural Language Processing (NLP), supporting a range of tasks such as bitext mining, machine translation, and cross-lingual information retrieval. Among these, comparable corpora—datasets containing text pairs with related but non-identical content across languages—are particularly valuable. Unlike parallel corpora, which consist of direct translations, comparable corpora naturally contain a mix of exact translations, paraphrases, and loosely related content, reflecting the linguistic and cultural variations between languages. This makes them a rich resource for training and evaluating multilingual NLP models (Lewis et al., 2020; Liu et al., 2020; Philippy et al., 2025).

However, existing document-level, cross-lingual corpora remain limited in scope. Many available resources are English-centric, primarily covering English alongside another high-resource language and/or are restricted to sentence-level alignments rather than full documents (Zweigenbaum et al., 2017; Artetxe and Schwenk, 2019). At the same time, large language models (LLMs) and modernized encoder architectures are advancing in their ability to process longer texts and numerous languages, further increasing the demand for multi-/cross-lingual, document-level corpora (Hengle et al., 2024; Wang et al., 2024; Zhang et al., 2024).

Beyond their NLP applications, cross-lingual document-level datasets also facilitate more linguistically motivated studies such as cross-cultural discourse analyses (Carbaugh and Cerulli, 2017) or comparative journalism research (Hanitzsch, 2019). More specifically, a German-French news article corpus could be used to examine how news narratives and framing strategies vary between the Germanophone and Francophone regions.

Given these potential interdisciplinary use cases, we collect comparable news articles in German and French from the online Swiss news outlet *20 Minuten*/*20 minutes*. As both editions are produced by the same publisher, with an internal article transfer workflow from one language to the other, they share a high degree of topical overlap, making them well-suited for comparable corpus creation. Our dataset comprises 15,000 article pairs, spanning nearly a decade (2015–2024). Each article pair consists of a German and a French news article published on the same day, covering the same or a highly related event. In addition to the document-level alignments, we release a sentence-aligned version of the dataset, which contains 117,126 sentences per language.

We release the dataset to the research community for non-commercial, scientific purposes[3].

[1]Dataset: https://huggingface.co/datasets/ZurichNLP/20min-XD

[2]Code: https://github.com/ZurichNLP/20min-XD

[3]See Appendix A for the detailed Copyright notice.

	Validation Set		Full Dataset		Top 15k	
Statistics	**German**	**French**	**German**	**French**	**German**	**French**
Total # of aligned articles	14	14	73,085	73,085	15,000	15,000
Total # of sentences	401	358	1,888,323	1,608,497	357,071	327,628
Total # of tokens	9,087	9,690	43,559,153	43,256,366	8,378,874	8,956,116
Total # of characters	38,523	38,519	189,598,932	174,789,207	36,924,383	36,387,070
Avg. title length in characters	59	54	51	53	51	54
Avg. title length in tokens	18	18	15	17	15	17
Avg. lead length in characters	146	155	152	146	152	150
Avg. lead length in tokens	39	43	39	40	38	41
Avg. content length in characters	2,547	2,542	2,391	2,192	2,258	2,222
Avg. content length in tokens	706	753	650	649	612	655
Avg. content length in sentences	29	26	26	22	24	22

Table 1: Detailed statistics of the validation, full, and top-15k subsets. The sentence segmentation was performed with spaCy '[de/fr]_core_news_sm' (Honnibal and Montani, 2017) models for sentence segmentation and tokenization with the `paraphrase-multilingual-mpnet` tokenizer.

2 Related Work

Switzerland's multilingual landscape, with four official languages, provides fertile ground for cross-lingual corpus creation. Several prior works have leveraged this linguistic diversity to construct multilingual datasets. For instance, SwissAdmin (Scherrer et al., 2014) is a sentence-aligned corpus of official Swiss government press releases available in German, French, Italian, and English. Similarly, the Bulletin Corpus (Volk et al., 2016) aligns issues of the *Credit Suisse Bulletin* across the same four languages.

20 Minuten has also served as a resource for previous NLP-related studies. Rios et al. (2021) constructed a dataset for automatic text simplification by pairing original German *20 Minuten* articles with their simplified counterparts. More recently, Kew et al. (2023) created a dataset aimed at automatic news summarization in German, further expanding the utility of Swiss news data in NLP research.

With this work, we aim to bridge these two subjects by introducing *20min-XD*, a French-German document-level comparable corpus, sourced from *20 Minuten* (German) and *20 minutes* (French).

3 Data Acquisition

To construct our dataset, we first scrape a total of 593,897 online news articles from both `www.20min.ch/` and `www.20min.ch/fr/`, covering the period from 01.01.2015 to 01.12.2024. In the following subsections, we describe the process applied to identify and align the semantically related articles.

3.1 Validation Set

To establish a gold standard for alignment evaluation, we selected all articles from a single publication day, resulting in 87 German and 70 French articles. Each French article was manually compared against the German articles to identify comparable pairs. While we did not strictly prohibit n:n pairings, the resulting validation set only contains 1:1 pairings. Through this process, we aligned 28 articles into 14 pairs, forming our validation set. Detailed statistics can be found in Table 1.

3.2 Automatic Article Alignment

Since manually aligning comparable articles across languages is time-intensive and requires proficiency in both German and French, we automate the process leveraging multilingual embedding models. Specifically, we encode portions of each article as numerical vectors and compute cosine similarity scores, which range from -1 to 1 ($*100$), to quantify their semantic similarity.

In order to find the most appropriate alignment methods for the *20 Minuten* articles, we conduct experiments on our validation set with different embedding models, alignment approaches, and similarity thresholds.

We choose not to embed the full article texts to ensure a fair comparison across the tested models, some of which have a sequence length constraint (3 out of the tested 5). The results on our validation set suggest that concatenating the article's title and lead provides a sufficiently strong signal for document alignment. This enables resource-efficient experimentation with encoder-based embedding models while avoiding length limitations.

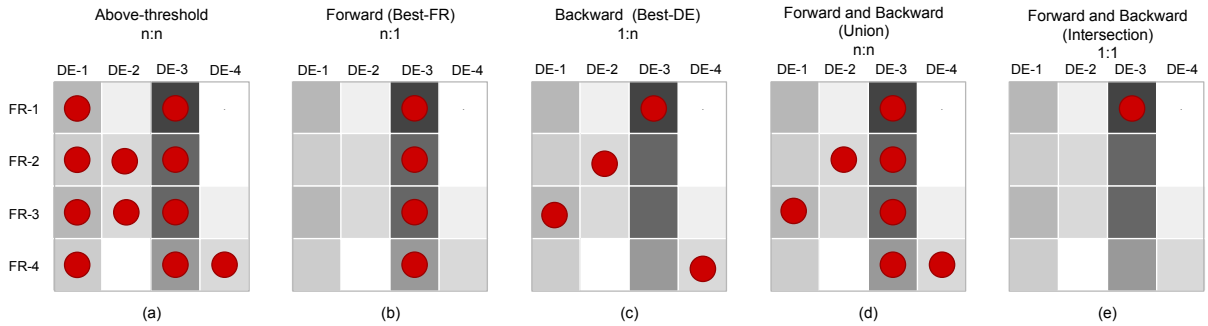


Figure 1: Matrix visualization of different alignment strategies.

Model	Above-threshold	Intersection	Union	Best-DE	Best-FR	Avg.
`paraphrase-multilingual-mpnet`	54.1	**64.7**	54.1	57.8	55.8	57.3
`gte-multilingual-base`	55.6	62.1	55.6	60.0	58.5	58.3
`LaBSE`	53.3	48.5	56.5	60.0	46.2	52.9
`sentence-swissBERT`	62.9	62.5	62.9	61.1	62.5	62.4
`gte-modernbert-base`	45.5	53.3	50.0	54.1	50.0	50.6

Table 2: F1 performance comparison of different models and different alignment approaches on the validation set. The corresponding thresholds are presented in Appendix B.

3.2.1 Models

We experiment with the set of models presented in Table 2: `paraphrase-multilingual-mpnet` is a state-of-the-art multilingual sentence-level paraphrase recognition model (Song et al., 2020); `gte-multilingual-base`, a long-context multilingual text representation model (Zhang et al., 2024); `sentence-swissBERT`, a sentenceBERT-based (Reimers and Gurevych, 2019) model trained on in-domain (*20 Minuten*) data (Grosjean and Vamvas, 2024); `gte-modernbert-base`, modernized, more efficient, long-context version of BERT that has been trained on predominantly English data (Warner et al., 2024).

Preliminary experiments with an LLM-based model (Wang et al., 2024) have shown that they outperform encoder-based models while also being able to process longer input sequences. They do, however, also increase the computational complexity of the embedding process, making it rather resource-intensive and barely feasible in terms of memory and time if scaled to a larger number of documents.

3.2.2 Alignment Strategies

Previous work in cross-lingual alignment has considered multiple possible alignment strategies that either expand or restrict the resulting number of alignments according to different categories as described in e.g. Jalili Sabet et al. (2020) for cross-lingual word alignment. Similarly to Hämmerl et al. (2024), we experiment with strategies that result in a range from weak to strong alignment, where strategies for weaker alignments typically allow a higher range of semantic similarity and multiple possible alignments, while strategies for stronger alignments are more restrictive towards a high semantic similarity and may only include one good alignment (Figure 1).

Above-Threshold considers all document pairs with a similarity score above a certain threshold as alignable, allowing for many-to-many (n:n) alignments. This means that any number of French articles can be linked to any number of German articles without additional constraints beyond the similarity threshold. While this approach captures a broad range of potential alignments, it does not enforce uniqueness or best-match constraints, leading to a higher number of alignments (Figure 1a).

Best-FR applies a many-to-one (n:1, German:French) constraint, where each FR article is aligned to the single DE article with which it has the highest cosine similarity, provided that the similarity exceeds the threshold. This ensures that each FR document has a single best-matching DE counterpart, but multiple French articles can still be mapped to the same German article. This approach prioritizes French articles selecting their closest German equivalent while allowing asymmetry in alignments (Figure 1b).

Best-DE follows the same principle as Best-FR

but from the German perspective, enforcing a one-to-many (1:n, German:French) constraint. This results in a setting where a single German article may be linked to multiple French articles, capturing scenarios where a single German document is the best translation candidate for multiple French counterparts (Figure 1c).

Union takes the union of Best-DE and Best-FR alignments, allowing many-to-many (n:n) alignments, but in a more restrictive manner than the Above-Threshold approach. Instead of considering all pairs above the threshold, it only retains document pairs where at least one side selects the other as its most similar document above the threshold (Figure 1d).

Intersection is the most restrictive strategy, enforcing a one-to-one (1:1) constraint. A valid alignment occurs only when the French article is the best match for the German article and vice versa provided their similarity score exceeds the threshold. This method forms the intersection of Best-DE and Best-FR, ensuring that alignments are bidirectional and mutually optimal (Figure 1e).

3.3 Setting a threshold

Since not every article has a comparable counterpart in the other language, we define a similarity score threshold above which two articles are considered alignable. This threshold must be exceeded in each of the alignment strategies described above. To determine the optimal threshold θ, we iterate through the range of 0 and 100 in steps of 0.5, selecting the one that maximizes the F1 score on our validation set:

$$\hat{\theta} = \arg\max_{\theta \in \{0, 0.5, \ldots, 100\}} F_1(\theta)$$

And we define F1 as follows, where P denotes predicted pairs and G gold pairs:

$$Prec = \frac{|P \cap G|}{|P|}$$

$$Recall = \frac{|P \cap G|}{|G|}$$

$$F_1 = 2 \cdot \frac{Prec \cdot Recall}{Prec + Recall}$$

This process is repeated for each of the embedding models described above. Our results show that `paraphrase-multilingual-mpnet` with the alignment strategy *intersection* at a similarity score threshold of 46, outperforms all other models on the validation set (see Table 2), making it our approach for article alignment.

It is worth noting that the number of samples in our validation set is small (87 German and 70 French articles). This could lead to statistical noise, exaggerating the apparent differences in the results, making them seem larger/smaller than they truly are.

3.4 Choosing A Time Window

To ensure precise alignment and reduce computational complexity, we restrict comparisons to articles published on the same date. This approach minimizes spurious matches between articles that discuss similar topics but are unrelated in terms of specific events or developments.

3.5 Post-Processing

After aligning the French and German articles, we clean the resulting corpus. Manual inspection indicates that faulty articles usually have a suspiciously high similarity score and contain an error message or the same text in the same language. We remove such pairs.

3.6 Sentence Alignment

To provide more fine-grained insights into the dataset, we conduct sentence-level analyses. To achieve this, we first segment articles into sentences using the spaCy '[de/fr]_core_news_sm' (Honnibal and Montani, 2017) models for German and French.

Once segmented, we perform cross-lingual sentence alignment, once again, applying the best performing approach described above: `paraphrase-multilingual-mpnet` with the *intersection* alignment strategy. While we consider only sentence pairs with a similarity score above 46 for our analyses, we release the sentence-aligned version of our corpus on all aligned sentences, including those whose similarity score does not exceed the threshold. This allows for more holistic future analyses, capturing not only the most strongly aligned sentences but also those with the weakest still detectable semantic similarity.

We post-process the sentence-level version of the dataset by removing sentence pairs that contain less than 30 characters, which entails names, trailing characters and source abbreviations.

Similarity Scores	German	French
Cosine: 98.48 (**max**) SentLengthCorr: 0.75 AlignRatio DE: 0.68 AlignRatio FR: 0.56 Monotonicity: 1.0	**Title:** Mobilität.: «Ab 2030 bieten wir nur noch vollelektrische Fahrzeuge an» **Lead:** Die Elektro-Revolution rollt. Traditionelle Autohersteller haben derzeit einen schweren Stand. Wir haben bei Helen Hu, Geschäftsführerin des Schweizer Ablegers von Volvo, seit 2010 in chinesischer Hand, nachgefragt, wie sie die Zukunft der Mobilität sieht.	**Title:** Mobilité: «A partir de 2030, nous ne proposerons plus que des véhicules entièrement électriques» **Lead:** La révolution électrique est en marche. Les constructeurs automobiles traditionnels ont actuellement la vie dure. Nous avons demandé à Helen Hu, directrice de la filiale suisse de Volvo, en mains chinoises depuis 2010, comment elle voit l'avenir de la mobilité.
Cosine: 84.05 (**mean among top-15k**) SentLengthCorr: -0.78 AlignRatio DE: 0.23 AlignRatio FR: 0.21 Monotonicity: -1.0	**Title:** LKW kreuzte Lieferwagen und stürzte dann ab **Lead:** Ein Lastwagen stürzte am Dienstag 300 Meter in die Tiefe. Der 66-jährige Fahrer wurde schwer verletzt. Jetzt gibt es erste Erkenntnisse, wie es zum Unfall kam.	**Title:** Un camion chute de 300 mètres, le chauffeur survit **Lead:** Un chauffeur de poids lourd a été grièvement blessé, mardi, après que son véhicule est sorti de la route, dans le canton d'Uri.
Cosine: 78.65 (**min among top-15k**) SentLengthCorr: -0.47 AlignRatio DE: 0.07 AlignRatio FR: 0.2 Monotonicity: -0.3	**Title:** GP Brasilien - Bottas gewinnt das Sprintrennen – Hamilton nach irrer Aufholjagd auf Rang 5 **Lead:** Am Samstag stand beim GP von Brasilien die Sprint-Entscheidung an. Die 3 WM-Punkte und die Pole-Position für das Rennen am Sonntag sicherte sich Valtteri Bottas.	**Title:** Automobile – Bottas prive Verstappen de la victoire au sprint et de la pole **Lead:** Valtteri Bottas s'est offert la course sprint et partira de la première case dimanche au Grand Prix du Brésil. Max Verstappen sera placé derrière lui et Lewis Hamilton 10e.
Cosine: 46.00 (**min among full dataset**)	**Title:** Sein Zwillingsbruder brachte ihn vor Gericht **Lead:** Hochriskante Börsengeschäfte ihres Verwaltungsratspräsidenten haben eine renommierte Churer Treuhandfirma in den Ruin getrieben. Der Beschuldigte musste vor Gericht erscheinen.	**Title:** Plombé par Kairos, Julius Bär doit se rattraper **Lead:** La filiale italienne de Julius Bär apparaît presque comme la source de tous les maux du gestionnaire de fortune zurichois.

Table 3: Comparison of the title and lead text of the aligned articles receiving the lowest, mean and highest cosine similarity scores from the top 15,000 aligned articles as well as the aligned articles with the lowest overall score from the full set of aligned articles, which is filtered from the final dataset.

3.7 Additional Measures of Similarity

In the corpus description in Section 4 we make use of additional cross-lingual similarity measures apart from the cosine distance that are based on the sentence alignments:

Alignable sentences per document To estimate how much text within an article is highly similar, we compute the relative percentage of alignable sentences. This measure is particularly interesting, as the full document is not considered during automatic article alignment, as described in Subsection 3.2. For each article, we define the alignable sentence ratio as:

$$\text{AlignRatio} = \frac{\text{NumAlignedSentences}}{\text{TotalSentences}}$$

Sentence length correlation If the sentence length, measured as the number of characters in the sentence, differs between the two languages in a systematic way, a high correlation between sentence lengths in aligned articles could be an additional indicator of semantic similarity. Hence, we compute the sentence length correlation of an article as a Pearson correlation.

Monotonicity We measure the cross-lingual monotonicity (degree by which aligned sentences appear in the same order) between an aligned article pair by calculating the Kendall rank correlation of the aligned sentences' position.

4 Dataset

Our alignment process results in 74,507 article pairs. During post-processing the corpus is filtered down to 73,085 article pairs. By agreement with *20 Minuten*, our dataset release is limited to 30,000 articles. Consequently, we select the top 15,000 article pairs sorted by their similarity score for publication, which we refer to as top 15k dataset in the following. Nonetheless, in the remainder of this paper, we will consider both the full dataset and the top 15k article pairs as subject of analysis. The detailed dataset statistics for both are presented in Table 1.

Out of the total 300,000+ sentences in each language from the top 15k dataset, we align 133,693 sentences per language, from which 117,126 are left after filtering. For the correlation studies in Section 4.2, we consider all the sentence pairs with similarity score above 46, totaling to 109,871 sentence pairs.

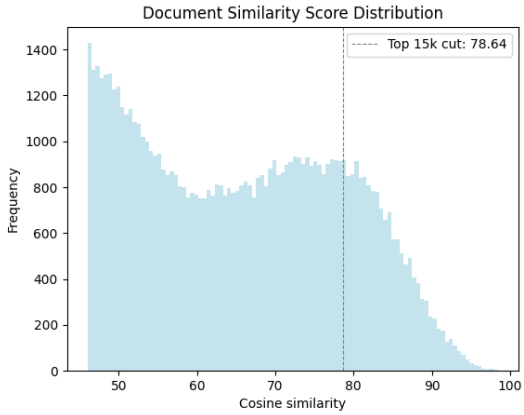


Figure 2: Document (cosine) similarity score distribution over all 74,085 article pairs divided into 100 bins ranging from the threshold of 46 to 100. The dashed line indicates the cut above which the top 15,000 article pairs form the final comparable dataset.

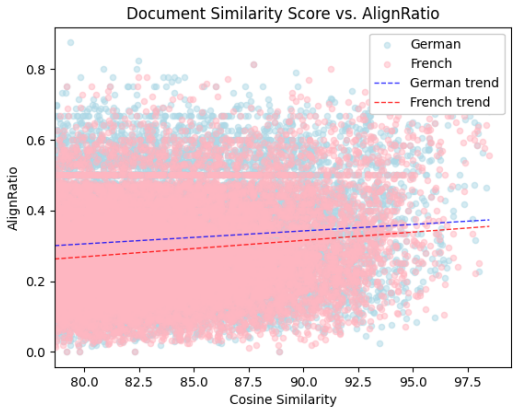


Figure 3: The document cosine similarities in comparison to the AlignRatio of each aligned article in German and French. Both languages show a positive trend line with weak positive correlation (FR: Pearson correlation coefficient $r = 0.145$; DE: $r = 0.103$).

4.1 Qualitative Analysis

Table 3 provides a qualitative comparison of article pairs with the lowest, mean, and highest cosine similarity scores in the top 15k dataset as well as the article pair with the lowest similarity score of all 75,085 initially aligned articles. The highest-scoring pairs exhibit strong lexical and syntactic similarities. The mean-scoring pairs effectively convey the same meaning but demonstrate noticeable differences of the order in which the information is presented. Only the last sentence in the German lead as well as the last phrase in the French lead introduce different information. The lowest-scoring pair in the top 15k dataset covers the same event but differs strongly in word choice and the order in which the information is conveyed. The lowest-scoring pair of the full set of aligned articles, while still loosely related (financial crises), differs in the actual event that is described (e.g., court case leading to a company's collapse vs. corporate struggle with subsidiary).

These results suggest that our dataset mostly consists of articles covering the same topic but with varying degrees of semantic overlap, text structure and length. In order to gain further insight into these features and their relationship to semantic similarity, we conduct a correlation study between the cosine scores of the aligned articles and the different measures described in Section 3.7.

4.2 Quantitative Analysis

4.2.1 Cosine Similarity Distribution

Figure 2 presents the distribution of cosine similarity scores among the aligned articles. The distribution exhibits a right-skewed pattern, suggesting that among the collection of scraped articles, French and German articles with moderate semantic relatedness are more prevalent than those with extremely high similarity scores. The number of articles first drops and then rises again with a rising cosine similarity before reaching a small peak at around 80, located almost exactly at our top 15,000 cutoff point. Following the cutoff, the frequency of article pairs declines sharply to a relatively low level towards higher similarity scores. This pattern loosely suggests the presence of two clusters of article pairs: one representing moderately related articles and another, less prominent, group of more closely related articles.

4.2.2 Correlation with AlignRatio

As a further measure of semantic similarity, we employ the alignment ratio (AlignRatio), which measures the proportionality of aligned sentences between the articles in the two languages, and examine how document similarity scores correlate. As shown in Figure 3, both German and French exhibit weak positive correlations between cosine similarity scores and AlignRatio ($r = 0.145$ for French, $r = 0.103$ for German). These findings suggest that articles with more alignments in the full text tend to have slightly higher semantic similarity. This supports our assumption that relying

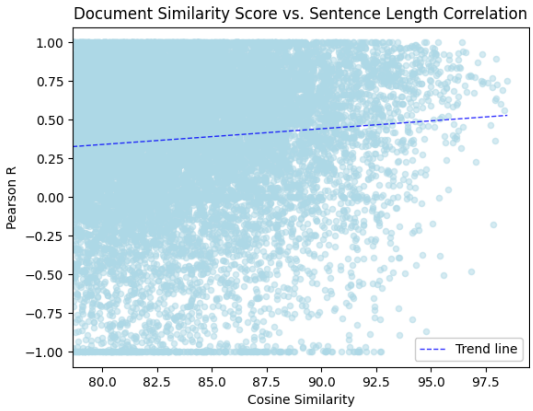


Figure 4: The document cosine similarities in comparison to the sentence length correlation of each aligned article. There is a very weak positive trend of correlation detectable between the two variables (Pearson correlation coefficient $r = 0.084$).

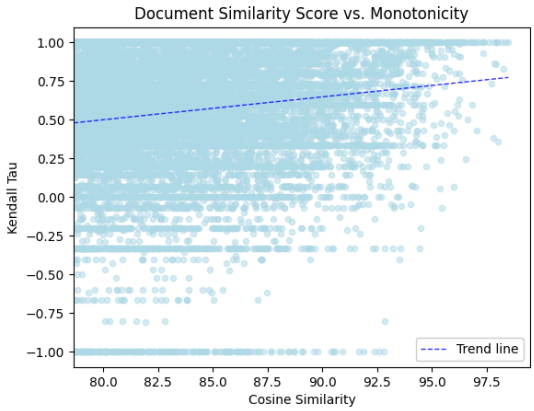


Figure 5: The document cosine similarities in comparison to the monotonicity score of each aligned article. A weak positive correlation trend is detectable between the two variables (Pearson correlation coefficient $r = 0.147$).

solely on the title and lead for the automatic alignment is sufficient but not perfect.

4.2.3 Correlation with Sentence Length

To analyze the relationship between document similarity scores and sentence length variations in aligned articles, we compute the correlation between cosine similarity scores and the sentence length correlation of each article pair. As illustrated in Figure 4, the results indicate a very weak positive correlation ($r = 0.084$).

4.2.4 Correlation with Monotonicity

We also investigate the relationship between document similarity scores and monotonicity, which quantifies the extent to which the order of information (= sentences) is preserved between aligned articles. Figure 5, presents the correlation between cosine similarity scores and monotonicity, showing a weak positive correlation ($r = 0.147$). This suggests, similarly to the previous results, that while higher document similarity scores are slightly associated with a more monotonic alignment of information, the effect is not strong. The clusters near -1.00 and 1.00 may indicate a high number of articles with only one or two aligned sentences — a pattern that could be worth to investigate further.

Given our qualitative analysis and correlation studies, we are confident our dataset maintains an adequate quality for a comparable corpus, covering the full range between direct translations and fairly unrelated text sequences. However, further work with these metrics could provide more insight. Specifically, the alignment ratio may serve as an indicator on which pieces of information are considered essential in both linguistic regions and which are missing from one or the other. Similarly, sentence length correlation could offer valuable perspectives in news-specific translation research. Lastly, monotonicity could be explored further by analyzing topic-specific trends, potentially revealing which topics tend to be translated in a more monotonic fashion than others.

5 Future Work

5.1 Comparing similarity of full text

While using only titles and leads was sufficient for aligning comparable articles, incorporating full article content into the similarity score calculation could provide a more granular and accurate insight into the degree of semantic similarity and relatedness of the articles. This approach could provide a more nuanced representation of narrative structure, argumentation, and topical emphasis. Although computationally intensive, modern embedding models such as `e5-instruct-7b` or `gte-multilingual-base` can theoretically process longer text spans, making full-text comparison increasingly feasible.

5.2 Multilingual long-context embedding models

Encoder-based embedding models are currently going through a renaissance with modernized imple-

mentations, such as ModernBERT (Warner et al., 2024), with significantly improved efficiency and ability to process longer text sequences. At this point in time, multilingual versions of this model specified for the text similarity task are scarce. Future work could explore extending ModernBERT to a multilingual setting and/or optimization for cross-lingual document alignment. Another potential direction is leveraging these modern architectures to develop a document-level counterpart to the (sentence-)swissBERT model.

5.3 Difference recognition

While semantic similarity has been a predominant focus in NLP, the ability to detect and quantify differences between texts—especially across languages—is an emerging research area (Vamvas and Sennrich, 2023). Inspired by diff-based operations in version control, this task could have implications for natural language versioning, collaborative document editing, and editorial workflows. Vamvas and Sennrich show that semantic similarity datasets can be repurposed for difference detection, but have to be synthetically altered to cover cross-linguality and longer text sequences.

Given the variation spectrum observed in our dataset (see Section 4), the diversity of near-translations and loosely related articles, an extension of our corpus with fine-grained annotations—at the paragraph, sentence, or even token level—could enable research into automatic cross-lingual difference recognition.

6 Conclusion

We introduce *20min-XD*, a new French-German document-level comparable dataset of news articles, sourced from the Swiss newspaper *20 Minuten/20 minutes*. The dataset consists of 15,000 aligned articles (or 117,126 aligned sentences) published over a ten-year period. To establish document-level and sentence-level alignment, we employ a multilingual paraphrase recognition model, which demonstrated strong performance during experiments on a manually curated validation set. Both qualitative and quantitative results show that our corpus captures a broad spectrum of cross-lingual similarity, from near-translations to more loosely related text pairs that still cover the same event, with varying degrees of alignable sentences, text lengths and monotonicity. We anticipate its use in future studies across a broad range of linguistically motivated studies.

Acknowledgments

This work was funded by the Swiss National Science Foundation (project InvestigaDiff; no. 10000503 for MW, JV, and RS, and project MUTAMUR; no. 213976 for RS). We sincerely thank everyone at 20 Minuten (TX Group) for their support and for making their data accessible to the research community, with special appreciation to Dean Cavelti for his patient communication. We are also grateful to Unitectra, particularly Peter Loch, for their valuable legal guidance. Finally, we extend our thanks to the Department of Computational Linguistics at the University of Zurich for their inspiring discussions and guidance, with special recognition to Sarah Ebling, Andrianos Michail, Patrick Haller and Anastassia Shaitarova.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Donal Carbaugh and Tovar Cerulli. 2017. *Cultural Discourse Analysis*.

Juri Grosjean and Jannis Vamvas. 2024. Fine-tuning the SwissBERT encoder model for embedding sentences and documents. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 41–49, Chur, Switzerland. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Hanitzsch. 2019. Comparative Journalism Research.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2024. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *Preprint*, arXiv:2408.10151.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 Minuten: A multi-task news summarisation dataset for German. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 1–13, Neuchatel, Switzerland. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

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

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. SwissAdmin: A multilingual tagged parallel corpus of press releases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1832–1836, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Jannis Vamvas and Rico Sennrich. 2023. Towards unsupervised recognition of token-level semantic differences in related documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

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. *Preprint*, arXiv:2412.13663.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Copyright Notice

The resulting dataset is released with the following copyright notice:

German / Deutsch (original):

English / Englisch:

B Experiments on Validation Set

Model	Above-threshold	Intersection	Union	Best-DE	Best-FR
`paraphrase-multilingual-mpnet-base-v2`	61.5	46.0	61.5	47.0	46.0
`LaBSE`	66.0	50.5	50.5	50.5	50.5
`sentence-swissBERT`	74.5	69.5	74.5	73.0	74.5
`gte-multilingual-base`	65.0	65.0	65.0	60.0	56.0
`gte-modernbert-base`	66.0	66.0	66.0	66.0	63.0

Table 4: Optimal threshold values for different models and alignment approaches. The corresponding F1 scores are presented in Table 2.

Assessing Open-Weight Large Language Models on Argumentation Mining Subtasks

Mohammad Yeghaneh Abkenar[*1,3]
yeghanehabkenar@uni-potsdam.de

Weixing Wang[*2,3]
weixing.wang@hpi.de

Hendrik Graupner[1,2,3]
hendrik.graupner@hpi.de

Manfred Stede[3]
stede@uni-potsdam.de

[1]Bundesdruckerei Gruppe GmbH Berlin, [2]Hasso Plattner Institute, [3]University of Potsdam

Abstract

We explore the capability of four open-weight large language models (LLMs) in argumentation mining (AM). We conduct experiments on three different corpora; persuasive essays (PE), argumentative microtexts (AMT) Part 1 and Part 2, based on two argumentation mining subtasks: (i) argument component type classification (ACTC), and (ii) argumentative relation classification (ARC). This work aims to assess the argumentation capability of open-weight LLMs, including Mistral 7B, Mixtral 8x7B, LLaMA2 7B and LLaMA3 8B in both, zero-shot and few-shot scenarios. Our results demonstrate that open-weight LLMs can effectively tackle argumentation mining subtasks, with context-aware prompting improving relation classification performance, though the models' effectiveness varies across different argumentation patterns and corpus types, suggesting potential for specialized adaptation in future argumentation systems. Our analysis advances the assessment of computational argumentation capabilities in open-weight LLMs and provides a foundation for future research.[1]

1 Introduction

Over the past few years, advancements in the broader field of natural language processing (NLP), such as pre-trained transformer-based models (Devlin, 2018), coupled with the increasing availability of diverse data, have significantly enhanced the potential for nearly every area of NLP, including argumentation mining (AM) (Stede and Schneider, 2018; Lawrence and Reed, 2020). AM, and specifically the problem of finding argumentation structures in text, has received much attention in the past decade. The objective of AM is to detect argumentation within text or dialogue, to create detailed representations of claims and their supporting or attacking arguments, and to analyze the reasoning patterns that validate the argumentation. Beyond academic interest, AM attracts significant attention for its diverse applications, as demonstrated by projects like IBM Debater (Bar-Haim et al., 2021), decision assistance (Liebeck et al., 2016), product reviews (Passon et al., 2018) and writing support (Wachsmuth et al., 2016).

[1]Code and data available on https://github.com/myeghaneh/OpenArgMinLLM/tree/main

2 Background and Related work

2.1 Argumentation Mining

Unlike many NLP problems, argumentation mining (AM) is not a single, straightforward task but rather a collection of interrelated subtasks. AM enhances sentiment analysis by delving deeper into the reasoning behind opinions. While sentiment analysis identifies "*what* people think about entity X," AM explores "*why* people think Y about X." One subtask we address is argument component type classification (ACTC), which identifies the type of argumentative discourse units, as defined by Hidey et al. (2017, p. 14) as follows:

- *Claim* (Conclusion): A statement in the text that articulates a perspective on a particular issue. It can include predictions, interpretations, evaluations, and expressions of agreement or disagreement with others' assertions.
- *Premise* (Evidence): A statement presented to strengthen a claim, designed to persuade the audience of its validity. Although premises may express opinions, their main function is to support or refute an existing proposition rather than introduce a new perspective.

We also cover argumentative relation classification (ARC) to identify relations among argumentative discourse units (ADUs) which is defined by Ali et al. (2022, p. 491) as follows:

- *Support* (For): The Support relation occurs when a premise enhances or reinforces a claim.

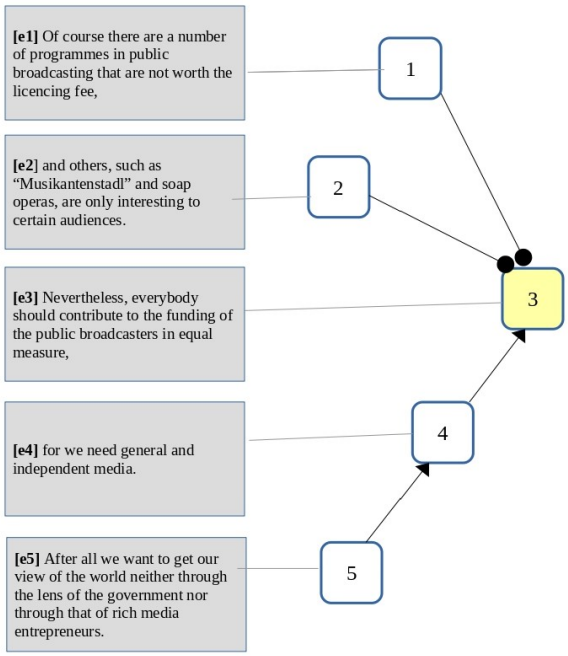


Figure 1: A simplified example from (Peldszus and Stede, 2015b), argumentative microtexts corpus, part 1. The argument structure consists of five elements (e1–e5) with argumentation component type annotation; premise (white boxes) and claim (yellow box) nodes, and supporting (arrow-head) and attacking (circle-head) relations

This can happen in various ways: If the claim is a proposition (such as a fact, opinion, or belief), the premise strengthens the claim's likelihood or truth. If the claim is an action, the premise provides justification or makes the action more acceptable. If the claim is an event, the premise increases the probability that the event occurred.

- *Attack* (Against): The Attack relation occurs when a premise undermines or contradicts a claim. This can manifest in several ways: If the claim is a proposition, the premise weakens the claim's likelihood or truth. If the claim is an action, the premise denies or challenges the justification for the action. If the claim is an event, the premise reduces the probability that the event occurred.
  Figure 1 provides a simplified example from our corpus, showcasing their component types and relationships. It also focuses on two subtasks and their interconnection, demonstrating how they are related and work together to form a final argument.

2.2 Using LLMs for AM

Recently, we saw huge breakthroughs in language modeling. Large Language Models such as GPT4 (Achiam et al., 2023), LLaMA3 (Dubey et al., 2024b), and Mistral (Jiang et al., 2023) have demonstrated strong capabilities in solving various NLP tasks. LLMs are capable of capturing the nuances, context, and semantics of the human language, allowing them to perform tasks such as text generation (Zhao et al., 2023), summarization (Jin et al., 2024; Chang et al., 2023; Zhang et al., 2024), translation (Wu et al., 2024; Xu et al., 2024; Li et al., 2024a), question answering (Li et al., 2024b; Wei et al., 2022), and more. As a result, there is an increasing interest in applying LLMs for computational argumentation tasks. For example, de Wynter and Yuan (2023) evaluated the ability of two LLMs to perform argumentative reasoning. Their experiments involved argumentation mining and argument pair extraction, assessing the LLMs' capability to recognize arguments under progressively more abstract input and output representations. However, their research is limited to the two closed-source language models GPT3 and GPT4. Chen et al. (2023) conducted a comprehensive analysis of LLMs on diverse computational argumentation tasks, their goal was to evaluate LLMs including ChatGPT, Flan models, and LLaMA2 models in both zero-shot and few-shot settings. However, their studies did not address the argumentative relation classification subtask and

they did not use some state-of-the-art models such as LLaMA3 and the Mistral family which according to Sinha et al. (2024) are also promising in various reasoning tasks.

To overcome the above limitations, we explore two key subtasks of argumentation: argumentation discourse unit classification and argument relation classification, using four open-source LLMs across three well-known argument mining corpora. We believe that argumentation mining subtasks are fundamentally different from argument pair extraction and argument generation. As such, argumentation mining subtasks need to be explored differently using various LLMs on the most well-known and important corpora with a similar structure.

3 Corpora and Task Definition

One approach to assessing the reasoning capabilities of LLMs is to evaluate concretely their performance on tasks that necessitate reasoning. We have chosen this approach, in order to measure the ability of different large language models in reasoning. In this paper, we conduct experiments on two central subtasks of argumentation mining using three well-known datasets ,which will be introduced in the next subsections.

3.1 Corpora

Dataset/Subtask	ACTC			ARC		
	Total	Claim	Premise	Total	Support	Attack
AMT1	576	112	464	455	284	171
AMT2	932	171	761	738	524	214
PE	6089	2257	3832	3821	3603	218

Table 1: Summary of sample number and label distributions of the three corpora.

Argumentative Microtexts Part 1(AMT1) The AMT1 corpus, created by (Peldszus and Stede, 2015a), includes 112 short texts (each about 3–5 sentences long) and 576 argumentative discourse units. They were originally written in German and have been professionally translated to English, as well as to Italian (Namor and Stede, 2019), Russian (Fishcheva and Kotelnikov, 2019) and recently to Persian (Abkenar and Stede, 2024) preserving the segmentation and if possible the usage of discourse markers and annotated with complete argumentation tree structures.

Argumentative Microtexts Part 2(AMT2). The second part of AMT, created by (Skeppstedt et al., 2018) using crowd-sourcing, includes 171 short texts with 932 argumentative discourse units in English which is annotated consistent with the approach utilized in the original corpus. One of the differences in this corpus is the existence of an implicit claim which is marked in the XML file.

Persuasive Essays(PE) The PE corpus comprises 402 argumentative essays (totaling 2235 paragraphs) written by English learners in response to specific prompts. Stab and Gurevych (2017) collected these essays from a website and annotated them with argumentation graphs. The essays begin with a question and include a major claim supported by evidence, which may have a substructure. Some sentences are non-argumentative, providing only background or minor elaborations. Each essay has a major claim, typically found at the end, supported by claims within the paragraphs. For consistency with other corpora, we treat "major claim" and "claim" as equivalent and classify argument components (ACs) at the paragraph level.

3.2 Tasks

Argument Component Type Classification (ACTC) Argumentative discourse units (ADUs) are minimal units of analysis, i.e., the smallest elements in a text that contribute to argumentative structure. In this paper, we define ACTC as the classification of these units as either "premise" or "claim"; we do not address the distinction between ADUs and non-argumentative material.

Argumentative Relation Classification(ARC) The goal of argumentative relation identification is to determine whether each pair of ADUs is argumentatively related or not (Rocha et al., 2018). We assume that the task of segmenting the text into ADUs has already been completed. Following (Stab and Gurevych, 2014), given an ordered pair of ADUs, the objective is to classify the relation between them as either "support" or "attack."

4 Methods

4.1 Vanilla Prompting

This approach involves asking the model to classify each ADU independently, without considering the whole context. As shown on the left side of Figure 2, we ask the model: "Please classify the following ADU q_i into one of the categories C_i." This is the same for the ARC, but we ask the same question on pairs of ADUs.

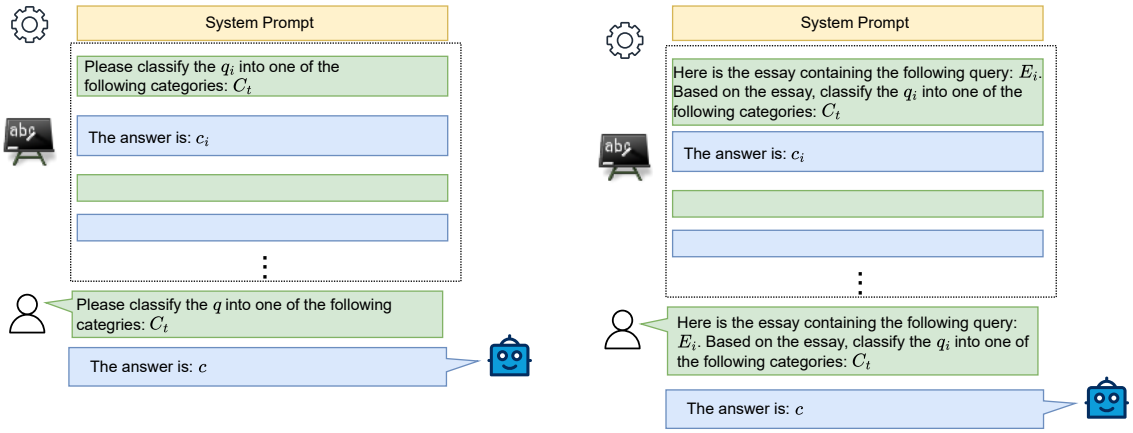


Figure 2: An overview of the prompting methods. Left: Vanilla Prompting. Right: Context-Aware Prompting

4.2 Context-Aware Prompting

This approach asks the model to classify each ADU based on its context in the text. As shown on the right side of Figure 2, we prompt the model with:

"Here is the essay containing the following query E_i. Based on the essay, classify q_i into one of the categories C_i." Unlike the standard method, where each ADU is classified independently, this context-aware prompting requires the model to consider the surrounding context of the essay or microtext for each ADU. For ARC, we ask the model to classify pairs of ADUs, still taking into account the context provided.

4.3 Prompt Design and Engineering

We designed our prompts to clearly communicate the task requirements while avoiding unnecessary complexity. For both vanilla and context-aware approaches, we provided the model with a system message identifying it as "an expert in linguistics and argumentation mining" to prime it for the specialized task.

For few-shot learning, we carefully selected demonstration examples to represent balanced class distributions and varying difficulty levels. When constructing demonstrations, we ensured that they represented diverse argumentative patterns and linguistic constructions present in the target corpus.

We performed preliminary experimentation to optimize prompt formatting, including the use of explicit indicators like "The answer is:" to guide the model's output format. This standardization facilitated easier evaluation and reduced parsing errors. Please refer to appendix B for the design of full prompts for all tasks.

4.4 Model Selection and Implementation

We test our two prompting methods with 4 advanced LLMs, namely **LLaMA 2-7B** (Touvron et al., 2023), **LLaMA 3-8B** (Dubey et al., 2024a), **Mistral-7B** (Jiang et al., 2023), and **Mixtral-8x7B** (Jiang et al., 2024). All samples for all three corpora are within the context window of each model. For comparison, we report the micro F_1-score, because the datasets are all imbalanced.

For implementation, we used the Hugging Face Transformers library[2] to access these models, running inference with a batch size of 1 and random seeds to minimize randomness in outputs. All experiments were conducted three times on NVIDIA A100 GPUs with 40GB of memory. We implemented automated post-processing of model outputs to extract predicted labels and compute metrics for evaluation. As shown in the highlighted part of the figures 3 and 4, performance remained consistently stable, showing minimal variation from stochastic effects.

5 Experiments and Results

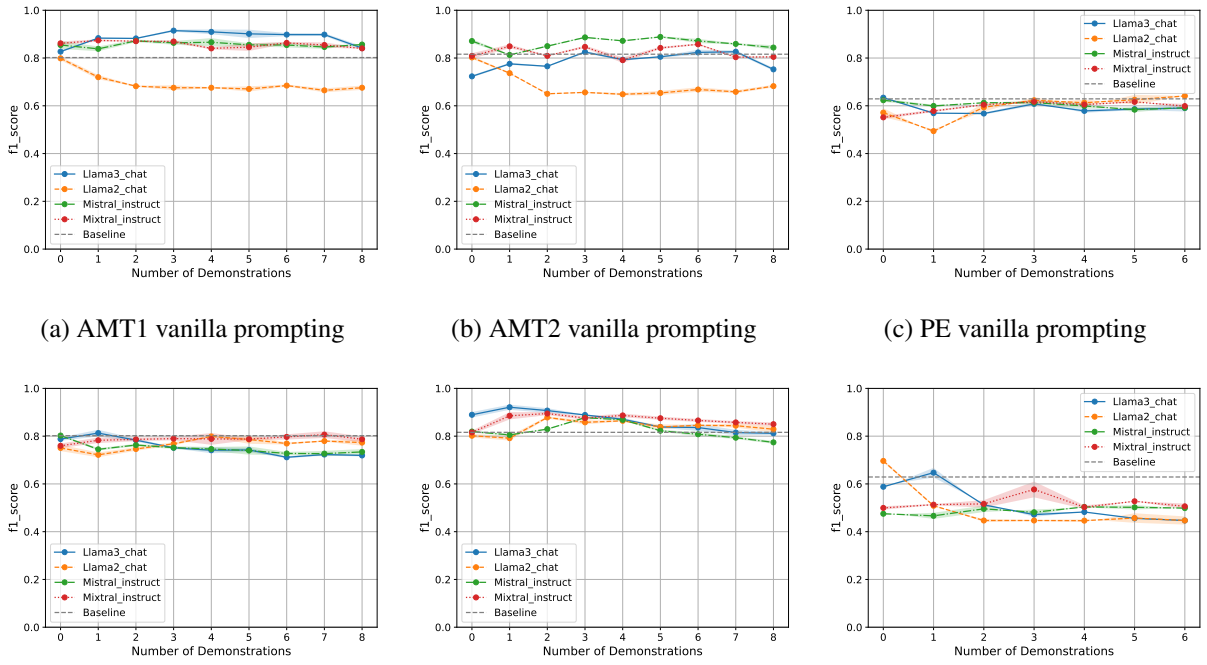
5.1 Baseline

For the experiments on ACTC, we employ a simple strategy of predicting the most frequent (majority) type observed for each ADU type in each of the corpora. As seen in the last row of the table 2 , this approach results in micro F_1-scores of 0.802 for AMT1, 0.816 for AMT2 and 0.629 for PE. Moreover, for the experiment on ARC, we followed the same strategy to calculate a baseline on relation types. This gave us: a micro F_1-score of 0.624 for MT1, 0.710 for MT2 and 0.942 for PE.

[2] https://huggingface.co/docs/transformers/de/

Model	AMT1				AMT2				PE			
	ACTC		ARC		ACTC		ARC		ACTC		ARC	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Mistral (Vanilla)	0.854	**0.745**$^+$	0.491	0.219	0.656	0.767	0.578	0.367	0.623	0.510	0.724	0.340
Mistral (Context)	0.802	0.463	**0.651**$^+$	**0.262**	0.456	0.502	0.693	0.434	0.475	0.428	0.792	0.370
Mixtral (Vanilla)	**0.861**$^+$	0.728	0.556	0.238	0.566	0.604	0.638	0.390	0.551	0.543	0.784	0.371
Mixtral (Context)	0.759	0.585	0.604	0.251	0.598	0.674	**0.734**$^+$	**0.451**$^+$	0.499	0.543	0.887	0.439
LLaMA2 (Vanilla)	0.798	0.489	0.216	0.118	0.465	0.471	0.291	0.236	0.571	0.544	0.350	0.210
LLaMA2 (Context)	0.750	0.574	0.017	0.002	0.577	0.656	0.222	0.154	**0.696**$^+$	0.546	0.632	0.270
LLaMA3 (Vanilla)	0.826	0.717	0.222	0.181	0.514	0.518	0.703	0.380	0.634	**0.612**$^+$	0.883	**0.583**$^+$
LLaMA3 (Context)	0.787	0.657	0.302	0.213	**0.671**	**0.816**$^+$	0.719	0.387	0.588	0.469	**0.931**	0.428
Majority Baseline	0.802	0.446	0.624	0.384	0.816	0.449	0.710	0.415	0.629	0.386	0.942	0.485

Table 2: Performance of different models across AMT1, AMT2 and PE corpora on ACTC, and ARC tasks. The bold values in the table represent the best result for each subtask and dataset, while the $^+$ indicates which of these results were able to outperform the baseline in zero-shot settings



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 3: ACTC. The first row shows the model performance using vanilla prompting on three datasets whereas the below row shows the performance with the context-aware prompting.

5.2 Zero-Shot Performance

Table 2 presents the results of zero-shot prompting. When comparing ACTC and ARC tasks, we find that context prompts generally improve ARC performance across most models and datasets, suggesting that context aids in better understanding the relationships between sentences. However, the ACTC task appears more sensitive to the introduction of context, with some models experiencing a performance drop. LLaMA3 stands out for maintaining strong performance across both tasks and all datasets when using context prompts. This suggests that LLaMA3 is more adaptable to varying prompting methods and datasets in AM tasks.

Considering these baselines, it is clear that the results on ARC are significantly better than those on ACTC, which could be due to the differences in task definitions and their subjective nature. For instance, identifying a claim within a text may be more subjective and context-dependent, requiring a deeper understanding of the argument. In contrast, determining whether two ADUs are supporting or attacking each other is relatively more straightforward and less ambiguous, making it easier for LLMs to classify them.

Regarding evaluation metrics, we observe that LLMs sometimes underperform the baseline on Micro F1 scores while showing stronger results

on Macro F1. This pattern suggests dataset imbalance, where the majority class dominates the Micro metric calculations. The stronger Macro F1 performance indicates that LLMs are better at handling minority classes when evaluating across all classes equally.

Given that LLMs are known to be effective few-shot learners (Brown et al., 2020), these promising zero-shot results suggest significant potential for further optimization through few-shot learning approaches, which we explore in the following section.

5.3 Performance and Number of Demonstrations

5.3.1 Results on ACTC

In the following sections, we only focus on Micro F1 scores. The full results on Macro F1 can be found in A. Figure 3 illustrates the performance on the ACTC task across the datasets, under different numbers of demonstrations. For the ACTC task, context-aware prompting can bring all models to a similar level. I.e., weaker models like LLaMA2 are enhanced while the stronger models are degraded. For example, considering three-shot learning on the AMT1 dataset, LLaMA3 can achieve 86% micro F_1-score using vanilla prompting (a), but the micro F_1-score drops to 81% with context-aware prompting (d). For the AMT2 dataset, we observe similar phenomena in (b) and (e), however, here LLaMA3 achieved the best results in the first shot using context-aware prompting.

In comparison, LLaMA2 improves from 79% (a) to 87% (d) by applying context-aware prompting. Moreover, we find that the application of context-aware prompting significantly reduces the performance disparity between the AMT1 and AMT2 datasets. This suggests that providing additional contextual information helps the models to handle variations between these datasets more effectively, resulting in a more uniform performance across different versions of the AM tasks.

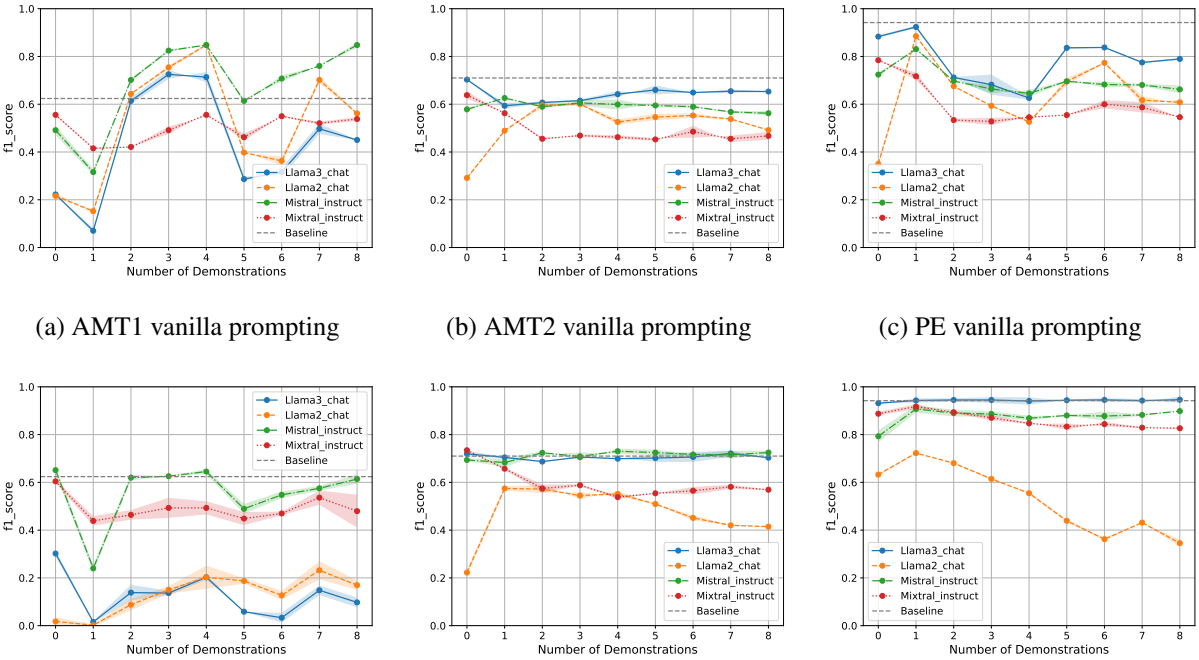
Our few-shot experiments on ACTC highlight the complexities of adapting to different argumentation styles. In the PE vanilla prompting setup in (c), model performance remains relatively stable across different numbers of demonstrations, with slight variations among models. This suggests that ACTC's argumentation structures may not be as easily influenced by increasing demonstration. However, in the context-aware prompting setting of PE, we see more fluctuations in (f), particularly in the early demonstrations. One possible explanation is the longer text length in PE dataset compared to the MT datasets. This corpus differs from the microtext corpora in that each paragraph can contain more than one claim, which impacts the weighting of component in the final F_1-micro score calculation. Furthermore, for the ACTC subtask in the PE dataset, the addition of contextual information could actually degrade the model's ability to solve the task effectively. The increased context could introduce more complexity, making it harder for models to solve ACTC task, which is one of subjective and complex task of Argumentation Mining that is align with finding in (Levy et al., 2024)

5.3.2 Results on ARC

For the ARC task, we see slightly different patterns of model performance in Figure 4. However, we still observe that context-aware prompting serves as an effective stabilizer for model performance. Comparing (a) and (d), we find that when models are prompted with additional contextual information, they exhibit reduced fluctuations in their performance regarding different numbers of demonstrations, suggesting that this approach helps mitigate the impact of noise brought by additional demonstrations. In contrast, vanilla prompting, which lacks this additional context, often results in more erratic performance across different numbers of demonstrations, likely because the models are more susceptible to the inherent variability and difficulty of the tasks. This fluctuation in vanilla prompting can be attributed to the models' struggle to consistently grasp the underlying patterns in the data without sufficient context, leading to inconsistent F_1-scores. By providing context-aware prompting, the models are better equipped to understand and process the tasks at hand, resulting in more stable and reliable outputs.

Our few-shot experiments with the ARC task highlight the challenge of transfer learning across different argumentation patterns. The PE corpus, with its academic writing style, showed the most consistent improvement with additional demonstrations, suggesting that formal argumentation patterns may be more learnable from examples. In contrast, the more varied AMT2 corpus showed less consistent improvement patterns, indicating that diverse argumentation styles may require more sophisticated adaptation approaches.

In comparing model architectures, we observed



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 4: ARC. The first row shows the model performance using vanilla prompting on three datasets where the below row shows the performance with the context-aware prompting.

that Mixtral's mixture-of-experts architecture consistently outperforms the others in the few-shot regime for relation classification, potentially due to its ability to activate different expert pathways for different relation types. This architectural advantage is particularly evident in the context-aware setting, where the model must integrate information across longer text spans.

5.4 Error Analysis and Qualitative Assessment

We conducted a detailed error analysis to understand when and why models fail at argumentation mining tasks. For ACTC, all models struggle most with claims that lack explicit stance indicators or that use hedging language. For instance, in AMT1, the sentence "Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins" was often misclassified as a claim due to its evaluative language, despite functioning as a premise in context.

For ARC, the most challenging cases involve implicit support or attack relations where no explicit discourse markers (like "because" or "however") are present. Models particularly struggle with relations that require domain knowledge to interpret correctly. Additionally, all models show a bias toward predicting the majority class, especially in the PE corpus, where support relations vastly outnumber attack relations.

Qualitatively, we observed that LLaMA3 produces more coherent explanations for its decisions when prompted to explain its reasoning, suggesting deeper understanding of argumentative structures. Mixtral exhibits greater sensitivity to subtle indicators of argumentative function, while Mistral performs better at identifying explicit discourse markers as indicators of relation type.

6 Discussion and Conclusion

6.1 Result Comparison with Literature

We assessed the reasoning abilities of four LLMs. Our evaluation focused on two sub-tasks in argumentation mining: ACTC and ARC. However, comparing these results with the state of the art is not straightforward, primarily due to the variations in how different metrics are evaluated and reported across studies. The LLMs performed particularly excelled in ARC in comparison to our majority baseline, and performed well in ACTC, surpassing or closely matching the results reported in (Abkenar et al., 2021) and (Chernodub et al., 2019) for AMT1 and PE, based on the micro F_1-score. However, statistical analysis of the LLMs' predictions shows that their performance generally

differs between AMT1 and AMT2, which we attribute to a difference in text quality due to the varying elicitation conditions. We also revealed that demonstrations serve as stabilizers rather than enhancers for both AM tasks.

6.2 Theoretical Implications

Our findings have several theoretical implications for understanding LLMs' capabilities in structured reasoning tasks. First, the models' strong zero-shot performance suggests they have acquired implicit knowledge of argumentation structures during pretraining, despite not being explicitly trained on argumentation tasks that we designed in this work. This supports the hypothesis that general language understanding includes some degree of argumentation comprehension. Second, the stabilizing rather than enhancing effect of demonstrations suggests that few-shot learning in AM primarily helps models understand task framing rather than teaching them new argument patterns. This challenges simplistic views of in-context learning as analogous to traditional learning from examples.

Limitations

We conducted our study on two central subtasks of AM. However, other subtasks, such as the identification of argument components and the evaluation of argument quality, need to be addressed. We also aim to evaluate more recent LLMs, such as DeepSeek (Guo et al., 2025) and Hermes (Teknium et al., 2024), which are potentially strong in reasoning. For future work, we intend to explore the impact of input length on model performance in AM subtasks. Additionally, our results focus exclusively on English argumentative corpora. We recommend that future research explores other languages, especially those underrepresented in argumentation mining.

Acknowledgments

We thank our colleagues in Innovations department of the Bundesdruckerei GmbH and the Hasso Plattner Institute for providing us with the opportunity to freely work on our research topics. Thank you for fostering an environment that encourages innovation and academic growth. The authors also sincerely thank the anonymous reviewers for their thoughtful recommendations that significantly improved this paper.

References

Mohammad Yeghaneh Abkenar and Manfred Stede. 2024. Neural mining of persian short argumentative texts. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 30–35.

Mohammad Yeghaneh Abkenar, Manfred Stede, and Stephan Oepen. 2021. Neural argumentation mining on essays and microtexts with contextualized word embeddings.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Basit Ali, Sachin Pawar, Girish Palshikar, and Rituraj Singh. 2022. Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 491–500.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project debater apis: Decomposing the ai grand challenge. *arXiv preprint arXiv:2110.01029*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

Adrian de Wynter and Tommy Yuan. 2023. I wish to have an argument: Argumentative reasoning in large language models. *arXiv preprint arXiv:2309.16938*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and et al. Angela Fan. 2024a. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024b. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-lingual argumentation mining for russian texts. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 134–144. Springer.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024a. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024b. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18608–18616.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153.

Ivan Namor and Manfred Stede. 2019. Mining italian short argumentative texts. In *Proceedings of the 5th Workshop on Argument Mining*.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Andreas Peldszus and Manfred Stede. 2015a. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Andreas Peldszus and Manfred Stede. 2015b. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from english to portuguese. In *Proceedings of the 5th Workshop on Argument Mining, 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. Evaluating open language models across task types, application domains, and reasoning types: An in-depth experimental analysis. *arXiv preprint arXiv:2406.11402*.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

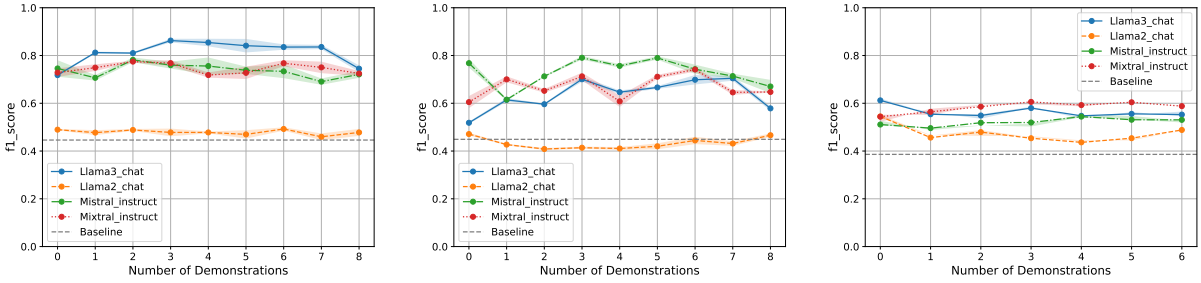
Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

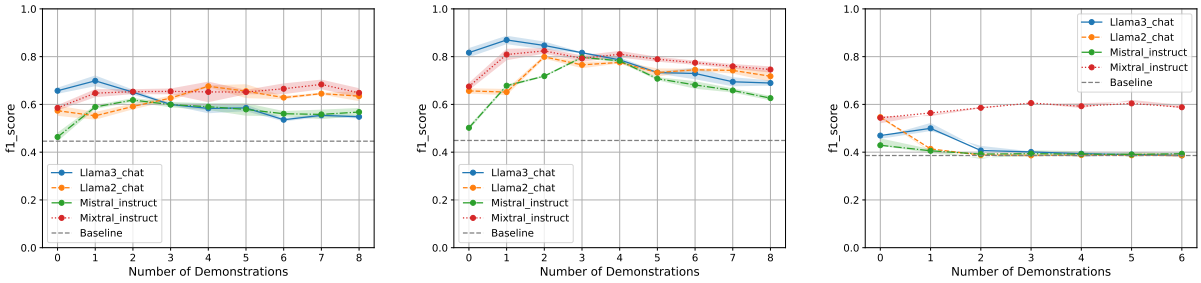
A Evaluation Results on Macro F1

B Prompt Design

In this section, we show four examples of our prompts designed for both vanilla and context-aware prompting methods.



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

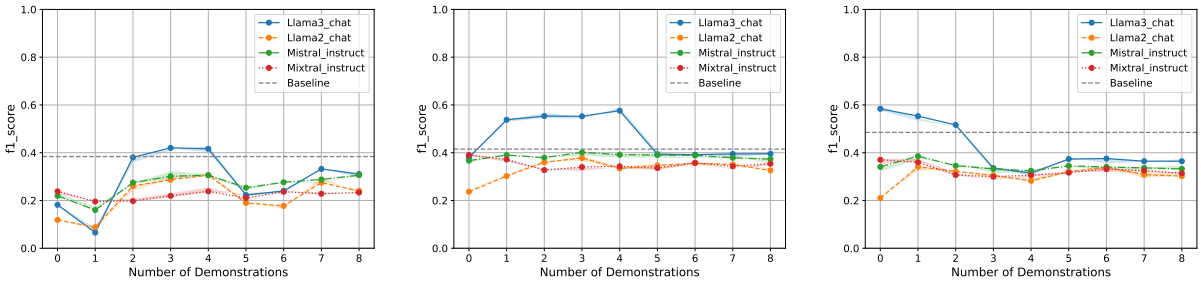


(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

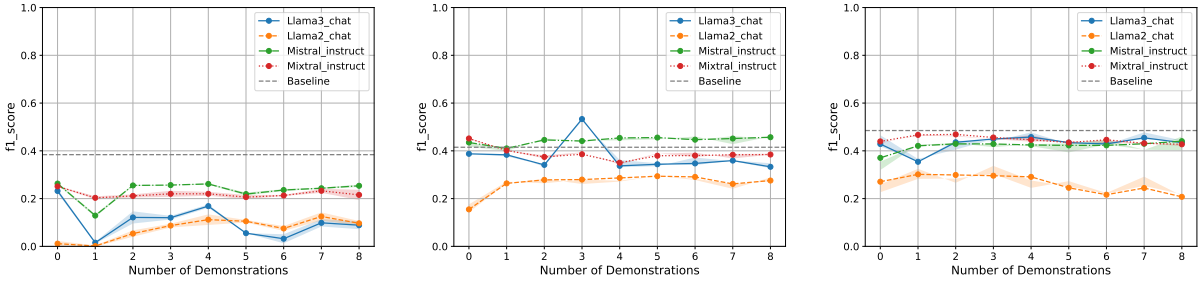
Figure 5: ACTC. The first row shows the model performance using vanilla prompting on three datasets whereas the below row shows the performance with the context-aware prompting.

System: You are an expert in linguistics and you are very good at argumentation mining. Now you are given a paragraph with indexs. Each sub-text is either the claim or premise. Your task is to find the claim in the paragraph. Provide the index of the claim in the text with < >. There is only one correct index.
Demo: Yes, it's annoying and cumbersome to separate your rubbish properly all the time. <2>Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. <3>But still Germany produces way too much rubbish, <4>and too many resources are lost when what actually should be separated and recycled is burnt. <5>We Berliners should take the chance and become pioneers in waste separation! The answer is: <5> One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt. <2>And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. <3>Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners. <4>Of course, first they'd actually need to be caught in the act by public order officers, <5>but once they have to dig into their pockets, their laziness will sure vanish! The answer is: <3> ...
Query: <1>For dog dirt left on the pavement dog owners should by all means pay a bit more. <2>Indeed it's not the fault of the animals, <3>but once you step in it, their excrement seems to stick rather persistently to your soles. The answer is:

Table 3: Example of Vanilla Prompting for ACTC task using AMT1 dataset.



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting



(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 6: ARC. The first row shows the model performance using vanilla prompting on three datasets where the below row shows the performance with the context-aware prompting.

System: You are an expert in linguistics and you are very good at argumentation Mining. Now you are given a sentence and a paragraph containing this sentence as a reference. Your task is to classify the sentence as either a Claim or a Premise according to the paragraph. Answer with <0> for Premise and <1> for Claim. There is only one Claim in the paragraph.

Demo: Please classify the sentence: Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. as either <1> for Claim or <0> for Premise in the given context: Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We Berliners should take the chance and become pioneers in waste separation!
The answer is: <0>

Please classify the sentence: And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. as either <1> for Claim or <0> for Premise in the given context: One can hardly move in Friedrichshain or Neuk0̈0f6lln these days without permanently scanning the ground for dog dirt. And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners. Of course, first they'd actually need to be caught in the act by public order officers, but once they have to dig into their pockets, their laziness will sure vanish!
The answer is: <0>
...

Query: Please classify the sentence: For dog dirt left on the pavement dog owners should by all means pay a bit more. as either <1> for Claim or <0> for Premise in the given context: For dog dirt left on the pavement dog owners should by all means pay a bit more. Indeed it's not the fault of the animals, but once you step in it, their excrement seems to stick rather persistently to your soles.
The answer is:

Table 4: Example of Context-aware Prompting for ACTC task using AMT1 dataset.

System: You are an expert in linguistics and you are very good at Relation Mining. Now you are given two sentences in an essay. Your task is to classify the relationship between the two sentences as 'Support' if Sentence 1 supports the stance of Sentence 2; or 'Attack' if Sentence 1 does not support Sentence 2. Provide only one word. DO NOT give explanation

Demo: Sentence 1:One who is living overseas will of course struggle with loneliness, living away from family and friends. Sentence 2:living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet.
The answer is: Attack

Sentence 1:What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. Sentence 2:through cooperation, children can learn about interpersonal skills which are significant in the future life of all students.
The answer is: Support
...

Query: Sentence 1:it also has to be affordable for the consumer. Sentence 2:When a product is commonly used, it becomes trustworthy for the society, no matter what quality it is.
The answer is:

Table 5: Example of Vanilla Prompting for ARC task using PE dataset.

System: You are an expert in linguistics and you are very good at Relation Mining. Now you are given two sentences in an essay. Your task is to classify the relationship between the two sentences as 'Support' if Sentence 1 supports the stance of Sentence 2; or 'Attack' if Sentence 1 does not support Sentence 2. Use the context as supporting context. Provide only one word. DO NOT give explanation.

Demo: Sentence 1:One who is living overseas will of course struggle with loneliness, living away from family and friends. Sentence 2:living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. Please classify the relationship as either Attack or Support based on the given context: Living and studying overseas It is every student's desire to study at a good university and experience a new environment. While some students study and live overseas to achieve this, some prefer to study home because of the difficulties of living and studying overseas. In my opinion, one who studies overseas will gain many skills throughout this experience for several reasons. First, studying at an overseas university gives individuals the opportunity to improve social skills by interacting and communicating with students from different origins and cultures. Compared to . . . . . in general life.
The answer is: Attack

Sentence 1:What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. Sentence 2:through cooperation, children can learn about interpersonal skills which are significant in the future life of all students. Please classify the relationship as either Attack or Support based on the given context: Should students be taught to compete or to cooperate?It is always said that competition can effectively promote the development of economy. In order to survive in the competition, companies continue to improve their products and service, and as a result, the whole society prospers. However, when we discuss the issue of competition or cooperation, ... in one's success.
The answer is: Support
. . .

Query: Sentence 1:it is necessary to make sure that people can live a long life. Sentence 2:animal experiments have negative impact on the natural balance. Please classify the relationship as either Attack or Support based on the given context: Using animals for the benefit of the human beings with the rapid development of the standard of people's life, increasing numbers of animal experiments are done, new medicines and foods, for instance. Some opponents says that it is cruel to animals and nature, however, I believe that no sensible person will deny that it is a dramatically cruel activity to humanity if the latest foods or medicines are allowed to be sold without testing on animals. In my essay, I will discuss this issue from twofold aspects.First of all, as we all know, animals are friendly and vital for people, because if there are no animals in the world, the balance of nature will break down, and we, human, will die out as well. The animal experiments accelerate the vanishing of some categories of animals. In other words, doing this various testing is a hazard of human's future and next generation.Though animal experiments have negative impact on the natural balance, it is necessary to make sure that people can live a long life. To begin with, it is indisputable that every new kind food or pill may be noxious, and scientists must do something to insure that the new invention benefits people instead of making people ill or even dying. The new foods or medicines are invented to promote the quantity of human's life. Thus even if they are volunteers; they cannot take the place of animals to test the new foods or medicines. Furthermore, it also have potentially harm for human's health without any testing. To sum up, I reaffirm that although there is some disadvantages of animals' profits, the merits of animal experiments still outweigh the demerits.
The answer is:

Table 6: Example of Context-aware Prompting for ARC task using PE dataset.

Detecting Greenwashing Hints in ESG Reports: Linguistic and Claim Analysis in Traffic-Related Emissions Disclosures

Johannes Florstedt and **Jonas Fahlbusch** and **Moritz Sontheimer**
Technische Universität Berlin
florstedt@campus.tu-berlin.de

Abstract

Detecting greenwashing in corporate Environmental, Social, and Governance (ESG) reports presents challenges due to data scarcity and ambiguity, particularly concerning complex topics like traffic emissions. This paper introduces a machine learning framework to identify potential greenwashing indicators by analyzing linguistic patterns and claim substantiation in 150 ESG reports from German DAX companies, 2020-2023. We evaluate sentiment polarity (VADER), linguistic specificity (ClimateBERT), and internal claim verification (Sentence-BERT). Key findings reveal two complementary signals suggesting areas for scrutiny: first, an external discrepancy where high reporting positivity coincides with lower relative external performance proxy scores (Refinitiv Emission Score), identifying specific firms potentially overstating performance; second, an internal inconsistency where low linguistic specificity correlates with weak internal claim substantiation, indicating potential *cheap talk*. While automated external claim verification proves difficult, the framework provides quantitative indicators to help stakeholders prioritize the assessment of ESG reports in the critical traffic sector.

1 Introduction

Heightened stakeholder demand for corporate Environmental, Social, and Governance (ESG) transparency has driven a significant increase in sustainability reporting. This trend, however, is accompanied by growing concerns regarding *greenwashing* – the practice where organizations portray their environmental performance more favorably than warranted by their actions (European Securities and Markets Authority, 2023). Evidence suggests this is a considerable issue, with studies finding misleading claims prevalent across various markets (UK Competition and Markets Authority, 2021; Australian Competition and Consumer Commission, 2023). The systematic detection of greenwashing is complicated by the lack of a universally standardized definition and the scarcity of reliably labeled datasets, which limits the applicability of conventional supervised machine learning methods (Calamai et al., 2025). As a result, research increasingly utilizes Natural Language Processing (NLP) techniques to analyze the extensive textual content of ESG reports for linguistic and semantic patterns that might indicate misrepresentation (Bingler et al., 2022; Vinella et al., 2024).

These detection challenges are particularly pronounced in the context of disclosures related to **traffic and transportation emissions**. This area holds critical importance due to the transport sector's substantial contribution to global greenhouse gas emissions (Shukla et al., 2022). Disclosures often involve complex Scope 3 emissions data (e.g., logistics, business travel, employee commuting), which are notoriously difficult to measure, report consistently, and verify externally (Berthe et al., 2025). This inherent complexity and potential data opacity may provide avenues for companies to engage in greenwashing within their traffic-related narratives (Robinson, 2022).

This paper presents and evaluates an ML framework specifically developed to identify potential greenwashing indicators within these traffic-related disclosures. We investigate the interplay between sentiment polarity and linguistic specificity, and their relationship to external performance proxies. Furthermore, we assess the degree to which specific claims are substantiated by internal textual evidence using semantic similarity techniques, while also exploring the practical challenges associated with attempts at external verification. A key objective is to understand how linguistic analysis and claim verification compare and potentially complement each other in highlighting potential greenwashing risks. Our aim is not to achieve definitive greenwashing classification, but rather to provide quantitative, data-driven indicators that enable stakeholders to prioritize and focus their scrutiny efforts more effectively.

2 Background and Related Work

Greenwashing often involves the strategic deployment of language, such as using excessively positive framing or ambiguous terminology, potentially to divert attention from unfavorable environmental performance (Delmas and Burbano, 2011). Indicators commonly associated with greenwashing include communication that appears overly positive relative to actual performance, the use of vague or non-specific language lacking concrete details, and the presentation of unsubstantiated claims regarding

environmental benefits (European Securities and Markets Authority, 2023; European Parliament, 2023). The traffic sector, characterized by its complex and often difficult-to-verify Scope 3 emissions footprint (Berthe et al., 2025), represents an area susceptible to such practices, as highlighted by public controversies involving the automotive and aviation industries (Robinson, 2022; Plucinska, 2023).

Existing ML approaches for detecting greenwashing signals are diverse. Supervised learning methods frequently grapple with the scarcity of labeled data, sometimes employing synthetically generated labels (Vinella et al., 2024), weak supervision based on aggregated firm-level scores (Sharma et al., 2024), or requiring substantial manual annotation efforts (Bingler et al., 2024). Unsupervised techniques often focus on identifying discrepancies between corporate narratives and external benchmarks. Common strategies found in the literature involve comparing report content and tone against quantitative ESG performance scores (Chen and Ma, 2024; Lagasio, 2024), while others analyze alignment with public discourse, such as media sentiment and topic coverage (Lipenkova et al., 2023; Zhao et al., 2023).

Common linguistic features analyzed include sentiment polarity (Chen and Ma, 2024; Zhao et al., 2023) and the degree of linguistic specificity versus vagueness (Bingler et al., 2024; Vinella et al., 2024). Domain-specific language models, particularly ClimateBERT (Webersinke et al., 2022), have shown improved effectiveness in analyzing the specialized vocabulary and context of climate-related text compared to general models (Bingler et al., 2024; Trajanov et al., 2023). While stylistic analysis is relatively common, the systematic evaluation of internal claim substantiation within reports seems less explored in the context of greenwashing detection. Automated fact-checking tools like LOKI (Li et al., 2024) offer potential pathways for external verification but face considerable hurdles when applied to the complex and nuanced nature of ESG claims (Leippold et al., 2024).

Our work integrates insights from these varied approaches. We employ a primarily unsupervised framework focusing on quantifiable indicators (positivity, specificity, internal consistency) tailored specifically to the traffic domain. We utilize accessible tools, including ClimateBERT variants and Sentence-BERT, and importantly, compare derived communication patterns against an external performance proxy.

3 Methodology

Our analysis is based on a corpus of 150 English-language ESG reports collected from German DAX companies for the years 2020 through 2023. Text was extracted from PDF documents using the `Kreuzberg` tool, chosen for its ability to produce cleaner textual output suitable for NLP tasks compared to some standard libraries. A multi-pipeline framework was implemented to analyze disclosures related to traffic emissions.

1. Filtering Traffic-Related Content: The core analysis focused on relevant text segments identified through a sequential filtering process applied to 500-character chunks (with a 20-character overlap, intended to preserve context across boundaries). First, the ClimateBERT Detector model (Bingler et al., 2024) classified chunks based on climate relevance, retaining those exceeding a confidence score threshold of 0.5. Second, these climate-relevant chunks were further filtered using a custom-developed traffic lexicon (keywords including 'fleet', 'electric vehicle', 'transport', 'fuel', 'logistics', 'business travel', 'commuting', 'aviation', 'shipping') to isolate segments specifically discussing traffic-related issues. This filtering cascade aimed to focus the analysis efficiently on the most pertinent text passages.

2. Language Analysis Module: This module evaluated the stylistic properties of the filtered chunks. Linguistic specificity was assessed using the ClimateBERT Specificity model (Bingler et al., 2024), classifying each chunk as either 'specific' (containing concrete data, metrics, or detailed actions) or 'non-specific' (general, vague statements). The proportion of 'specific' chunks per document was calculated to derive a document-level Specificity Score (0-100). Sentiment polarity was determined using VADER (Hutto and Gilbert, 2014), selected for its capability to handle contextual nuances like negation and intensifiers found in narrative text. The average VADER compound score across a document's filtered chunks was linearly transformed into a Positivity Score (0-100 scale, where 50 indicates neutrality).

3. Claim Verification Module: This module examined the substantiation of claims. For *internal verification*, potential claim sentences (identified heuristically via modal/assertive keywords + traffic terms) and potential proof sentences (identified via evidence-related keywords) were extracted. Sentence-BERT (Reimers and Gurevych, 2019), specifically the efficient `all-MiniLM-L6-v2` model, generated embeddings for claims and proofs. Cosine similarity was computed between each claim and all potential proof sentences from the same report. The highest similarity score to a non-identical proof sentence was considered the measure of internal support. An average Internal Claim Score (0-100) per document summarized this semantic coherence. While pragmatic, these heuristic extraction steps influence the inputs to the similarity assessment and represent a known limitation. For *external verification*, a limited, exploratory analysis was performed on a small set of claims (prioritizing those with low internal scores) using the public LOKI web interface (Li et al., 2024) to investigate the feasibility and challenges of automated web-based verification.

4. Performance Proxy: We utilized the Refinitiv Emission Score (0-100), accessed via the Refinitiv Eikon database, as an external proxy for corporate environmental performance. This score was chosen due to its focus on emissions within the broader ESG context, its consideration of Scope 1-3 emissions data (though not specifically isolating traffic), and its methodology

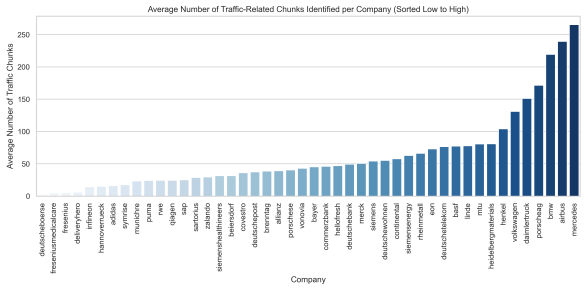


Figure 1: Average Number of Traffic-Related Chunks Identified per Company (Sorted Low to High).

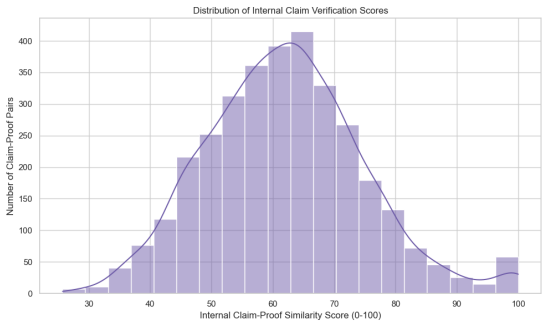


Figure 2: Distribution of Internal Claim Verification Scores (Individual Claim-Proof Pairs).

which integrates company disclosures with external controversy screening, offering a relatively comprehensive benchmark available for this study (LSEG Data & Analytics, 2024).

5. Analysis: The core analysis involved calculating correlations (Pearson, Spearman) between the company-level average indicators (Positivity, Specificity, Claim Score, Refinitiv Score). Visual discrepancy analysis using scatterplots was employed to identify specific companies exhibiting patterns potentially indicative of greenwashing risk relative to the observed trends.

4 Results

4.1 Reporting Intensity and Linguistic Style

The analysis revealed substantial variation in the extent to which companies elaborated on traffic-related climate issues. Figure 1 illustrates the wide range in the average number of filtered, relevant text chunks per company, with firms in transport-intensive sectors generally providing more content, though significant intra-sector variation exists. This variability in reporting intensity affects the statistical robustness of metrics for companies with minimal relevant text.

On average, the linguistic style within these disclosures tended towards positive sentiment (mean Positivity Score 70.5) and moderate specificity (mean Specificity Score 65.4%). Importantly, no statistically significant correlation was found between a company's average positivity score and its average specificity score, suggesting these represent largely independent dimensions

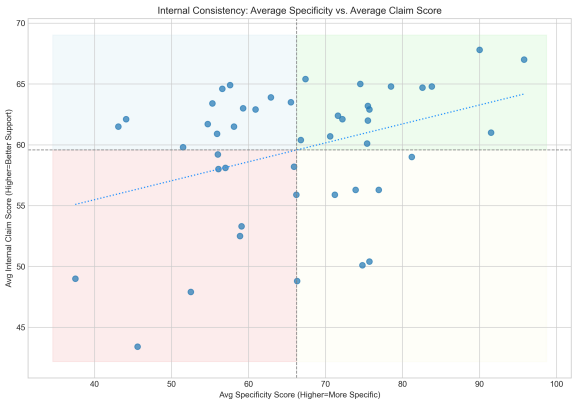


Figure 3: Internal Consistency: Avg Specificity vs. Avg Claim Score. Bottom-left (red) suggests potential inconsistency.

of communication style in this context.

4.2 Internal Claim Verification

The internal claim verification assessed semantic similarity between identified claims and potential supporting sentences within the same document. The distribution of individual claim-proof similarity scores (Figure 2) was centered around a mean of 62.0 (0-100 scale). This suggests that, typically, claims found moderately related textual evidence within the report. However, the broad distribution of scores indicates varying degrees of internal substantiation, with some claims finding strong semantic links while others lacked clear support.

4.3 Correlation Analysis and Discrepancies

Analysis of company-level average indicators over the 2020-2023 period revealed significant relationships:

Internal Consistency Signal: A statistically significant, moderate positive correlation was observed between Average Specificity and Average Internal Claim Score (Spearman rho=0.310, p=0.036; Pearson r=0.363, p=0.013). This key finding indicates that companies employing more specific and detailed language in their traffic disclosures also tend to exhibit stronger internal semantic coherence, meaning their claims are better supported by other statements within the report (visualized in Figure 3). This linkage between linguistic style and internal evidence provides a measurable indicator of reporting consistency.

External Alignment Signal: Average Positivity showed a significant positive correlation with the Average Refinitiv Emission Score proxy (Spearman rho=0.332, p=0.024; Pearson r=0.472, p=0.001). On average, companies assessed as having better emissions performance (via the proxy) tended to use more positive language in their traffic-related sections. No significant correlations were found between Specificity or Claim Score and the Refinitiv score. Analyzing discrepancies from the main Positivity-Refinitiv trend is crucial here. Figure 4 identifies companies (marked with red

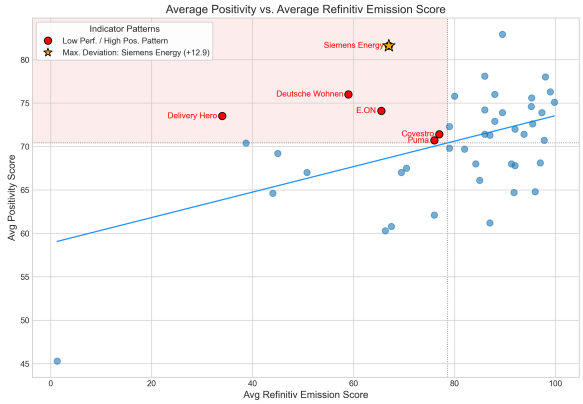


Figure 4: External Discrepancy: Avg Positivity vs. Avg Refinitiv Score. Upper-left (red circles) flags potential risk.

circles: Delivery Hero, Dt. Wohnen, E.ON, Covestro, Puma, Siemens Energy) situated in the upper-left area, characterized by high reporting positivity despite lower relative performance proxy scores. This pattern aligns with theoretical greenwashing risk profiles (Delmas and Burbano, 2011). Siemens Energy exhibited the largest positive deviation from the overall trend line.

5 Discussion

The framework applied in this study provides quantitative indicators and reveals communication patterns that can aid in the systematic identification of potential greenwashing risks within the challenging domain of traffic-related ESG disclosures.

A key finding is the significant positive correlation between linguistic specificity and internal claim substantiation, serving as an informative **internal consistency check**. Reports characterized by both vague language (low specificity) and weak internal support for claims (low internal claim score) – corresponding to the bottom-left quadrant in Figure 3 – represent a pattern suggesting potential concern. This combination might indicate instances of *cheap talk*, where commitments are stated vaguely or lack concrete detail and verifiable grounding within the report itself. Identifying such internal inconsistencies allows analysts to focus attention on disclosures that appear potentially insubstantial or poorly documented.

The analysis of **external alignment** revealed that higher reporting positivity, on average, correlated with better assessed performance according to the Refinitiv Emission score proxy. This suggests that positive communication is not solely the domain of poorer performers. However, the true value lies in identifying deviations from this general trend. Companies exhibiting high positivity relative to their performance proxy score (upper-left area in Figure 4) display a pattern consistent with established definitions of greenwashing risk (Delmas and Burbano, 2011) – potentially creating an impression of sustainability leadership not fully matched by the external benchmark. While acknowledging the proxy's limitations, this discrepancy analysis provides a data-driven basis for flagging specific companies (e.g., those marked red, and particularly the largest deviator, Siemens Energy) whose optimistic framing merits closer qualitative investigation.

Importantly, these two signals – internal inconsistency and external discrepancy – offer **complementary diagnostic perspectives**. A report might be flagged by one signal but not the other. Using both allows for a more comprehensive risk assessment. For instance, a report could be internally coherent but externally misaligned, or vice versa. This multi-signal approach enhances the ability of stakeholders (investors, regulators, researchers) to prioritize their limited resources, directing in-depth qualitative analysis and verification efforts towards the reports and companies exhibiting the most salient risk indicators. Effective use of these signals can streamline the otherwise daunting task of evaluating large volumes of ESG reporting.

The research also highlights persistent **methodological challenges**. NLP models for specificity or sentiment analysis are not infallible and can misinterpret context, particularly in formal reporting language. Heuristic methods for filtering content or extracting claims, while computationally efficient, inherently limit precision and recall. The exploratory external verification using LOKI confirmed substantial difficulties in reliably automating fact-checking for nuanced ESG claims via standard web search; the tool struggled with context, comparative language, and source reliability, limiting the utility of its outputs without careful manual validation. These limitations underscore that automated tools are best viewed as aids to, rather than replacements for, critical human analysis.

Despite these limitations, the framework provides a valuable advancement by offering structured, data-driven indicators. It moves the assessment of reporting credibility beyond subjective interpretation towards identifying specific, quantifiable patterns associated with potential greenwashing risk in the critical domain of corporate traffic emissions reporting.

6 Conclusion

This research developed and evaluated a machine learning framework to identify potential greenwashing indicators in traffic-related ESG disclosures. By analyzing linguistic style (positivity, specificity) and internal claim substantiation, and correlating these with an external performance proxy, we identified two complementary signals meriting further scrutiny: 1) *Internal inconsistency* (low specificity combined with weak internal claim support), potentially indicating *cheap talk*, and 2) *External discrepancy* (high reporting positivity relative to assessed performance). These quantitative indicators provide stakeholders with a data-driven methodology to prioritize the assessment of reporting credibility, contributing to efforts towards greater transparency and

accountability in this vital sustainability domain.

Limitations

The findings should be interpreted considering several limitations. **Scope and Data:** The analysis focused on English-language reports from German DAX companies (2020-2023) and specifically on traffic-related disclosures, limiting broader generalizability. The lack of a standardized, labeled greenwashing dataset necessitated using proxy indicators. A key limitation is the reliance on the Refinitiv Emission Score as an external performance proxy. This score reflects overall corporate emissions performance and is not specific to traffic-related activities. Comparing communication patterns within the traffic domain to this aggregate score assumes a degree of correlation between general performance and specific reporting, an assumption which requires caution as traffic-specific trends might diverge. Furthermore, any ESG score represents a specific assessment methodology with its own potential biases. **Methodology and Tools:** Standard PDF-to-text conversion potentially introduced noise and missed non-textual information. Resource constraints led to heuristic methods for filtering and claim/proof identification, impacting precision/recall. The accuracy of employed NLP models (e.g., ClimateBERT Specificity, VADER) affects result reliability. Sentiment analysis tools may misinterpret neutral technical language. **Verification Challenges:** Internal claim scores reflect semantic similarity based on heuristically extracted sentences, not guaranteed factual accuracy. Exploratory external verification using the public LOKI interface revealed significant limitations in reliably assessing specific, complex ESG claims against web data due to issues with context, source evaluation, and reasoning capabilities. **Conceptual Ambiguity:** Defining and operationalizing greenwashing remains inherently challenging, limiting objective measurement. The identified indicators signal risk, not definitive proof of intent.

Future Work

The implementation of the EU's Corporate Sustainability Reporting Directive (CSRD), mandating standardized, machine-readable formats (XHTML/iXBRL) and detailed Scope 3 data (European Parliament and Council of the European Union, 2022; European Commission, 2023), offers significant opportunities. Future research should leverage these formats to potentially overcome current text extraction issues and enable more robust analysis of granular data. Applying this framework to CSRD reports will allow investigation into whether reporting patterns evolve under this stricter regulation.

Methodological advancements could involve replacing heuristic steps with more sophisticated NLP techniques for claim extraction (cf. Stammbach et al., 2022) and contextual filtering, possibly using semantic topic modeling. Refining specificity analysis (e.g., distinguishing numerical vs. qualitative detail) could yield richer insights. Addressing the challenge of reliable external verification remains crucial, likely requiring integration of curated authoritative databases or domain-specific knowledge graphs, moving beyond generic web search tools. Expanding this analytical approach to other sectors, regions, and ESG topics will further contribute to understanding and enhancing corporate sustainability reporting credibility.

References

Australian Competition and Consumer Commission. 2023. Greenwashing by businesses in australia. findings of the accc's internet sweep of environmental claims. Last accessed on April 11, 2025.

Tegwen Berthe, Sandrine Nguiakam, and Mathieu Journeau. 2025. Measuring scope 3 emissions: implications challenges for investors. Technical report, Amundi Research Center.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2025. Corporate greenwashing detection in text – a survey. arXiv preprint arXiv:2502.07541.

Yan Chen and Ding Ma. 2024. Detection of greenwashing in esg reports of chinese listed companies based on word2vec and tf-idf. In *Proceedings of the 2024 International Conference on Data Mining*, New York, NY, USA. Association for Computing Machinery.

Magali A Delmas and Vanessa Cuerel Burbano. 2011. The drivers of greenwashing. *California management review*, 54(1):64–87.

European Commission. 2023. Commission delegated regulation (eu) 2023/2772 adopting european sustainability reporting standards (esrs) - set 1. Official Journal of the European Union, L, 2023/2772 (Published 22 December 2023).

European Parliament. 2023. Green claims directive: Protecting consumers from greenwashing. Last accessed on April 11, 2025.

European Parliament and Council of the European Union. 2022. Directive (eu) 2022/2464 as regards corporate sustainability reporting (csrd). Official Journal of the European Union, L 322, p. 15–80.

European Securities and Markets Authority. 2023. Guidelines on greenwashing in sustainability reporting. Last accessed on April 11, 2025.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Valentina Lagasio. 2024. Esg-washing detection in corporate sustainability reports. *International Review of Financial Analysis*, 96:103742.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *Preprint*, arXiv:2401.12566.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2024. Loki: An open-source tool for fact verification. *Preprint*, arXiv:2410.01794.

Janna Lipenkova, Guang Lu, and Susie Rao. 2023. Detecting greenwashing signals through a comparison of esg reports and public media.

LSEG Data & Analytics. 2024. Esg scores | lseg. LSEG webpage. Last accessed on April 11, 2025.

J. Plucinska. 2023. Greenwashing cases against airlines in europe and the us. Reuters. Last accessed on April 11, 2025.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

D. Robinson. 2022. 10 companies called out for greenwashing. Published on Earth.org. Last accessed on April 13, 2025.

Ujjwal Sharma, Stevan Rudinac, Joris Demmers, Willemijn van Dolen, and Marcel Worring. 2024. Greenscreen: A multimodal dataset for detecting corporate greenwashing in the wild. In *International Conference on Multimedia Modeling*, pages 96–109. Springer.

P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley. 2022. *Chapter 10: Transport*. Cambridge University Press.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *arXiv preprint*.

Dimitar Trajanov, Georgi Lazarev, Lubomir Chitkushev, and Irena Vodenska. 2023. Comparing the performance of chatgpt and state-of-the-art climate nlp models on climate-related text classification tasks. In *Proceedings of the 4th International Conference on Environmental Design (ICED2023)*.

UK Competition and Markets Authority. 2021. Global sweep finds 40% of firms' green claims could be misleading. Last accessed on April 11, 2025.

Avalon Vinella, Margaret Capetz, Rebecca Pattichis, Christina Chance, Reshmi Ghosh, and Kai-Wei Chang. 2024. Leveraging language models to detect greenwashing. *Preprint*, arXiv:2311.01469.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. *Preprint*, arXiv:2110.12010.

Yue Zhao, Leon Kroher, Maximilian Engler, and Klemens Schnattinger. 2023. Detecting greenwashing in the environmental, social, and governance domains using natural language processing. In *KDIR*, pages 175–181.

Enhancing Multilingual LLM Pretraining with Model-Based Data Selection

Bettina Messmer[*1], **Vinko Sabolčec**[*1], **Martin Jaggi**[1]
[1]EPFL
Correspondence: firstname.lastname@epfl.ch

Abstract

Dataset curation has become a basis for strong large language model (LLM) performance. While various rule-based filtering heuristics exist for English and multilingual datasets, model-based filtering techniques have primarily focused on English. To address the disparity stemming from limited research on non-English languages, we propose a model-based filtering framework for multilingual datasets that aims to identify a diverse set of structured and knowledge-rich samples. Our approach emphasizes transparency, simplicity, and efficiency, leveraging Transformer- and FastText-based classifiers to ensure the broad accessibility of our technique and data. We conduct comprehensive ablation studies on the FineWeb-2 web crawl dataset across diverse language families, scripts, and resource availability to demonstrate the effectiveness of our method. Training a 1B-parameter Llama model for 70B and 119B tokens, our approach can match the baseline MMLU score with as little as 15% of the training tokens, while also improving across other benchmarks. These findings provide strong evidence for the generalizability of our approach to other languages. As a result, we extend our framework to 20 languages, for which we release the refined pretraining datasets.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive performance improvements when trained on increasingly larger datasets and model sizes (Brown et al., 2020). While Brown et al. (2020) already observed the importance of using a cleaned version of Common Crawl for improved performance, the high cost of LLM training has further motivated research into better pretraining quality filters.

Deduplication and heuristic-based dataset cleaning have become standard practices in data curation (Rae et al., 2021; Raffel et al., 2020; De Gibert et al., 2024). These quality filters are often complemented by additional filters, such as the removal of personally identifiable information (PII) (Penedo et al., 2024a) or model-based toxicity filtering (Soldaini et al., 2024). Recently, model-based filtering has also emerged as a promising method for quality filtering. The release of FineWeb-Edu (Penedo et al., 2024a) demonstrated that pretraining on just 10% of the tokens (38B) from an English dataset filtered using a model-based approach can achieve performance comparable to models trained on 350B tokens of unfiltered data. Moreover, when trained on equivalent amounts of data, this model largely outperforms the baseline. Concurrently, the release of DataComp-LM (DCLM) (Li et al., 2024b) showed that competitive performance can be achieved using a simple and efficient model-based approach, namely a FastText (Joulin et al., 2017) classifier trained on a carefully selected training dataset.

However, these recent advances have primarily focused on English data. This emphasis risks further widening the disparity in LLM performance between languages, as less than half of internet content is written in English[1]. To address this concern, we aim to extend model-based filtering frameworks to multilingual datasets. While model perplexity-based filtering is commonly applied to multilingual datasets (Wenzek et al., 2019; Laurençon et al., 2022; Nguyen et al., 2023), the current state-of-the-art, FineWeb-2 (Penedo et al., 2024c), primarily relies on heuristic-based filters. In this work, we focus on model-based filtering with a quality definition that emphasizes: 1) structured data and 2) knowledge-rich data samples, to enhance multilingual pretraining datasets.

To achieve this, we leverage embedding-based classification models. Firstly, we adopt the Fast-

[*]Equal contribution

[1]w3techs.com/technologies/overview/content_language

Text quality filtering approach from DCLM to develop a unified framework for multilingual datasets that span diverse language families, scripts, and resource availability, focusing on Chinese, German, French, Arabic, and Danish as representative languages for our experiments. Additionally, we extend this embedding-based approach by incorporating Transformer (Vaswani et al., 2023) embeddings, specifically XLM-RoBERTa (Conneau et al., 2020), for filtering.

In summary, our contributions are as follows:

- We propose a transparent, simple, and unified framework for multilingual model-based filtering at web scale, enabling data curation across diverse language families, scripts and resource availability.
- We present comprehensive per-language ablation studies of embedding-based multilingual quality filtering on top of the FineWeb-2 dataset (Penedo et al., 2024c), achieving performance comparable to the baseline while using as little as 15% of the tokens.
- We evaluate the impact of different training datasets for data selection classifiers on the downstream performance of LLMs.
- We release the refined pretraining dataset filtered using our proposed framework, FineWeb2-HQ[2], which spans 20 languages and is distributed under the *Open Data Commons Attribution License (ODC-By) v1.0*, along with the codebase[3], to advance multilingual language modeling.

2 Related Work

Data Curation. In order to pretrain LLMs on a large amount of diverse texts, Common Crawl[4] is often used as the base dataset. However, early works already observed that performing quality filtering on Common Crawl is crucial for model performance (Brown et al., 2020). There exist various data curation approaches, such as deduplication (Lee et al., 2022), PII removal (Subramani et al., 2023), or toxicity filtering (Arnett et al., 2024). Another important aspect is quality filtering of the documents. For this, the definition of quality is an important aspect. A common approach is to use heuristics to remove documents outside of the target distribution, such as filtering based on average word length, existence of punctuation, or document length (Rae et al., 2021; Raffel et al., 2020). Another approach is to define model-based filters, where research has focused on perplexity measure of the text (Wenzek et al., 2019; Marion et al., 2023; Ankner et al., 2024), distributional similarity measures (Brown et al., 2020; Xie et al., 2023; Li et al., 2024b) and LLM-based quality assessment (Gunasekar et al., 2023; Wettig et al., 2024; Sachdeva et al., 2024; Penedo et al., 2024a). In this work, we build upon previous curated datasets based on heuristic filtering, specifically FineWeb-2 (Penedo et al., 2024c), and focus on model-based distributional similarity filtering for structured and knowledge-rich documents relying on textual embedding representation.

Curated English datasets. One of the early curated datasets was C4 (Raffel et al., 2020), followed by MassiveText (Rae et al., 2021). RefinedWeb (Penedo et al., 2023) was an important step forward, demonstrating that filtered web data can outperform selected high-quality data sources. While these datasets have not been made fully publicly available, their filtering techniques have been expanded upon in recent fully public datasets, such as Dolma (Soldaini et al., 2024), FineWeb, and FineWeb-Edu (Penedo et al., 2024a). While FineWeb primarily relies on filter heuristics for data quality, Dolma adopts model perplexity filtering. FineWeb-Edu takes model-based filtering a step further and relies on LLM-based quality assessment. Similarly, a concurrent work, DCLM, has achieved competitive performance using FastText (Joulin et al., 2017) classifier trained on a carefully selected training dataset. In this work we adapt and extend this approach to the multilingual context.

Curated Multilingual Datasets. Analogously to the English datasets, there have been efforts in the multilingual space. An influential work has been CCNet (Wenzek et al., 2019), whose language identification and model perplexity filter for data quality has been re-used in later datasets. Again, while CCNet was not published directly, but rather provided the tools for data cleaning, RedPajama (Together Computer, 2023) is a prominent multilingual dataset relying on these filtering techniques. While RedPajama offers data in 5 European languages, other datasets, such as OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021; Abadji et al., 2022), mC4 (Xue et al., 2021), ROOTS (Lau-

[2] huggingface.co/datasets/epfml/FineWeb2-HQ
[3] github.com/epfml/fineweb2-hq
[4] commoncrawl.org

rençon et al., 2022), MADLAD-400 (Kudugunta et al., 2023), CulturaX (Nguyen et al., 2023), and HPLT (de Gibert et al., 2024), focus on expanding beyond, spanning a variety of language families and scripts. While they offer refined datasets for hundreds of languages, FineWeb-2 (Penedo et al., 2024c) pushes the limit to thousands of languages and further improves the performance. Our work also focuses on filtering quality samples across various language families and scripts. However, we limit our scope to 20 languages, as the number of documents drops quickly and there is trade-off between retaining a sufficient number of pretraining tokens and ensuring data quality (Muennighoff et al., 2023; Held et al., 2025). In our results, we observe the greatest benefits using stricter data filtering.

Multilingual Embedding Models. Early word embedding models like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) lacked contextual understanding. FastText (Bojanowski et al., 2017) built upon them and improved performance by incorporating subword information. Transformer (Vaswani et al., 2023) models like BERT (Devlin et al., 2019) and GPT (Radford et al., 2018) then revolutionized the field with context-aware embeddings. Multilingual models like mBERT, XLM (Lample and Conneau, 2019), and XLM-RoBERTa (Conneau et al., 2020) further advanced cross-lingual understanding, with recent open-source LLMs pushing performance even higher (Llama Team, 2024; Mistral AI, 2025). Using such models, documents as well as representative samples can be mapped into a shared embedding space to estimate their similarity. Focusing on transparency, simplicity and efficiency in our work, we use FastText and XLM-RoBERTa for our filtering, and analyze the trade-off between computational complexity and filtering performance.

Multilingual Evaluation. Evaluating LLMs requires diverse benchmarks testing linguistic and cognitive abilities like reading comprehension, reasoning, and knowledge. While English benchmarks like MMLU (Hendrycks et al., 2020) and ARC (Clark et al., 2018) exist, other languages often use translations from English, e.g., XNLI (Conneau et al., 2018) and machine-translated version of MMLU (Lai et al., 2023). However, translations can be problematic, failing to capture cultural nuances or introducing "translationese" (Romanou et al., 2024). Recent work by Romanou et al. (2024); Singh et al. (2024a) emphasizes the need for culturally sensitive, natively collected benchmarks. Task difficulty and task formulation also impact model performance when trained for shorter durations (Kydlíček et al., 2024). In our work, we follow the recent evaluation tasks selection and methodology by Kydlíček et al. (2024) to assess our model-based filtering approaches across multiple languages.

3 Methods

In this work, we present our model-based filtering approaches. Our methodology is structured into two key components: 1) we select suitable training datasets, aiming to identifying a diverse set of structured and knowledge-rich samples and 2) we describe the different models, namely FastText and Transformer embedding-based filters, used to capture and leverage these characteristics.

3.1 Classifier Training Dataset

Representative Sample Selection. Our goal is to identify a diverse set of structured and knowledge-rich samples, especially within a multilingual context. We define two criteria for our training datasets: 1) the samples must be informative and well-structured and 2) the datasets must be available in multiple languages. While some multilingual benchmark datasets meet these criteria precisely, it is important to note that we do not train the LLM directly on this data. Instead, we train a proxy model to assess pretraining data quality. Nevertheless, we must remain cautious about potentially increased pretraining data contamination stemming from this approach, as discussed in Appendix B.6.

Based on our criteria, we selected the following datasets as representative examples:

- ***Aya Collection.*** A prompt completion dataset comprising ∼514M samples covering a wide variety of tasks, generated using instruction-style templates in 101 languages (Singh et al., 2024b).
- ***Aya Dataset.*** Human-annotated instruction fine-tuning dataset consisting of ∼202K prompt-completion pairs in 65 languages (Singh et al., 2024b).
- ***MMLU.*** Originally for English language, the dataset contains ∼14K multiple-choice knowledge questions in diverse subjects and areas (Hendrycks et al., 2020). Multilingual version was translated into 14 languages by professional translators (OpenAI, 2024).

- ***OpenAssistant-2.*** The dataset contains ∼14K user-assistant conversations with multiple messages in 28 languages (Fischer et al., 2024).
- ***Include-Base-44.*** Multiple-choice questions focused on general and regional knowledge, as well as reasoning, extracted from academic and professional exams. Spanning 44 languages, it includes a total of ∼23K samples (Romanou et al., 2024).

Representative Sample Collection. *MMLU* and *Include-Base-44* are highly curated benchmark datasets, containing structured, knowledge-rich samples. The *Aya Dataset* is human-curated, while *OpenAssistant-2* is partially human-curated and partially generated by large language models (LLMs). In contrast, the *Aya Collection* consists of various AI-generated samples without quality guarantee, though it represents the largest and most multilingual corpus of the five.

To address this quality difference, we create two *Multilingual Knowledge Collection (MKC)* configurations:

- ***MKC***: *Include-Base-44*, *OpenAssistant-2*, *MMLU*, and the *Aya Dataset*
- ***MKC***$^+$: *MKC* and the *Aya Collection*

This allows us to evaluate the trade-off between data quality and scale.

Dataset Creation. For our model-based filtering approaches, our goal is to identify documents from the pretraining dataset that are most similar to our representative samples, with the notion of similarity determined by the specific classifier used. We can measure the similarity to our training dataset directly, for example, by computing the cosine similarity to our training samples in the embedding space. Alternatively, following the approach of Li et al. (2024b), the task can be framed as a binary classification problem, with the representative samples as the positive class. For the negative class, we can simply subsample documents from our pretraining dataset, under the assumption that the majority of these documents are neither well-structured nor knowledge-rich. We use both approaches for our classifiers.

To create the binary classification training dataset, we selected 80K random examples from the training set (*MKC* or *MKC*$^+$) as positive samples and 80K random examples from FineWeb-2 as negative samples. For smaller datasets, namely *Include-Base-44* and *OpenAssistant-2*, the entire dataset was used. The same training dataset was utilized across all model-based filtering approaches, disregarding negative samples when unnecessary. Additionally, we created a training dataset for each language individually to avoid leaking language-specific biases to data of other languages.

Sample Pre-processing. We applied no pre-processing to the FineWeb-2 (negative) samples but performed minimal pre-processing on the representative (positive) samples. For instance, in datasets like *MMLU* or *OpenAssistant-2*, we concatenated various sample components, namely the question-answer pairs, and user-assistant conversation parts to obtain the full samples. For the *Aya Collection*, we resolved encoding issues in non-Latin languages and removed samples containing <*unk*> tokens, which were particularly prevalent in Arabic data (37.1%) and less present in the data of all 20 languages that we used (5.5%).

3.2 FastText-based Filtering (FT)

To efficiently process datasets with over 100 million documents (Penedo et al., 2024c), similar to DCLM (Li et al., 2024b), we used a binary FastText classifier (Joulin et al., 2017). This classifier runs on CPU and can be easily deployed across multiple cores, for example using DataTrove (Penedo et al., 2024b).

We trained our FastText classifier on the processed training set using 2-gram features (4-gram for Chinese). Additional details about the training process are given in Appendix A.1. These classifiers were then used to assign scores to all documents in the pretraining dataset. To filter the dataset, we applied a score threshold based on the desired retention percentage of documents. This approach balances dataset size and the predicted quality of the samples.

3.3 Transformer Embedding-based Filtering

To leverage rich semantic information based on contextual relationships, we utilized the Transformer model embeddings. Specifically, we selected a pretrained XLM-RoBERTa base model (Conneau et al., 2020) due to its support of 100 languages, a relatively small size of approximately 279M parameters, and its transparent training procedure. This choice enabled us to process web-scale data efficiently without being restricted to a single language and to align with our commitment to open science.

To retain general embeddings that can be reused

across methods, we opted against fine-tuning the model. For each document from our datasets, we computed the 768-dimensional embedding by mean pooling the embeddings of the output sequence. Since the model has a fixed maximum sequence length of 512 tokens, we considered only the first 512 tokens of each document, assuming they are representative of the entire document.

After computing the embeddings of our corpora, we experimented with two methods: 1) classification of embeddings using a multi-layer perceptron and 2) cosine similarity between the embeddings. As in the FastText approach, we scored each document and applied a threshold to retain the desired percentage of the highest-scoring documents.

Multi-Layer Perceptron (MLP). We trained a single-hidden-layer neural network with a hidden dimension of 256, the ReLU activation function, a dropout rate of 20%, and the sigmoid function on the output. The network was trained for 6 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with a constant learning rate 0.0003 and binary cross-entropy loss. We computed document scores using the output layer of the MLP model, which used XLM-RoBERTa document embeddings as input.

Cosine Similarity (CS). We computed the document scores as the maximum cosine similarity between its embeddings and a set of K randomly sampled positive sample embeddings. We experimented with varying values of K, including 1024, 2048, 4096, 8192, and 16384. However, we did not observe a significant differences in the documents with high scores across these variations when manually inspecting the data. To strike a balance between the diversity of the positive samples and computational efficiency, we chose $K = 8192$ for our experiments.

4 Experiments

4.1 Experimental Setup

Technical Details. We evaluate 1B-parameter Llama models (Llama Team, 2024) to demonstrate the effectiveness of our model-based filtering approaches. The models are trained on either 70B or 119B tokens, balancing token quality and diversity. The smaller dataset (70B tokens) exposes the model to each token at most once (with a few exceptions where some tokens appear twice due to the dataset containing fewer than 70B tokens). The larger dataset (119B tokens) simulates longer training, resulting in increased token repetition. Training utilizes the HuggingFace Nanotron library (Hugging Face, 2024a) with the AdamW optimizer (Loshchilov and Hutter, 2019) and a WSD learning rate schedule (Hägele et al., 2024).

To minimize the need for costly hyperparameter tuning, we maintain a consistent setup across all experiments. Specifically, we adopt the DeepSeek scaling law (DeepSeek-AI et al., 2024) with a batch size of 1.6M tokens, learning rate of 0.0008, and 2000 warmup steps. We provide our Nanotron config in Appendix A.2.

As base dataset, we use FineWeb-2 (Penedo et al., 2024c), which has been shown to provide a strong baseline across a variety of languages. Since FineWeb-2 is globally deduplicated, we rehydrate both filtered and unfiltered data using the hyperparameters recommended by Penedo et al. (2024c).

To validate our method on English, we use three datasets: FineWeb (Penedo et al., 2024a) as the baseline, along with FineWeb-Edu (Penedo et al., 2024a) and DCLM (Li et al., 2024b), both of which represent the current state-of-the-art. Tokenization is performed using the multilingual Mistral v3 (Tekken) tokenizer (Mistral AI, 2024). All experiments are conducted using 80 NVIDIA GH200 chips.

Evaluation. Our evaluation prioritizes a diverse range of tasks to ensure the models retain well-rounded capabilities, rather than focusing exclusively on knowledge-based tasks. Specifically, we include tasks covering reading comprehension, general knowledge, natural language understanding, common-sense reasoning, and generative tasks in the target language. To evaluate our approach, we use the HuggingFace LightEval library (Fourrier et al., 2023).

For French, Chinese, and Arabic, we utilize the FineTasks (Kydlíček et al., 2024) multilingual evaluation suite, which is designed to provide meaningful signals even for models trained in the order of 100B tokens. We select analogous tasks for German and Danish. For English, we rely on the SmolLM tasks suite (Hugging Face, 2024b). A complete list of tasks and their evaluation metrics for each language is provided in Appendix C.

Model Selection. We follow the approach used in FineTasks for filter selection, computing a global rank score across individual metrics and languages to determine the optimal approach.

4.2 Experimental Results & Discussion

4.2.1 Model Selection

In Section 3, we introduced several model-based filtering approaches. *But which of these performs the best?* We evaluate which combination of our defined classifier training datasets (*MKC* or *MKC*$^+$) and filtering methods (*FT*, *MLP* or *CS*) achieve the highest performance. Table 1 presents the overall ranking across our representative language selection (Chinese, German, French, Arabic, Danish) and training runs of 70B and 119B tokens. Analogous to the DCLM filtering recipe (Li et al., 2024b), the results are based on a dataset that retains 10% of the documents for the high-resource datasets (Chinese, German, French) and keeps 56% and 65% of the documents for the lower-resource languages (Arabic and Danish, respectively). These percentages maintain approximately 70B tokens, under the assumption of uniform token distribution across documents. We also exclude approaches that use *MKC* for training on Danish, as it lacks sufficient training data. For detailed, per-language results, please refer to Appendix B.1.

Table 1 demonstrates that *MLP MKC*$^+$ approach outperforms all other approaches. Interestingly, the high- and low-scored samples presented in Appendix D align with the observed rankings. Figure 1 further highlights the strong performance of *MLP MKC*$^+$, particularly for high-resource languages, where it largely outperforms the baseline. For lower-resource languages—where less data was filtered—the performance gains are less pronounced. Notably, *FT* filtering is also competitive. Given the computational expense of XLM-RoBERTa embeddings, FastText can be a promising alternative in resource-constrained setups.

Approach	Average Rank
MLP MKC$^+$	4.35
MLP MKC	6.11
FT MKC$^+$	7.17
FT MKC	8.04
CS MKC	8.10
Baseline	8.72
CS MKC$^+$	8.79

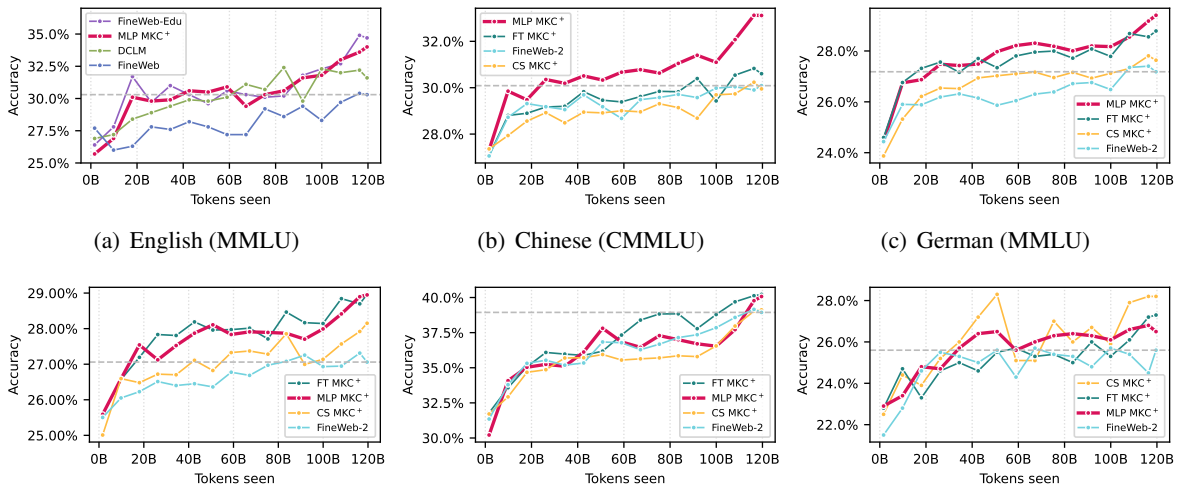
Table 1: Benchmark performance comparison (average rank) between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 70B and 119B tokens were consumed.

4.2.2 Threshold Selection

In Section 4.2.1, we base our model selection on experiments that retain top 10% of the data for high-resource languages. *But is this the optimal threshold?* Following the methodology of Li et al. (2024b), we analyze the impact of varying filter strengths on performance for Chinese, German, and French, using our *MLP* and *FT* filtering methods. The results are summarized in Table 2, with a comprehensive analysis, including results for *CS*, provided in Appendix B.2 (Table 11). Consistent with their findings, we observe that retaining top 10% of the data is a competitive threshold, particularly for approaches using the *MKC*$^+$ dataset. Interestingly, approaches using *MKC* perform better with higher retention. Motivated by the observed bias in certain approaches favoring the selection of shorter documents, we examine how this bias interacts with performance when retaining more documents. As demonstrated in Figure 2 for German, Appendix B.2 for other languages, and the retained token counts in Table 12, the *MLP MKC* approach shows a tendency to retain shorter documents, while achieving higher performance with an increased number of retained documents. In contrast, the *CS* and *FT* filtering methods present mixed results, suggesting that the optimal threshold selection may be influenced by additional factors.

Approach	Threshold	Average Rank
MLP MKC$^+$	10%	8.85
MLP MKC$^+$	15%	9.44
MLP MKC	20%	11.37
MLP MKC	15%	11.70
MLP MKC	10%	11.95
MLP MKC$^+$	20%	11.97
FT MKC$^+$	10%	13.92
FT MKC	15%	14.62
FT MKC	10%	14.74
FT MKC	20%	15.62
FT MKC$^+$	15%	16.27
FT MKC$^+$	20%	16.51
Baseline	–	18.55

Table 2: Benchmark performance comparison (average rank) between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*) trained on *MKC*$^+$ or *MKC*, retaining top 10%, 15% or 20% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated for Chinese, German and French after 70B and 119B tokens were consumed.



(a) English (MMLU) (b) Chinese (CMMLU) (c) German (MMLU)

(d) French (MMLU) (e) Arabic (MMLU) (f) Danish (ARC)

Figure 1: Benchmark performance comparison (accuracy) during training on 119B tokens between the baseline methods (FineWeb, DCLM, FineWeb-Edu, and FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*), trained on *MKC*$^+$. When using our approaches, the data retention rates are set to 10% for English, Chinese, German, and French, 56% for Arabic, and 65% for Danish. For English, Chinese, German, and French, baseline-level performance is observed around 20B tokens consumed (16.7% of the total).

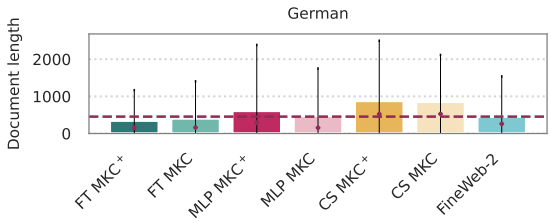


Figure 2: Comparison of average document length and standard deviation in FineWeb-2 before and after filtering using one of our approaches retaining top 10% of the documents. The average document length of FineWeb-2 is represented as a red horizontal line, while the medians are shown as red dots. Document length is measured based on number of space-separated tokens.

4.2.3 Training Data Analysis

The experiments in Sections 4.2.1 and 4.2.2 are based on the training datasets *MKC* and *MKC*$^+$. *But is the diversity introduced by combining various base datasets truly necessary?* We evaluate the impact of each base dataset individually and compare it to the combined *MKC*$^+$ dataset. For this ablation study, we use our best filtering method (*MLP* with a top 10% retention) and train the models on 30B tokens. This token count is chosen to match the size of the smallest filtered dataset, ensuring consistency across comparisons. The results, presented in Table 3, show that despite the absence of a quality guarantee for all samples in the *Aya Collection*, this dataset yields strong performance, making our approach applicable for various languages. Overall, we observe that the diversity resulting from combining all individual training datasets gives the best results.

Interestingly, models trained exclusively on *Include-Base-44* and *OpenAssistant-2* perform worse overall than the baseline. This may be due to the nature of these datasets. For instance, *Include-Base-44* is relatively small and domain-specific, e.g., consisting primarily of driving license exam questions in its German subset. Similarly, *OpenAssistant-2* includes a limited number of samples, with fewer than 2K positive samples per training set, which likely negatively impacts classifier performance. Again, we relate model performance to the average document length bias in Appendix B.3 and confirm the findings from Section 4.2.2, suggesting that factors beyond the retained document length bias may influence performance.

4.2.4 Impact on multilingual model training

Although not the primary focus of our work, we believe that refined datasets can contribute to advancing the performance of multilingual models. To investigate this, we conducted an ablation study by training a 1B-parameter model on 595B tokens (5×119B), covering all five languages: Chinese, German, French, Arabic and Danish. We trained two models—the first one using our filtered FineWeb-2 dataset and the second one using unfiltered FineWeb-2 data. We then compared these

Dataset	Average Rank
MKC$^+$	2.52
Aya Collection	2.91
Aya Dataset	3.17
MMLU	3.57
Baseline	4.09
OpenAssistant-2	4.53
Include-Base-44	5.42

Table 3: Benchmark performance comparison (average rank) between the baseline (FineWeb-2) and the *MLP* filtering method trained on either *MKC*$^+$ as a whole or its individual dataset components, retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish. The average rank is computed across FineTasks performance of 1B-parameter models trained on each language with 30B tokens.

results for each language against their monolingual counterparts trained on 119B tokens.

The results for French are presented in Table 4. We observe that the multilingual LLM outperforms its monolingual counterpart on our filtered datasets, whereas the monolingual model achieves better performance than the multilingual model on the FineWeb-2 dataset. This trend is consistent across all languages except Chinese. Detailed results for the other languages are provided in Appendix B.4.

Dataset	Ours$_M$	Ours	FW-2	FW-2$_M$
Average Rank	1.8333	2.0556	3.0000	3.1111
Belebele	**0.3667**	0.3533	0.3444	0.3511
HellaSwag	0.5270	**0.5380**	0.5180	0.4970
X-CSQA	0.2740	0.2740	**0.2870**	0.2750
XNLI 2.0	**0.7660**	0.7400	0.7180	0.7330
FQuAD	**0.3212**	0.2803	0.2401	0.2459
MMLU	0.2841	**0.2895**	0.2706	0.2735
Mintaka	0.0456	0.0438	**0.0712**	0.0579
X-CODAH	**0.2900**	0.2667	0.2633	0.2567
ARC (Challenge)	0.2970	**0.3180**	0.2850	0.2670

Table 4: Benchmark performance comparison for French of multilingual LLMs (M) trained on FineWeb-2 or the refined dataset using our *MLP MKC*$^+$ approach (retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish) trained on 595B tokens, against their monolingual counterparts trained on 119B tokens. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

In addition to these experiments, we explore how replay of original FineWeb-2 data affects performance in Appendix B.5, measure data contamination in Appendix B.6, and validate our approach on English data in Appendix B.7.

5 Dataset Information

Based on the results of our experiments, we create the dataset, named FineWeb2-HQ, by filtering all available FineWeb-2 data (version `2.0.1`) in 20 languages using the *MLP MKC*$^+$ approach with 10% retention rate. The statistics of the resulting dataset are presented in Table 5. We release the dataset under the *Open Data Commons Attribution License (ODC-By) v1.0* license on Hugging Face[5]. The main use case of the dataset is LLM pretraining, however, the dataset may also be used for other natural language processing tasks.

Language	Number of documents	Disk size
Russian	55,220,956	1.2TB
Chinese	54,211,986	784GB
German	43,095,728	618GB
Spanish	40,057,637	515GB
Japanese	34,185,427	393GB
French	32,248,772	483GB
Italian	21,180,304	269GB
Portuguese	18,135,468	222GB
Polish	13,384,885	168GB
Dutch	12,920,963	160GB
Indonesian	8,911,149	125GB
Turkish	8,578,808	100GB
Czech	5,995,459	104GB
Arabic	5,560,599	94GB
Persian	5,107,187	69GB
Hungarian	4,527,332	79GB
Swedish	4,382,454	61GB
Greek	4,346,440	84GB
Danish	4,082,751	61GB
Vietnamese	4,003,956	59GB

Table 5: Statistics (number of documents and disk size) of the dataset resulting from filtering FineWeb-2 using the *MLP MKC*$^+$ approach with 10% retention rate in 20 languages.

6 Conclusion

In this work, we introduced a novel framework for model-based filtering of web-scale multilingual pretraining datasets, demonstrating consistent improvements on LLM benchmarks across a wide range of languages. Our Transformer embedding-based classifier, *MLP MKC*$^+$, outperforms state-of-the-art methods on both English and multilingual datasets, even when decontaminating the datasets or using them for training multilingual LLMs. This

[5] huggingface.co/datasets/epfml/FineWeb2-HQ

demonstrates that simple classifiers can achieve competitive results. While our FastText-based filtering approach performed well and shows promise in resource-constrained setups, *MLP MKC*$^{+}$ consistently outperformed all other methods and can be easily scaled to other languages. These results motivate us to expand our framework to 20 languages and release the corresponding refined pretraining datasets and our code, contributing to the advancement of multilingual language modeling.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, arXiv:2201.06642.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L. Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *Preprint*, arXiv:2405.20541.

Catherine Arnett, Eliot Jones, Ivan P. Yamshchikov, and Pierre-Carl Langlais. 2024. Toxicity of the Commons: Curating Open-Source Pre-Training Data. *arXiv preprint arXiv:2410.22587*.

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)*, page 749–775. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Preprint*, arXiv:1607.04606.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and

Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Preprint*, arXiv:2003.05002.

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. *Preprint*, arXiv:1803.05457.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. 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)*. Association for Computational Linguistics.

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.

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, and 1 others. 2024. A new massive multilingual dataset for high-performance language technologies. *arXiv preprint arXiv:2403.14009*.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *Preprint*, arXiv:2401.02954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. FQuAD: French Question Answering Dataset. *Preprint*, arXiv:2002.06071.

Sophie Fischer, Federico Rossetto, Carlos Gemmell, Andrew Ramsay, Iain Mackie, Philip Zubel, Niklas Tecklenburg, and Jeffrey Dalton. 2024. Open assistant toolkit–version 2. *arXiv preprint arXiv:2403.00586*.

Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. 2023. LightEval: A lightweight framework for LLM evaluation.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. 2024. Scaling laws and compute-optimal training beyond fixed training durations. *arXiv preprint arXiv:2405.18392*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. *Preprint*, arXiv:2011.03080.

William Held, Bhargavi Paranjape, Punit Singh Koura, Mike Lewis, Frank Zhang, and Todor Mihaylov. 2025. Optimizing pretraining data mixtures with llm-estimated utility. *arXiv preprint arXiv:2501.11747*.

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*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Preprint*, arXiv:2305.08322.

Hugging Face. 2024a. Nanotron. Accessed 30 Jan. 2025.

Hugging Face. 2024b. SmolLM - blazingly fast and remarkably powerful. Accessed 30 Jan. 2025.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *Preprint*, arXiv:2402.12840.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *Preprint*, arXiv:2309.04662.

Hynek Kydlíček, Guilherme Penedo, Clémentine Fourier, Nathan Habib, and Thomas Wolf. 2024. FineTasks: Finding signal in a haystack of 200+ multilingual tasks. Accessed 30 Jan. 2025.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Preprint*, arXiv:1901.07291.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, and 1 others. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. *Preprint*, arXiv:1910.07475.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, and 1 others. 2024b. DataComp-LM: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. 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 1274–1287, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2021b. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Llama Team. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *Preprint*, arXiv:2309.04564.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

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

Mistral AI. 2024. v3 (tekken) tokenizer. Accessed 30 Jan. 2025.

Mistral AI. 2025. Mistral small 3. Accessed 30 Jan. 2025.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *Preprint*, arXiv:1906.05394.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *Preprint*, arXiv:2305.16264.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

OpenAI. 2024. MMMLU. Accessed 30 Jan. 2025.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024a. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. DataTrove: large scale data processing. Accessed 30 Jan. 2025.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024c. FineWeb2: A sparkling update with 1000s of languages. Accessed 30 Jan. 2025.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172.

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

Pluto-Junzeng. 2019. pluto-junzeng/chinesesquad. Accessed 30 Jan. 2025.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *Preprint*, arXiv:2402.09668.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv preprint arXiv:1907.10641*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering. *Preprint*, arXiv:2210.01613.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024a. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *Preprint*, arXiv:2106.12066.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset. Accessed 30 Jan. 2025.

Ankit Kumar Upadhyay and Harsit Kumar Upadhya. 2023. Xnli 2.0: Improving xnli dataset and performance on cross lingual understanding (xlu). *Preprint*, arXiv:2301.06527.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *Preprint*, arXiv:2402.09739.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *Preprint*, arXiv:2302.03169.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Additional Experimental Details

A.1 FastText Training Details

The FastText classifier was trained on the processed training set using `2-grams`, a `minCount` of 1, and the `softmax` loss function. All other parameters were automatically tuned using the FastText library. For Chinese, fixed parameters were used: 30 training epochs and a learning rate of 0.1 to ensure training stability. Additionally, `4-grams` and a `minCount` of 0 were selected based on manual evaluation of the results.

Prior to training the FastText models, we pre-processed the training data by removing `newlines`.

A.2 Nanotron Configuration

To facilitate the reproducibility of our model training, we provide the Nanotron (Hugging Face, 2024a) configuration used in our experiments.

```yaml
checkpoints:
  checkpoint_interval: 1000
  checkpoints_path: checkpoints/
  checkpoints_path_is_shared_file_system: false
  resume_checkpoint_path: null
  save_initial_state: false
data_stages:
  - data:
      dataset:
        dataset_folder: template
      num_loading_workers: 1
      seed: 42
    name: General purpose training (Single dataset)
    start_training_step: 1
general:
  benchmark_csv_path: null
  consumed_train_samples: null
  ignore_sanity_checks: true
  project: template
  run: template
  seed: 42
  step: null
lighteval: null
logging:
  iteration_step_info_interval: 1
  log_level: info
  log_level_replica: info
model:
  ddp_bucket_cap_mb: 25
  dtype: bfloat16
  init_method:
    std: 0.025
  make_vocab_size_divisible_by: 1
  model_config:
    bos_token_id: 1
    eos_token_id: 2
    hidden_act: silu
    hidden_size: 1536
    initializer_range: 0.02
    intermediate_size: 6144
    is_llama_config: true
    max_position_embeddings: 1024
    num_hidden_layers: 24
    num_attention_heads: 16
    num_key_value_heads: 16
    pad_token_id: null
    pretraining_tp: 1
    rms_norm_eps: 1.0e-06
    rope_scaling: null
    tie_word_embeddings: true
    use_cache: true
    vocab_size: 131072
```

```yaml
optimizer:
  optimizer_factory:
    adam_beta1: 0.9
    adam_beta2: 0.95
    adam_eps: 1.0e-08
    name: adamW
    torch_adam_is_fused: true
  learning_rate_scheduler:
    learning_rate: 0.0008
    lr_decay_starting_step: 61001 # for 119B tokens (36001 for 70B tokens, 15001 for 30B tokens)
    lr_decay_steps: 12000 # for 119B tokens (7000 for 70B tokens, 4000 for 30B tokens)
    lr_decay_style: 1-sqrt
    lr_warmup_steps: 2000
    lr_warmup_style: linear
    min_decay_lr: 0.00
  zero_stage: 0
  clip_grad: 1.0
  weight_decay: 0.1
  accumulate_grad_in_fp32: true
parallelism:
  dp: 80
  expert_parallel_size: 1
  pp: 1
  pp_engine: 1f1b
  tp: 1
  tp_linear_async_communication: true
  tp_mode: REDUCE_SCATTER
profiler: null
tokenizer:
  tokenizer_max_length: null
  tokenizer_name_or_path: mistralai/Mistral-Nemo-Base-2407
  tokenizer_revision: null
tokens:
  batch_accumulation_per_replica: 1
  limit_test_batches: 0
  limit_val_batches: 0
  micro_batch_size: 20
  sequence_length: 1024
  train_steps: 73000 # for 119B tokens (43000 for 70B tokens, 19000 for 30B tokens)
  val_check_interval: -1
```

B Additional Results

B.1 Model Selection - Per Language Results

For completeness, we present the individual benchmark results of the 1B-parameter model trained on 119B tokens for each language in the following tables: Table 6 for Chinese, Table 7 for French, Table 8 for German, Table 9 for Arabic, and Table 10 for Danish.

Approach	*MLP MKC*$^+$	*MLP MKC*	*CS MKC*	*FT MKC*	*FT MKC*$^+$	Baseline	*CS MKC*$^+$
Average Rank	1.7333	2.4333	4.0667	4.0667	4.4667	5.2333	6.0000
AGIEval	**0.2995**	0.2948	0.2897	0.2919	0.2817	0.2853	0.2773
Belebele	**0.3300**	0.3233	0.3178	0.3133	0.3133	0.3056	0.3022
C^3	**0.4550**	0.4480	0.4400	0.4500	0.4400	0.4400	0.4370
C-Eval	**0.3095**	0.3060	0.2760	0.2903	0.2906	0.2878	0.2805
CMMLU	**0.3312**	0.3259	0.3041	0.3043	0.3060	0.3009	0.2995
CMRC 2018	0.2224	0.2125	0.1614	**0.2251**	0.2164	0.1949	0.1866
HellaSwag	0.3790	**0.3800**	0.3530	0.3680	0.3660	0.3510	0.3370
M3Exam	**0.3319**	0.3245	0.3084	0.3201	0.3245	0.3216	0.3245
X-CODAH	0.3033	0.3000	**0.3233**	0.3100	0.2900	0.2967	0.3067
X-CSQA	**0.2740**	0.2680	0.2690	0.2610	0.2520	0.2510	0.2650
XCOPA	0.6200	**0.6400**	0.6180	0.5740	0.5740	0.6000	0.5620
OCNLI	0.5470	0.5470	0.5340	0.5250	**0.5600**	0.5420	0.5060
Chinese-SQuAD	0.0929	**0.1097**	0.0865	0.0889	0.0850	0.0777	0.0585
XStoryCloze	**0.5800**	0.5630	0.5710	0.5560	0.5610	0.5580	0.5570
XWINO	0.6429	0.6528	**0.6587**	0.6131	0.5992	0.6429	0.6111

Table 6: Benchmark performance comparison in Chinese between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining 10% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 119B tokens were consumed.

Approach	*FT MKC*$^+$	*MLP MKC*$^+$	*MLP MKC*	*FT MKC*	*CS MKC*	*CS MKC*$^+$	Baseline
Average Rank	3.2222	3.5000	3.5556	3.7778	4.0000	4.6667	5.2778
Belebele	0.3378	0.3533	**0.3678**	0.3489	0.3444	0.3344	0.3444
HellaSwag	**0.5380**	**0.5380**	0.4990	0.5150	0.5280	0.5070	0.5180
X-CSQA	0.2820	0.2740	0.2730	**0.2990**	0.2850	0.2900	0.2870
XNLI 2.0	0.7340	0.7400	0.7430	0.7230	**0.7450**	0.7330	0.7180
FQuAD	0.2597	0.2803	**0.3032**	0.2981	0.2411	0.2476	0.2401
MMLU	0.2896	0.2895	**0.2925**	0.2886	0.2806	0.2815	0.2706
Mintaka	0.0710	0.0438	0.0334	0.0670	0.0610	**0.0976**	0.0712
X-CODAH	**0.3000**	0.2667	0.2867	0.2767	**0.3000**	0.2800	0.2633
ARC (Challenge)	0.3120	**0.3180**	0.3090	0.3060	0.2950	0.2830	0.2850

Table 7: Benchmark performance comparison in French between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining 10% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 119B tokens were consumed.

Approach	*MLP MKC*$^+$	*FT MKC*$^+$	*FT MKC*	*CS MKC*	*MLP MKC*	*CS MKC*$^+$	Baseline
Average Rank	3.1250	3.1250	3.5000	3.7500	4.5000	4.7500	5.2500
MMLU	**0.2940**	0.2879	0.2926	0.2770	0.2905	0.2764	0.2718
ARC (Challenge)	0.2760	0.2850	0.2820	**0.2880**	0.2830	0.2640	0.2680
Mintaka	0.0580	0.0548	0.0735	0.0576	0.0494	**0.0766**	0.0498
Belebele	**0.3611**	0.3578	0.3544	0.3544	0.3567	0.3422	0.3544
X-CODAH	0.3367	0.3500	0.3300	0.3567	0.3400	**0.3600**	0.3467
X-CSQA	0.2978	**0.3008**	0.2877	0.2887	0.2857	0.2918	0.2787
HellaSwag	0.4640	0.4710	**0.4870**	0.4820	0.4540	0.4390	0.4470
XNLI 2.0	0.6620	0.6530	0.6740	0.6440	0.6610	0.6520	**0.6890**

Table 8: Benchmark performance comparison in German between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining 10% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 119B tokens were consumed.

Approach	*MLP MKC*$^+$	*MLP MKC*	*FT MKC*$^+$	Baseline	*CS MKC*$^+$	*CS MKC*	*FT MKC*
Average Rank	2.7812	3.2500	3.6875	3.9688	3.9688	5.0312	5.3125
EXAMS	0.3537	**0.3656**	0.3552	0.3582	0.3443	0.3262	0.3346
MMLU	0.4007	0.3909	**0.4023**	0.3894	0.3912	0.3781	0.3885
ARC (Easy)	**0.4330**	0.4230	0.4210	0.4120	0.4020	0.3940	0.4080
AlGhafa SciQ	0.6915	**0.7005**	0.6965	0.6854	0.6724	0.6683	0.6804
Belebele	0.3456	0.3356	0.3322	0.3311	0.3356	**0.3567**	0.3233
SOQAL	**0.7333**	0.6867	0.7000	0.7200	0.7267	0.6867	0.7133
MLQA	0.2386	**0.2402**	0.1928	0.1901	0.2189	0.2154	0.1793
TyDi QA	**0.1547**	0.1476	0.1230	0.1441	0.1223	0.1097	0.1182
AlGhafa RACE	0.3720	**0.3740**	0.3640	0.3710	0.3590	0.3660	0.3730
ARCD	**0.3638**	0.3505	0.3235	0.3354	0.3358	0.3432	0.3043
X-CODAH	0.2600	0.2533	0.2567	**0.2633**	**0.2633**	0.2500	0.2600
AlGhafa PIQA	0.6360	0.6320	**0.6400**	0.6240	0.6320	0.6320	0.6370
X-CSQA	0.2740	0.2810	0.2770	**0.2900**	0.2880	0.2720	0.2770
XNLI 2.0	0.6570	0.6910	0.6990	**0.7010**	0.6910	0.6900	0.6770
HellaSwag	0.4270	0.4220	0.4280	0.4250	0.4260	**0.4320**	0.4150
XStoryCloze	0.6150	0.6100	0.6100	0.6070	0.6130	**0.6180**	0.5930

Table 9: Benchmark performance comparison in Arabic between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining 56% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 119B tokens were consumed.

Approach	*CS MKC*$^+$	*MLP MKC*$^+$	*FT MKC*$^+$	Baseline
Average Rank	1.0000	2.5000	3.1667	3.3333
ARC (Challenge)	**0.2820**	0.2650	0.2730	0.2560
HellaSwag	**0.4950**	0.4850	0.4750	0.4750
Belebele	**0.3333**	0.3289	0.3189	0.3289

Table 10: Benchmark performance comparison in Danish between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, and *CS*) trained on *MKC*$^+$ or *MKC*, retaining 65% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated after 119B tokens were consumed.

B.2 Threshold Selection

To confirm that the *CS* filtering method is not competitive with *MLP* and *FT*, even when a higher percentage of documents is retained, we present the complete threshold selection results, including the *CS* method, in Table 11 in addition to the results shown in Section 4.2.2 (Table 2).

We provide further results on the variation in the average length of documents retained by our model-based filtering approaches for Chinese, French, Arabic, and Danish. These results complement the findings for German discussed in Section 4.2.2 and are shown in Figure 3. Table 12 lists the actual dataset sizes (number of retained tokens) after tokenization for all languages.

Approach	Threshold	Average Rank
MLP MKC$^+$	10%	11.73
MLP MKC$^+$	15%	12.13
MLP MKC	20%	15.07
MLP MKC	15%	15.09
MLP MKC$^+$	20%	15.40
MLP MKC	10%	16.09
FT MKC$^+$	10%	18.61
CS MKC	15%	19.02
CS MKC	20%	19.24
FT MKC	15%	19.84
FT MKC	10%	20.02
CS MKC	10%	20.67
FT MKC	20%	20.80
FT MKC$^+$	15%	22.05
FT MKC$^+$	20%	22.52
CS MKC$^+$	15%	24.66
CS MKC$^+$	20%	25.08
Baseline	–	25.54
CS MKC$^+$	10%	26.94

Table 11: Benchmark performance comparison (average rank) between the baseline (FineWeb-2) and our proposed filtering methods (*FT*, *MLP*, *CS*) trained on *MKC*$^+$ or *MKC*, retaining top 10%, 15% or 20% of the documents. The average rank is computed across FineTasks performance of 1B-parameter models evaluated for Chinese, German and French after 70B and 119B tokens were consumed.

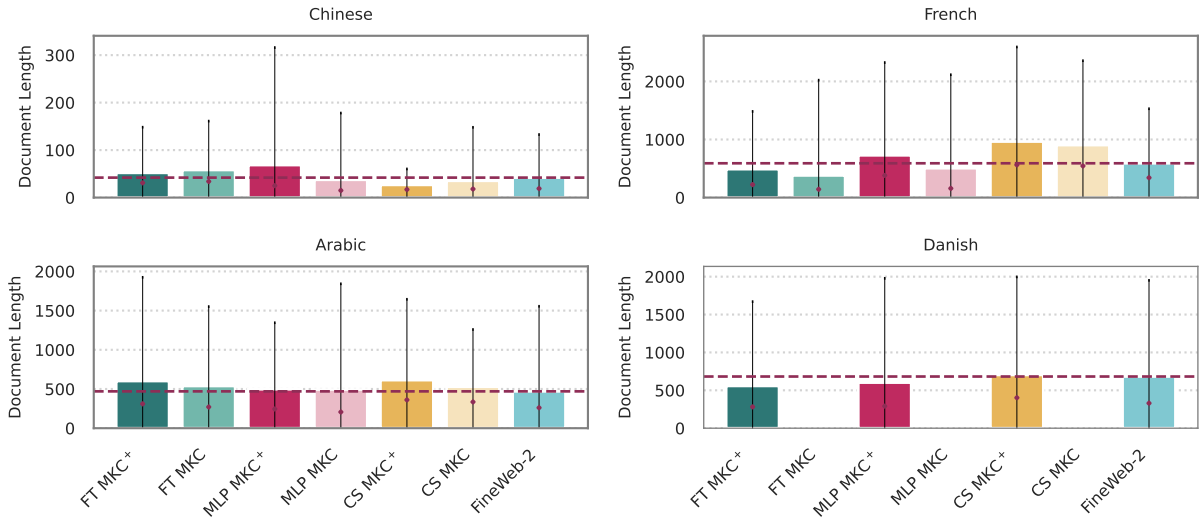


Figure 3: Comparison of average document length and standard deviation in FineWeb-2 before and after filtering using one of our approaches retaining top 10% of the documents for Chinese and French, 56% for Arabic and 65% for Danish. The average document length of FineWeb-2 is represented as a red horizontal line, while the medians are shown as red dots. Document length is measured based on number of space-separated tokens.

Approach	Chinese	French	German	Arabic	Danish
MLP MKC$^+$	150B (9%)	89B (12%)	119B (12%)	78B (61%)	71B (66%)
MLP MKC	105B (7%)	72B (10%)	87B (9%)	75B (59%)	–
FT MKC$^+$	221B (14%)	70B (10%)	63B (6%)	77B (61%)	70B (65%)
FT MKC	190B (12%)	43B (6%)	65B (7%)	80B (63%)	N/A
CS MKC$^+$	170B (11%)	126B (17%)	166B (17%)	82B (65%)	77B (71%)
CS MKC	161B (10%)	132B (18%)	172B (18%)	83B (65%)	–
Baseline	1597B	730B	973B	127B	108B

Table 12: Comparison of retained tokens in FineWeb-2 before and after filtering using one of our proposed approaches retaining top 10% of the documents for Chinese, French and German, 56% for Arabic and 65% for Danish. The token counts correspond to the size of the tokenized datasets, processed with the multilingual Mistral v3 (Tekken) tokenizer (Mistral AI, 2024).

B.3 Training Data Analysis

We give details on the variation in the average length of documents retained by our model-based filtering method *MLP* for Chinese, French, Arabic, and Danish with different training datasets. The results are shown for German in Figure 4 and for all other languages in Figure 5.

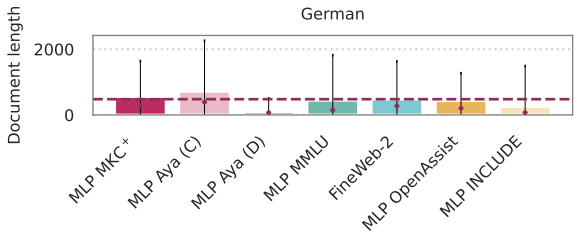


Figure 4: Comparison of average document length and standard deviation in FineWeb-2 before and after filtering using *MLP* filtering method retaining top 10% of the documents with different training datasets. The average document length of FineWeb-2 is represented as a red horizontal line, while the medians are shown as red dots. Document length is measured based on number of space-separated tokens.

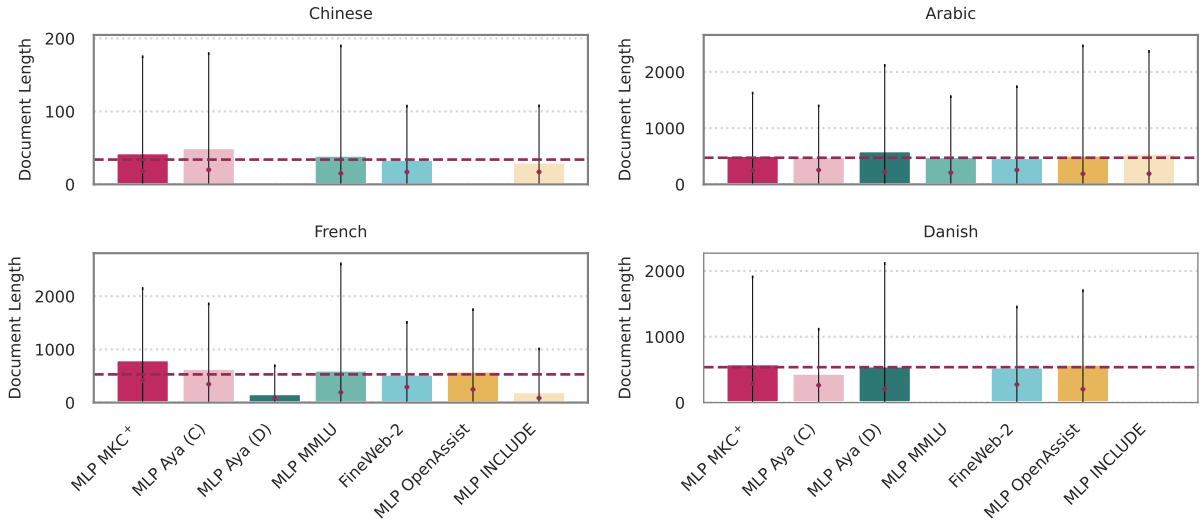


Figure 5: Comparison of average document length and standard deviation in FineWeb-2 before and after filtering using *MLP* filtering method retaining top 10% of the documents for Chinese and French, 56% for Arabic and 65% for Danish with different training datasets. The average document length of FineWeb-2 is represented as a red horizontal line, while the medians are shown as red dots. Document length is measured based on number of space-separated tokens.

B.4 Impact on multilingual model training

This section presents the results of our *MLP MKC*$^+$ approach on multilingual model training for Chinese (Table 13), Arabic (Table 14), German (Table 15), and Danish (Table 16), in addition to the results for French discussed in Section 4.2.4.

Dataset	Ours	Ours (M)	FW-2 (M)	FW-2
Average Rank	1.5667	2.1667	2.9000	3.3667
AGIEval	**0.2995**	0.2863	0.2894	0.2853
Belebele	0.3300	**0.3456**	0.3189	0.3056
C^3	**0.4550**	0.4520	0.4480	0.4400
C-Eval	**0.3095**	0.2848	0.2683	0.2878
CMMLU	**0.3312**	0.3064	0.2967	0.3009
CMRC 2018	0.2224	**0.2689**	0.2090	0.1949
HellaSwag	**0.3790**	0.3740	0.3740	0.3510
M3Exam	**0.3319**	0.3040	0.3304	0.3216
X-CODAH	0.3033	**0.3067**	0.2800	0.2967
X-CSQA	0.2740	**0.2810**	0.2780	0.2510
XCOPA	**0.6200**	0.6020	0.5860	0.6000
OCNLI	**0.5470**	0.5320	0.4910	0.5420
Chinese-SQuAD	0.0929	**0.1304**	0.1017	0.0777
XStoryCloze	**0.5800**	0.5760	0.5650	0.5580
XWINO	0.6429	0.6409	**0.6468**	0.6429

Table 13: Benchmark performance comparison for Chinese of multilingual LLMs trained on FineWeb-2 or the refined dataset using our *MLP MKC*$^+$ approach (retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish) trained on 595B tokens, against their monolingual counterparts trained on 119B tokens. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

Dataset	Ours (M)	Ours	FW-2	FW-2 (M)
Average Rank	1.9688	2.0000	2.7500	3.2812
EXAMS	0.3336	0.3537	**0.3582**	0.3076
MMLU	0.3828	**0.4007**	0.3894	0.3599
ARC (Easy)	0.4190	**0.4330**	0.4120	0.3760
AlGhafa SciQ	0.6764	**0.6915**	0.6854	0.6563
Belebele	**0.3511**	0.3456	0.3311	0.3344
SOQAL	0.7000	**0.7333**	0.7200	0.6533
MLQA	0.2208	**0.2386**	0.1901	0.2085
TyDi QA	**0.1634**	0.1547	0.1441	0.1429
AlGhafa RACE	**0.3830**	0.3720	0.3710	0.3770
ARCD	0.3377	**0.3638**	0.3354	0.2970
X-CODAH	**0.2767**	0.2600	0.2633	**0.2767**
AlGhafa PIQA	0.6170	**0.6360**	0.6240	0.6160
X-CSQA	0.2860	0.2740	**0.2900**	0.2660
XNLI 2.0	0.7080	0.6570	0.7010	**0.7340**
HellaSwag	**0.4390**	0.4270	0.4250	0.4240
XStoryCloze	**0.6370**	0.6150	0.6070	0.6160

Table 14: Benchmark performance comparison for Arabic of multilingual LLMs trained on FineWeb-2 or the refined dataset using our *MLP MKC*$^+$ approach (retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish) trained on 595B tokens, against their monolingual counterparts trained on 119B tokens. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

Dataset	Ours (M)	Ours	FW-2	FW-2 (M)
Average Rank	1.5000	2.1250	2.9375	3.4375
MMLU	0.2918	**0.2940**	0.2718	0.2691
ARC (Challenge)	0.2740	**0.2760**	0.2680	0.2640
Mintaka	**0.0821**	0.0580	0.0498	0.0660
Belebele	**0.3956**	0.3611	0.3544	0.3633
X-CODAH	**0.3500**	0.3367	0.3467	0.3167
X-CSQA	**0.3048**	0.2978	0.2787	0.2787
HellaSwag	**0.4690**	0.4640	0.4470	0.4430
XNLI 2.0	0.6420	0.6620	**0.6890**	0.6340

Table 15: Benchmark performance comparison for German of multilingual LLMs trained on FineWeb-2 or the refined dataset using our *MLP MKC*$^+$ approach (retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish) trained on 595B tokens, against their monolingual counterparts trained on 119B tokens. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

Dataset	Ours (M)	Ours	FW-2 (M)	FW-2
Average Rank	1.6667	2.1667	3.0000	3.1667
ARC (Challenge)	**0.2920**	0.2650	0.2600	0.2560
HellaSwag	0.4710	**0.4850**	0.4560	0.4750
Belebele	**0.3700**	0.3289	0.3311	0.3289

Table 16: Benchmark performance comparison for Danish of multilingual LLMs trained on FineWeb-2 or the refined dataset using our *MLP MKC*$^+$ approach (retaining top 10% of the documents for Chinese, German, and French, 56% for Arabic, and 65% for Danish) trained on 595B tokens, against their monolingual counterparts trained on 119B tokens. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

B.5 Replay of Original Data

We explore whether incorporating small percentage of raw data can help improve performance. We do this for our best FastText (*FT MKC*$^+$) and Transformer approaches (*MLP MKC*$^+$). Table 17 presents the results of experiments where 5% and 10% unfiltered data were mixed into the training dataset, alongside results from training without any data mixing. Both *FT MKC*$^+$ and *MLP MKC*$^+$ approaches show mixed signal, although *MLP MKC*$^+$ approach demonstrates little difference between mixing 5% unfiltered data

and no mixing, indicating that it retains enough diversity.

Approach	Mixture Rate	Average Rank
MLP MKC$^+$	5%	5.09
MLP MKC$^+$	0%	5.16
MLP MKC$^+$	10%	5.40
FT MKC$^+$	10%	7.17
FT MKC$^+$	0%	7.51
FT MKC$^+$	5%	8.66

Table 17: Benchmark performance comparison (average rank) of our *MLP MKC*$^+$ and *FT MKC*$^+$ approaches, retaining top 10% of the documents while mixing in 0%, 5% or 10% of the FineWeb-2 dataset. The average rank is computed across FineTasks performance of 1B-parameter models evaluated for Chinese, German, or French, after consuming 70B and 119B tokens.

B.6 Data Contamination Analysis

To ensure the validity of our approach, we conduct decontamination experiments, as web crawl data may include evaluation benchmark tasks. While Li et al. (2024b) addressed similar concerns, our approach follows the methodology of Brown et al. (2020). Specifically, we perform 13-gram decontamination of the LLM training data separately for English and French evaluation benchmarks. However, unlike the original approach, we remove the entire document if it is flagged as contaminated, using the implementation provided in DataTrove (Penedo et al., 2024b).

Tables 18 and 19 present the results of decontamination experiments for English and French, respectively. We conducted the experiments using the following approaches (resulting in document contamination rates): baseline FineWeb English (0.16%), *MLP MKC*$^+$ English with 10% retention (0.19%), baseline FineWeb-2 French (0.14%), and *MLP MKC*$^+$ French with 10% retention (0.14%). As in our previous experiments, we train the models on 119B tokens. Additionally, we compare the results against equivalent training runs without decontamination to further analyze its impact. For an example of a contaminated sample, see Appendix E.

For English models, decontamination slightly reduces performance both for our approach and baseline FineWeb data. However, even when decontaminated, our approach still outperforms training on non-decontaminated baseline data. For French models, performance of our approach is comparable between decontaminated and non-decontaminated datasets, with both continuing to outperform baseline FineWeb-2 data. Interestingly, decontaminated baseline data yields better results than its non-decontaminated counterpart.

Dataset	Ours	Ours (D)	FW*	FW* (D)
Average Rank	1.5000	2.1111	3.0556	3.3333
ARC (Challenge)	**0.3550**	0.3440	0.3010	0.2880
ARC (Easy)	**0.6670**	0.6520	0.5880	0.5700
CommonsenseQA	0.3870	**0.4000**	0.3850	0.3820
HellaSwag	**0.6040**	**0.6040**	0.5930	0.5890
MMLU	**0.3400**	0.3220	0.3030	0.3050
OpenBookQA	**0.3860**	0.3840	0.3560	0.3740
PIQA	0.7510	0.7590	**0.7620**	0.7600
WinoGrande	**0.5720**	0.5550	0.5550	0.5570
TriviaQA	**0.0820**	0.0380	0.0370	0.0250

Table 18: Benchmark performance comparison in English for our *MLP MKC*$^+$ approach (retaining top 10% of the documents), both decontaminated (D) and non-decontaminated, against the baseline FineWeb datasets, also in decontaminated and non-decontaminated variants. The average rank is computed across SmolLM task performance for 1B-parameter models trained on 119B tokens.

Dataset	Ours	Ours (D)	FW-2 (D)	FW-2
Average Rank	2.0556	2.0556	2.7222	3.1667
Belebele	0.3533	0.3400	**0.3778**	0.3444
HellaSwag	**0.5380**	0.5350	0.5180	0.5180
X-CSQA	0.2740	0.2810	0.2730	**0.2870**
XNLI 2.0	**0.7400**	**0.7400**	0.7070	0.7180
FQuAD	0.2803	0.2620	**0.2890**	0.2401
MMLU	**0.2895**	0.2875	0.2711	0.2706
Mintaka	0.0438	**0.0797**	0.0658	0.0712
X-CODAH	0.2667	**0.2900**	0.2800	0.2633
ARC	**0.3180**	0.3110	0.2880	0.2850

Table 19: Benchmark performance comparison in French for our *MLP MKC*$^+$ approach (retaining top 10% of the documents), both decontaminated (D) and non-decontaminated, against the baseline FineWeb-2 datasets, also in decontaminated and non-decontaminated variants. The average rank is computed across FineTasks performance for 1B-parameter models trained on 119B tokens.

B.7 Approach Validation on English

We explore whether our approach transfers to English and assess the performance. Table 20 presents the performance of *MLP MKC*$^+$ with 10% retention applied to the English FineWeb dataset (Penedo et al., 2024a). Our method is compared against FineWeb and baselines using model-based filtered datasets, including DCLM (Li et al., 2024b) and FineWeb-Edu (Penedo et al., 2024a). To save computational resources, we use the 6 most recent FineWeb and FineWeb-Edu dumps and the first partition of DCLM[6], which we denote with *. Each of these subsets contains more than 119B tokens, with FineWeb retaining this size even after applying our filtering retaining top 10% of the documents.

While each approach demonstrates strengths in different benchmarks, as seen from Table 20 and Figure 1, the overall average rank results indicate that our method outperforms all other baselines.

Dataset	Ours	DCLM*	FW-Edu*	FW*
Average Rank	1.8333	2.3889	2.4444	3.3333
ARC (Challenge)	0.3550	0.3530	**0.3850**	0.3010
ARC (Easy)	0.6670	0.6470	**0.6970**	0.5880
CommonsenseQA	0.3870	**0.4100**	0.3770	0.3850
HellaSwag	**0.6040**	0.5960	0.5700	0.5930
MMLU	0.3400	0.3160	**0.3470**	0.3030
OpenBookQA	0.3860	0.3840	**0.4180**	0.3560
PIQA	0.7510	0.7510	0.7410	**0.7620**
WinoGrande	**0.5720**	0.5610	0.5660	0.5550
TriviaQA	0.0820	**0.1240**	0.0320	0.0370

Table 20: Benchmark performance comparison for English of our *MLP MKC*$^+$ approach (retaining top 10% of the documents) against baseline datasets: FineWeb, DCLM, and FineWeb-Edu. The average rank is computed across SmolLM task performance for 1B-parameter models trained on 119B tokens.

[6] huggingface.co/datasets/mlfoundations/dclm-baseline-1.0-parquet

C List of evaluation benchmarks and metrics

We provide a detailed overview of the evaluation benchmarks used to assess our models' performance, along with their respective evaluation metrics in Table 21. For non-English tasks and English MMLU, we use the *cloze* multiple-choice prompt, which allows the model to directly predict each option instead of using the standard prompt format with A/B/C/D letter prefixes as targets. This approach was chosen because it has been shown to serve as a more reliable performance indicator earlier in training (Kydlíček et al., 2024). We evaluate the models in a 0-shot setting.

Table 21: List of evaluation benchmarks and metrics used in our setup for Chinese, French, German, Arabic, Danish, and English.

Benchmark	Chinese	French	German	Arabic	Danish	English	Evaluation metric
AGIEval (Zhong et al., 2023)	✓						Normalized accuracy
AlGhafa ARC (Almazrouei et al., 2023)				✓			Normalized accuracy
AlGhafa PIQA (Almazrouei et al., 2023)				✓			Normalized accuracy
AlGhafa RACE (Almazrouei et al., 2023)				✓			Normalized accuracy
AlGhafa SciQ (Almazrouei et al., 2023)				✓			Normalized accuracy
ArabicMMLU (Koto et al., 2024)				✓			Normalized accuracy
ARC (Clark et al., 2018)						✓	Normalized accuracy
ARCD (Mozannar et al., 2019)				✓			F1 score
Belebele (Bandarkar et al., 2024)	✓	✓	✓	✓	✓		Normalized accuracy
C^3 (Sun et al., 2020)	✓						Normalized accuracy
C-Eval (Huang et al., 2023)	✓						Normalized accuracy
Chinese-SQuAD (Pluto-Junzeng, 2019)	✓						F1 score
CMMLU (Li et al., 2024a)	✓						Normalized accuracy
CMRC 2018 (Cui et al., 2019)	✓						F1 score
CommonsenseQA (Talmor et al., 2019)						✓	Normalized accuracy
EXAMS (Hardalov et al., 2020)				✓			Normalized accuracy
FQuAD (d'Hoffschmidt et al., 2020)		✓					F1 score
HellaSwag (Zellers et al., 2019)						✓	Normalized accuracy
M3Exam (Zhang et al., 2023)	✓						Normalized accuracy
Meta MMLU (Llama Team, 2024)		✓	✓				Normalized accuracy
Mintaka (Sen et al., 2022)		✓	✓				F1 score
MLMM ARC (Lai et al., 2023)		✓	✓		✓		Normalized accuracy
MLMM HellaSwag (Lai et al., 2023)	✓	✓	✓	✓	✓		Normalized accuracy
MLQA (Lewis et al., 2020)				✓			F1 score
MMLU (Hendrycks et al., 2020)						✓	Normalized accuracy
OCNLI (Hu et al., 2020)	✓						Normalized accuracy
OpenBookQA (Mihaylov et al., 2018)						✓	Normalized accuracy
PIQA (Bisk et al., 2019)						✓	Normalized accuracy
SOQAL (Mozannar et al., 2019)				✓			Normalized accuracy
TriviaQA (Joshi et al., 2017)						✓	Quasi-exact match
TyDi QA (Clark et al., 2020)				✓			F1 score
WinoGrande (Sakaguchi et al., 2019)						✓	Normalized accuracy
X-CODAH (Lin et al., 2021a)	✓	✓	✓	✓			Normalized accuracy
XCOPA (Ponti et al., 2020)	✓						Normalized accuracy
X-CSQA (Lin et al., 2021a)	✓	✓	✓	✓			Normalized accuracy
XNLI 2.0 (Upadhyay and Upadhya, 2023)		✓	✓	✓			Normalized accuracy
XStoryCloze (Lin et al., 2021b)	✓			✓			Normalized accuracy
XWINO (Tikhonov and Ryabinin, 2021)	✓						Normalized accuracy

D FineWeb documents in different scoring approaches

To illustrate the types of documents each classifier scores highly or poorly, we present the highest- and lowest-scoring FineWeb examples for each of our classifier approaches (*FT MKC*$^+$, *MLP MKC*$^+$, *CS MKC*$^+$). These examples were selected from the randomly chosen FineWeb test dataset (10K) used to validate the training of our model-based classifiers.

D.1 FastText Classifier (FT)

Highest score:

hi. i couldn't solve my problem because it has two conditional logical propositions. the problem is:can anyone help me about this, thanks =)we're expected to know that: . is equivalent tofind a logically equivalent proposition for:by first writing its contrapositive, and then applying demorgan's lawand the equality forthey were trying to be helpful by outlining the steps we should follow,. . but i think they made it more confusing.i don't see the purpose of using the contrapositive here.. . i wouldn't have done it that way.besides, the statement is a tautology . . .which gives us: .and this is a tautology: "a thing implies itself" ... which is always true.i don't know of any "logically equivalent proposition" we can write . . .

Lowest score:

|starts||23 sep 2016 (fri) (one day only)|want to travel soon but don't wish to fork out a fortune for flights? check out today's promotion from jetstar featuring promo fares fr $35 all-in valid for travel period commencing 12 october 2016don't miss out! all-in frenzy fares to hong kong, penang and more from $35.sale ends 23 sep, 11pm!|travelling||price||travel period||find flight||penang||$35|| [...]

D.2 Multi-Layer Perceptron (MLP)

Highest score:

Naqhadeh County is a county in West Azerbaijan Province in Iran. The capital of the county is Naqadeh. At the 2006 census, the county's population was 117,831, in 27,937 families. The county is subdivided into two districts: the Central District and Mohammadyar District. The county has two cities: Naqadeh and Mohammadyar.

Lowest score:

Custom Wedding Gifts
Personalized photo frames, albums & keepsakes. Heirloom quality!
Custom Engraved Journals
Handmade in Florence Italy. Dozens of sizes and paper styles!
Awesome Leather Journals
Personalized, Customizable, Artisan made in Santa Fe, NM.
Ink Rendering from Photos
100% Hand painted with unique style by pro artists. From $49.

D.3 Cosine Similarity (CS)

Highest score:

When you are renting a 5, 10, 15, 20, 30 or 40 yard dumpster, you want a company you can trust with prices that make you smile. Give us a call today and see the difference we can make in your next construction or clean out project.
Simply give us a call and we will help you figure out your dumpster rental needs.
Our dumpsters usually go out same-day or next-day depending on when you call.
We provide top-notch service, while going easy on your bottom line. What more could you ask for?
Our trained operators are here to give you a fast and hassle-free experience from start to finish.[...]

Lowest score:

Cooperative flat 206/J
- Cooperative flat 201/J - Sold
2(1)+kitchenette, 50,1 m2Cooperative flat 202/J - Sold
2(1)+kitchenette, 44,9 m2Cooperative flat 203/J - Sold
2(1)+kitchenette, 50,6 m2Cooperative flat 204/J - Sold
1+kitchenette, 27,1 m2Cooperative flat 205/J - Sold
2(1)+kitchenette, 50,1 m2Cooperative flat 206/J - On sale
3+kitchenette 86,7 m2[...]

E Example of a contaminated document

We present an example of a FineWeb document that was removed during our decontamination pipeline.

MMLU contaminated document (matched 13-gram in bold):

Here is our diagram of the Preamble to the Constitution of the United States. It is based on our understanding of the use of "in order to" as a subordinating conjunction that introduces a series of infinitival clauses (without subjects) that, in turn, modify the compound verbs "do ordain" and "establish."
See A Grammar of Contemporary English by Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. Longman Group: London. 1978. p. 753.
We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, **provide for the common defence, promote the general Welfare, and secure the Blessings** of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.
If you have alternative rendering for this sentence, we would be happy to hear of it. Use the e-mail icon to the left.

Fine-tuning Whisper on Low-Resource Languages for Real-World Applications

Vincenzo Timmel[1], Claudio Paonessa[2], Manfred Vogel[1], Daniel Perruchoud[1], Reza Kakooe[1]

[1]University of Applied Sciences and Arts Northwestern Switzerland
[2]Noxenum.io

{vincenzo.timmel, manfred.vogel, daniel.perruchoud, reza.kakooee}@fhnw.ch, claudio.paonessa@noxenum.io

Abstract

This paper presents a new approach to fine-tuning OpenAI's Whisper model for low-resource languages by introducing a novel data generation method that converts sentence-level data into a long-form corpus, using Swiss German as a case study. Non-sentence-level data, which could improve the performance of long-form audio, is difficult to obtain and often restricted by copyright laws. Our method bridges this gap by transforming more accessible sentence-level data into a format that preserves the model's ability to handle long-form audio and perform segmentation (by predicting timestamps) without requiring non-sentence-level data. Our data generation process improves performance in several real-world applications and leads to the development of a new state-of-the-art speech-to-text (STT) model for Swiss German. We compare our model with a non-fine-tuned Whisper and previous state-of-the-art Swiss German STT models, where our new model achieves higher BLEU scores. Our results also indicate that the proposed method is adaptable to other low-resource languages, supported by written guidance and code that allows the creation of fine-tuned Whisper models, which keep segmentation capabilities and allow the transcription of longer audio files using only sentence-level data with high quality.

1 Introduction

Swiss German refers to the dialects spoken in the German-speaking regions of Switzerland. Due to the limited number of speakers, linguistic resources are scarce, also because Swiss German exists only as a spoken language, without any formal grammar or standardized written form. As a result, Swiss German STT systems are typically formulated as speech translation tasks, where Swiss German audio is transcribed into standard German text (Plüss et al., 2021, 2022, 2023). Generally, "Swiss German transcription" or "Swiss German ASR" refers to converting spoken Swiss German directly into written Standard German, combining transcription and translation in a single step.

The Whisper models (Radford et al., 2022) developed by OpenAI are trained on a large-scale corpus of audio recordings and corresponding transcriptions obtained by web crawling. The dataset used for the multilingual version of Whisper includes samples from almost 100 different languages. After English and Chinese, German represents the third-largest part of the dataset, with 13'344 hours. The unexpectedly high transcription quality of Whisper for Swiss German audio and video (see Table 4) and observed distinct hallucinations (discussed in Section 5) prove the presence of Swiss German audio in the original training dataset. While the model performs remarkably well for Swiss German, there is still considerable room for improvement for practical applications that require higher transcription quality, such as judicial interrogation transcripts or medical diagnosis and treatment orders.

In addition, fine-tuning solutions which fail to predict timestamps make it impossible to use Whisper for subtitling, multi-speaker conversation pattern analysis and other real-world applications[1]. Table 1 illustrates how Whisper Large-v2 fine-tuned on sentences loses is capabilities to predict timestamps and starts to fail when the audio segment is getting longer and more difficult to predict, even though when evaluating it on the test-split of the sentence-level dataset, the fine-tuned model shows much improvement over the original Whisper Large-v2.

In this paper, we focus on fine-tuning OpenAI's Whisper model for low-resource languages in real-world applications, using Swiss German as a case study. We evaluate the segmentation capabilities

[1]https://huggingface.co/blog/fine-tune-whisper

Table 1: Comparison of a rapidly spoken Swiss-German saying (Straub, 2024), transcribed by Whisper Large-v2. The model, fine-tuned on sentence-level data, fails to predict timestamps and performs worse than the original Large-v2.

Input Audio
...
Ich zeig der, wo de Bartli de Moscht holt.
...
Whisper Large-v2
...
`[00:00:08]` Ich zeige dir, wo Bartli den Most holt. `[00:00:11]`
...
Whisper Large-v2 (fine-tuned on sentences)
...
Ich zeige dir, wo es die Bartli in den Most holt.
...

of Whisper after fine-tuning and demonstrate the beneficial effect of fine-tuning on long-form audios generated from sentence-level data. Finally, we evaluate the impact of the amount of training data on model performance when fine-tuning Whisper.

In particular, we address the following key research questions:

- How can sentence-level datasets be adapted to effectively train Whisper models for longer audio sequences while maintaining segmentation and transcription quality?
- How does fine-tuning Whisper affect its segmentation capabilities, especially when moving from sentence-level to long-form data?
- How does fine-tuning Whisper with additional datasets, such as pseudo-labeled long-form audio, affect its performance in various real-world scenarios?

By exploring these research questions, this paper provides insights into improving STT systems for low-resource languages through innovative data generation and fine-tuning strategies.

2 Related work

Despite the challenge of scarce data resources, recent advancements in speech translation for Swiss German have been substantial, driven in part by recent collections of high-quality sentence-level datasets from Swiss parliaments minutes and crowdsourcing initiatives such as SwissDial (Dogan-Schönberger et al., 2021), the Swiss Parliaments Corpus SPC (Plüss et al., 2021), SDS-200 (Plüss et al., 2022), and STT4SG-350 (Plüss et al., 2023).

Prior to the release of Whisper, a commonly applied foundation model for building ASR systems was XLS-R (Babu et al., 2021). XLS-R is based on the wav2vec 2.0 architecture (Baevski et al., 2020) and was pre-trained on 436K hours of speech data in 128 languages. Previous research on Swiss German speech recognition has therefore often used the pre-trained XLS-R backbone (Plüss et al., 2022, 2023; Schraner et al., 2022; Paonessa et al., 2023).

There are many papers and blog posts on the fine-tuning of Whisper (Gandhi, 2022; Singh, 2023; de Bruin, 2023; OpenAI, 2023; Ma et al., 2023; Shamsian et al., 2024; Do et al., 2023; Ferraz et al., 2023; Sicard et al., 2023) and some focused on low-resource settings (Piñeiro-Martín et al., 2024; Hsu et al., 2024; Liu et al., 2024; Pillai et al., 2024). The problem of language forgetting is also discussed extensively, and it is shown that fine-tuning on a new language yields the best performance for the new language, but degrades the capabilities on existing languages (Qian et al., 2024). However, it is rare to find papers that explicitly address the problem of fine-tuning Whisper for transcription of longer audios. Many papers also do not evaluate or discuss the segmentation capabilities after fine-tuning. Finally, it is uncommon to see evaluations of fine-tuned Whisper on out-of-distribution datasets.

3 Approach

3.1 Data Generation

The Whisper model works with a fixed input length of 30 seconds. Samples shorter than 30 seconds must be padded by appending zeros (silence). However, available datasets often consist of sentence-level samples, which are usually much shorter than 30 seconds. This is also true for Swiss German speech translation corpora such as SPC, SDS-200 and STT4SG-350. Using such training data to fine-tune Whisper is challenging because it requires significant padding for each sample. This carries the potential risk of compromising the model's ability to robustly handle long-form audio and predict timestamps, a crucial aspect of the model for many use cases. We therefore start with available sentence-level pairs of Swiss German audio and standard German transcriptions (see Table 3) and concatenate multiple sentences to synthetically generate long-form audios with corresponding segment

timestamps, as shown in Figure 1.

Figure 1: Illustration of generated long-form training data from sentence-level audios. Although timestamps are available via the length of the audio, they are not displayed here.

The data generation strategy contains the following steps:

- **Timestamp Correction:** By using Voice Activity Detection (VAD), specifically leveraging Silero Models[2], we correct the start and end timestamps of the resulting audio segments.
- **Noise Overlapping:** By simply concatenating two audio samples, the transitions often become noticeable because they abruptly change noise characteristics. To improve the transitions between consecutive samples, we employ a random overlapping technique that leverages the silence intervals detected by VAD at the beginning and end of each sample. By taking advantage of these silence parts, this enhancement accurately simulates consecutive audio segments. Together with Timestamp Correction, it also allows to create a speech overlap, such that two speakers speak over each other.
- **Speaker Retention:** For samples that include speaker identification, the probability of retaining the same speaker in successive samples is 50%. This enhancement helps to create more realistic sequences in which speaker changes occur at a pace with natural speech patterns.

Later in the study, in Table 5, we show the influence of these data generation strategies on different datasets. In Figure 2 the general approach is shown. The beginning and the end of speech are detected and then, when concatenating sentences together, we can either:

- do **Concat**, concatenate files as they are,
- detect the end and start of speech and make the non-speech **Overlap** by up to 200ms when concatenating audios,
- introduce a **Negative Overlap**, so that the speech of two sentences overlaps by 200ms.

Figure 2: Illustration of the logical structure for stitching together sentences using VAD and overlap mechanisms. With the help of a VAD model, we precisely mark the start and end of speech. This allows us to vary the length of pauses between sentence and even introduce an overlap.

3.2 Training Details

For model initialization, we use the Whisper Large-v2 weights, as initial tests showed it outperformed Whisper Large-v3 on the Swiss German datasets used. And we take advantage of its strong baseline performance for Swiss German by using the German language tag **DE** (see Section 4.2).

Using gradient check-pointing (Sohoni et al., 2022), gradient accumulation, and an 8-bit optimizer (Dettmers et al., 2021), we achieve an effective batch size of 256 on a single NVIDIA A100 40 GB GPU. Gradient check-pointing is applied to both the encoder and decoder, with 16 gradient accumulation steps and a per-step batch size of 16. Additionally, we apply stochastic depth (Huang et al., 2016) to the encoder and decoder blocks. This setup results in a mixed-precision training run that takes about 42 hours.

We use a learning rate scheduler with a linearly increasing warm-up phase followed by a linear decay to zero, as described in the original Whisper training procedure (Radford et al., 2022). During training, each sample has a 50% chance of containing timestamps and a 50% chance of containing prompts, mentioned in a comment on the OpenAI Whisper repository [3].

[2] https://github.com/snakers4/silero-models

[3] https://github.com/openai/Whisper/discussions/838

Table 2: Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Max. Learning Rate	2.0×10^{-4}
Weight Decay	0.1
Warmup Updates	128
AdamW Specific Parameters	
β_1	0.9
β_2	0.98
ϵ	1.0×10^{-9}

Following the improved training procedure of Whisper Large-v2, we apply SpecAugment (Park et al., 2019) during training with the same parameters as in (Radford et al., 2022), summarized in Table 2.

3.3 Train, Validation and Test Data

For our training data, we use the Swiss German sentence-level datasets (Plüss et al., 2021, 2022, 2023) with the predefined train and validation sets. For the train and validation split, unless mentioned otherwise, we use our data-generation pipeline explained in section 3.1. As additional training and validation data, we use Swiss Broadcasting Corporation (SRG) shows, pseudo-labeled (PL) by transcription with Whisper Large-v2. We selected 17 TV series, in which Swiss German is spoken.

As test data, we use the predefined split of the Swiss German sentence-level datasets (not processed by our pipeline and thus stay as single sentences) and a Dataset-A containing a manually transcribed doctor-patient conversation obtained from a confidential phone call. Due to data privacy, this dataset cannot be disclosed and remains a closed source dataset. As an additional test set, we use SRG data from 5 TV series for which manual transcriptions are available, i.e.: *Einstein*, *Puls*, *Impact Investigativ*, *SRF Kids News*, and *SRF ohne Limit*. In contrast to the pseudo-labeled SRG train and validation data, we use as test set subtitles manually created by SWISS TXT, a subsidiary of SRG.

Because we have reasonable suspicion (see Section 5) that OpenAI has data from the SRG in its Whisper training corpus, we only considered SRG data for the validation and test set broadcasted after the release of Whisper Large-v2 to allow a fair comparison with our baseline, the Whisper Large-v2 base model.

The total hours of data used for training, validation and testing are presented in Table 3.

Table 3: Overview of the datasets used for training, validation, and testing, including totals per split.

Name (Variant)	Split	Hours	# Speakers
SDS-200 (Clean)	Train	50	1,799
STT4SG-350 (All)	Train	276	219
SPC (0.9 IOU)	Train	176	194
SRG (PL)	Train	406	–
Total		908	$>$ 2,212
SDS-200 (Clean)	Val	5.2	288
STT4SG-350 (All)	Val	21	219
SRG (PL)	Val	20	–
Total		46.2	$>$ 507
SDS-200 (Clean)	Test	5.2	281
STT4SG-350 (All)	Test	34	56
SPC (0.9 IOU)	Test	6	26
SRG (SWISS TXT)	Test	20	–
Dataset-A	Test	0.22	2
Total		65.42	$>$ 365

In Figure 3, we analyze the relationship between the BLEU score (Papineni et al., 2002) on the STT4SG-350 test set and the amount of training data used for fine-tuning. For training, we used the 502 hours long-form corpus consisting of SDS-200, STT4SG-350 and SPC, but we do not include the pseudo-labeled data to show what can be expected from high-quality labeled data. The models were trained using hierarchically nested datasets, each partition holding approximately 20% of the training data. Training was run until the word error rate (WER) on the validation set showed no more improvement. Once the training had stabilized, the best performing model on WER was selected.

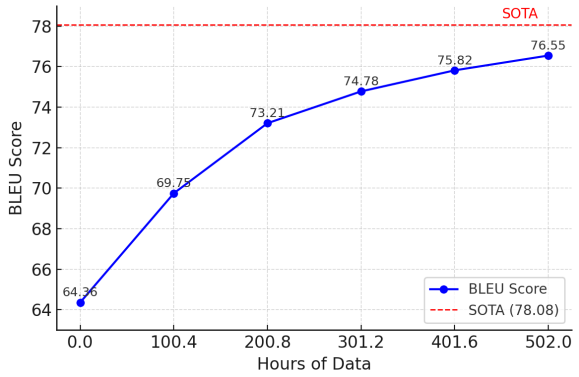


Figure 3: BLEU score on the STT4SG-350 test set vs. amount of training data (given in Table 3) used for fine-tuning. The model evaluated at 0 hours of training data corresponds to the original Whisper Large-v2. The SOTA model is discussed in section 4.4.

Based on the unabated rise of the curve in Figure 3, it's reasonable to deduct that more training data will improve model performance further.

4 Experiments

4.1 Evaluation

In all experiments, the computed metrics are derived from the WhisperX[4] versions of the models using fp16 precision. The beam size is configured to 5, **VAD** is enabled and the language tag is **DE**. For the sentence-level datasets, we report the WER or the BLEU metric or both, if the layout allows it. For the long-form test dataset we replace the BLEU metric with SubER (Wilken et al., 2022); this allows us to incorporate a metric that also assesses the quality of segmentation by timestamp prediction. Before calculating the metrics, all sentences are transformed to lowercase, and punctuation is removed. For the BLEU metric we specifically use sacreBLEU (Post, 2018) with default parameters.

4.2 Base Results

In Table 4 we compare the performance of the original Whisper Large models (v2 and v3) on our datasets. Interestingly, Whisper Large-v3 exhibits an improvement over its predecessor, Large-v2, only on the STT4SG-350 test set. It is noteworthy that the 24,605 samples in the STT4SG-350 test set yielded identical transcripts for both models in 11,340 instances (46.1%). Conversely, both the SPC and SDS-200 test sets reveal a slight decline in performance for Whisper Large-v3. The gap between the two models is most evident on the SRG dataset, where the Large-v2 model yields noticeably better performance.

[4] https://github.com/m-bain/whisperX

Table 4: Performance of the original Whisper models on various sentence-level test sets. SubER is shown on the long-form SRG data.

Test Dataset	Model	WER	BLEU	SubER
SPC	Large-v2	**28.21**	**58.08**	-
	Large-v3	28.94	57.90	-
SDS-200	Large-v2	**27.69**	**57.35**	-
	Large-v3	27.88	57.00	-
STT4SG-350	Large-v2	22.41	**64.36**	-
	Large-v3	**22.01**	64.13	-
SRG (SWISS TXT)	Large-v2	**28.42**	**63.61**	**30.63**
	Large-v3	38.69	56.31	42.58

Based on this comparison, there is no indication of a noticeable improvement on Swiss German audio when using Whisper Large-v3 instead of Large-v2.

4.3 Segmentation Forgetting & Out-of-Distribution Performance Degradation

To assess the impact of fine-tuning without timestamps on both segmentation capabilities and performance on out-of-distribution data, we conducted experiments comparing models trained on different data generation parameters with our pipeline. Table 5 presents the BLEU and SubER metrics for each model across various test datasets. The following parameters for the data generation are compared:

- **Sentence-level**: Training is conducted at the sentence level with padding.
- **Concat**: No adjustments are made; sentences are simply concatenated together.
- **Speaker_ret**: There is a 50% chance that two consecutive sentences originate from the same speaker without reusing any sentences.
- **Overlap**: There is a 50% chance that two sentences overlap, meaning the speech of the sentence begins immediately as the speech of the preceding sentence ends.
- **Neg_overlap**: There is a 10% chance that when audios overlap, the speech of two separate sentences overlaps by 200ms, simulating scenarios where two speakers talk simultaneously.

- **All**: This method combines all the aforementioned preprocessing techniques.

Table 5: Comparison of BLEU and SubER metrics for Whisper Large-v2, fine-tuned on sentence-level (first row) and differently generated long-form data (Large-v2 refers to the original model without fine-tuning). The results on SRG and Dataset-A show that the sentence-level fine-tuning performs noticeably worse compared to training on long-form data.

Test Dataset	Data/Model	BLEU	SubER
STT4SG-350	Sentence-level	**77.38**	-
	Concat	76.08	-
	Speaker_ret	76.91	-
	Overlap	72.12	-
	Neg_overlap	76.53	-
	All	76.55	-
	Large-v2	64.36	-
SRG (SWISS TXT)	Sentence-level	47.57	51.34
	Concat	51.34	42.87
	Overlap	49.68	44.15
	Speaker_ret	52.51	41.07
	Neg_overlap	50.63	43.56
	All	51.62	41.64
	Large-v2	**63.61**	**30.63**
Dataset-A	Sentence-level	35.01	55.44
	Concat	46.80	41.94
	Speaker_ret	46.79	41.99
	Overlap	45.84	43.62
	Neg_overlap	45.40	43.03
	All	47.22	41.15
	Large-v2	**48.89**	**39.12**

The model fine-tuned solely on padded sentence-level samples exhibits a substantial decline in timestamp prediction accuracy, evidenced by a SubER score exceeding 51 on the SRG dataset, even with the help of VAD, which substitutes timestamps. Because we use WhisperX to evaluate the models, the timestamps are given by the VAD-Model. Without WhisperX, the metrics of the sentence-level Model would be much worse.

While the sentence-level model attains with 77.38 the highest BLEU score on the STT4SG-350 dataset, it performs poorly on longer audio sequences in SRG and Dataset-A, especially when the segmentation quality is taken into account. This underscores its limitations with out-of-distribution data. In comparison, models trained with the generated long-form dataset — especially 'All' — demonstrate better generalization, maintaining higher BLEU scores and lower SubER scores across different datasets, nearly reaching the original Whisper Large-v2 on the Dataset-A on BLEU.

The original Large-v2 model without fine-tuning, outperforms the sentence-level model on long-form datasets. This suggests that fine-tuning exclusively on sentence-level, which was data processed into long-form, degrades the performance on unseen datasets. Incorporating long-form audio and datasets with diverse distribution into the training process is essential for preserving segmentation capabilities and ensuring robust performance across diverse data distributions.

4.4 Overcome Shortcomings

Despite the measures taken to simulate long-form audio, our fine-tuning procedure leads to a reduction in segmentation and transcription quality when applied to real long-form audio, as shown by the SubER metric given in Table 5 for the Dataset-A and SRG datasets. To address this issue, we enrich the training dataset by incorporating samples from the specific distribution of the intended prediction target, in our case, pseudo-labeled SRG data mentioned in Section 3.

As part of our methodology to mitigate language forgetting in the final model training, we use the Mozilla Common Voice 13 German dataset (Ardila et al., 2020). Using the data generation pipeline described in Section 3.1, we curated a subset of 15,000 samples, each lasting 30 seconds, resulting in 125 hours of additional training data. The train, validation, and test set splits were taken as defined by Mozilla Common Voice Version 13.

This leads to a model fine-tuned on long-form audio with our generated corpus based on the three sentence-level datasets (SPC, SDS-200, and STT4SG-350), the pseudo-labeled dataset SRG (PL) based on Swiss German TV shows, and the German part of the Common Voice 13 training data, concatenated as described above. This strategy significantly improves the model performance, as shown in Table 6, and leads to a new state-of-the-art model for Swiss German speech-to-text, referred to as **SOTA**.

4.5 Dialect Comparison

Since the STT4SG-350 test dataset contains identical sentences in 7 different dialects, it allows a fair comparison of model performance in terms of dialect-specific accuracy.

The results in Table 7 show large differences in the performance of the original Large-v2 model across the different dialect regions. In contrast, our fine-tuned SOTA model exhibits improved WER over a much narrower range across all dialects. For the reader, the improvements over other Swiss-

Table 6: Our new Whisper Large-v2 based SOTA model, fine-tuned on long-form audio created with data generation method 'All' and supplemented with SRG (PL) and Common Voice 13 de, compared to the original Whisper Large-v2

Test Dataset	Model	WER	SubER	BLEU
SPC	Our SOTA	**20.98**	-	**68.34**
	Large-v2	28.21	-	58.08
SDS-200	Our SOTA	**16.70**	-	**72.69**
	Large-v2	27.92	-	57.00
STT4SG-350	Our SOTA	**12.11**	-	**78.08**
	Large-v2	22.41	-	64.13
SRG (SW-TXT)	Our SOTA	**26.31**	**29.76**	**64.67**
	Large-v2	28.42	30.63	63.61
Dataset-A	Our SOTA	**34.50**	**35.31**	**51.40**
	Large-v2	38.00	39.12	48.89
CV13 de	Our SOTA	**6.42**	-	-
	Large-v2	6.53	-	-

German ASR models based on Wav2Vec (XLS-R) and Transformer (TF) architectures (Schraner et al., 2022) are also shown.

Table 7: WER for Swiss German dialects on the STT4SG-350 test set for selected models; XLS-R and TF are older models based on Wav2Vec (XLS-R) and Transformer (TF) architectures (Schraner et al., 2022).

Dialect	Large-v2	SOTA	XLS-R	TF
BS	25.02	**12.72**	16.30	21.24
BE	25.92	**13.68**	15.74	20.96
GR	19.59	**11.45**	14.32	17.29
IS	17.63	**10.73**	13.26	16.37
OS	21.27	**12.45**	16.45	18.58
VS	29.31	**12.72**	17.75	22.64
ZH	18.29	**11.03**	13.41	17.30

5 Conclusions

A key advantage of OpenAI's Whisper model is its ability to process audio of arbitrary length with built-in segmentation capabilities. However, fine-tuning such a model on sentence-level datasets while preserving these features is a significant challenge.

This paper demonstrates the potential of fine-tuning Whisper for low-resource languages, using Swiss German as a case study, and addresses the three research questions posed in the introduction. First, the paper shows how sentence-level datasets can be effectively adapted for training on longer audio sequences through a novel data generation pipeline, including techniques such as timestamp correction, noise overlapping, and speaker retention. These methods enable the generation of realistic long-form audio data that preserves segmentation and transcription quality.

Second, the fine-tuning approach significantly improves the model's segmentation capabilities, particularly for long-form data, compared to sentence-level models. By evaluating the segmentation performance with SubER metrics, the study highlights the benefits of incorporating diverse training data and demonstrates improved robustness for timestamp prediction and audio segmentation.

Finally, the inclusion of additional datasets, such as pseudo-labeled long-form audio from Swiss Broadcasting Corporation shows, improves the model's performance in real-world applications. We also show how to maintain performance in other languages by supplementing the training data with samples from those languages, thereby mitigating catastrophic forgetting. The results show that this method generalizes well to out-of-distribution datasets, achieving state-of-the-art performance in Swiss German speech-to-text tasks and suggesting broader applicability to other low-resource languages.

In addition, we have highlighted that while the model may improve on data from the same distribution as the training data, in reality the model preforms worse on out-of-distribution data, as shown in table 5. This underscores the importance of creating or acquiring evaluation datasets that closely mimic the intended deployment environment, ensuring the ASR system's robustness and usefulness in real-world applications.

Our research lays the groundwork for future work on data preparation and fine-tuning for OpenAI's Whisper model, especially in low-resource settings. For this we provide a simple framework, addressing catastrophic forgetting through long-form data generation and pseudo-labeling, enabling robust transcription even with limited datasets. The code for our data generation procedure [5] and model fine-tuning [6] is publicly available.

Additionally, we observed distinct hallucinations of Whisper Large-v2 mentioning Swiss subtitling companies, such as being able to reliably trigger Whisper to transcribe "Untertitel von SWISS TXT"

[5] https://github.com/i4ds/Whisper-prep
[6] https://github.com/i4ds/Whisper-finetune

- a watermark of SWISS TXT that is only present in the subtitle files, never in the audio - when asked to transcribe the title music of the SRF Meteo show or when music is being played in "SRF bi de Lüt".

6 Future work

As we have extensively analyzed and evaluated different methods to generate long-form data from sentence-level data, the combination of the data generation and methods to avoid catastrophic forgetting, as presented by (Qian et al., 2024) by using Elastic Weight Consolidation (Kirkpatrick et al., 2017), could be a next research topic. Another potential next step involves diversifying the data sources by augmenting the pseudo-labeled datasets with additional real-world data, including a broader range of TV programs, varied conversational contexts, and noisy environments. This expansion aims to enhance the robustness and generalization capabilities of the models. Notably, preliminary experiments indicate that a fine-tuned Whisper Large-v3 model performs particularly well on conversational speech, highlighting its potential superiority in this context and emphasizing the need for a large corpus of freely spoken dialogues.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

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. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Theo de Bruin. 2023. Asr-whisper-finetuning. https://github.com/Theodb/ASR-whisper-finetuning.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861.

Andrea Do, Oscar Brown, Zhengjie Wang, Nikhil Mathew, Zixin Liu, Jawwad Ahmed, and Cheng Yu. 2023. Using fine-tuning and min lookahead beam search to improve whisper. *arXiv:2309.10299*.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *Preprint*, arXiv:2103.11401.

Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. 2023. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. *arXiv:2311.01070*.

Sanchit Gandhi. 2022. Fine-tune whisper for multilingual asr with transformers. https://huggingface.co/blog/fine-tune-whisper.

Ming-Hao Hsu, Kuan Po Huang, and Hung yi Lee. 2024. Meta-whisper: Speech-based meta-icl for asr on low-resource languages. *Preprint*, arXiv:2409.10429.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. 2016. Deep networks with stochastic depth. *Preprint*, arXiv:1603.09382.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Rao Ma, Mengjie Qian, Mark J. F. Gales, and Kate M. Knill. 2023. Adapting an asr foundation model for spoken language assessment. *arXiv:2307.09378*.

OpenAI. 2023. Parameter-efficient fine-tuning of whisper-large v2 in colab on t4 gpu using peft+int8 training. https://github.com/openai/whisper/discussions/988.

Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect transfer for Swiss German speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15240–15254, Singapore. Association for Computational Linguistics.

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

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Leena G Pillai, Kavya Manohar, Basil K Raju, and Elizabeth Sherly. 2024. Multistage fine-tuning strategies for automatic speech recognition in low-resource languages. *Preprint*, arXiv:2411.04573.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docio-Fernández, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. In *Interspeech 2024*, pages 1235–1239. ISCA.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. In *Swiss Text Analytics Conference 2021*, Proceedings of the Swiss Text Analytics Conference 2021.

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.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. 2024. Learn and don't forget: Adding a new language to asr foundation models. In *Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss german speech to text system evaluation. *Preprint*, arXiv:2207.00412.

Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Keyword-guided adaptation of automatic speech recognition. *arXiv:2406.02649*.

Clement Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. Spaiche: Extending state-of-the-art asr models to swiss german dialects. *Preprint*, arXiv:2304.11075.

Vaibhav Singh. 2023. Faster whisper finetuning with lora powered by peft. `https://github.com/Vaibhavs10/fast-whisper-finetuning`.

Nimit S. Sohoni, Christopher R. Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré. 2022. Low-memory neural network training: A technical report. *Preprint*, arXiv:1904.10631.

Jakob Straub. 2024. Die 20 besten schweizer sprichwörter und alltagsweisheiten. *Lingoda.com*. Accessed: 2024-12-01.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

GOOSVC: Version Control for Content Creation with Generative AI

David Grünert[1,2], **Alexandre de Spindler**[1] and **Volker Dellwo**[2]
[1]Zurich University of Applied Sciences, Winterthur, Switzerland
[2]Department of Computational Linguistics, University of Zurich
{grud,desa}@zhaw.ch

Abstract

This paper introduces GOOSVC[1], a version control system for content creation using generative AI. As generative AI models become integral to creative workflows, managing iterative changes, branching, and merging of content is challenging. Current version control systems are not designed for these workflows, which involve multiple AI assistants exchanging text, images, or other artifacts. In this paper, we identify the core requirements for such a system and show how GOOSVC meets them. Our system provides full traceability and versioning of both artifacts and conversation states, allowing seamless integration of multiple AI assistants into creative workflows.

1 Introduction

Generative AI has rapidly evolved into a powerful creative partner in domains such as marketing, design, data science, and creative writing (Davenport and Mittal, 2022). Individuals and teams often rely on large language models (LLMs) or multimodal AIs to brainstorm ideas, refine concepts, and generate or revise content (White et al., 2023). Despite these successes, complex workflows, in which users combine multiple AI assistants, pose significant challenges. Studies confirm that creative work requires fluid human–AI co-creation (McGuire et al., 2024; Rezwana and Maher, 2023). Such co-creation is rarely linear: users frequently need to revise, branch, or revert to earlier prompts, and must manage a growing collection of text, images, audio files, or other artefacts in the process (Cygnis, 2024; Kumar and Suthar, 2024; Coca-Cola, 2023). Because of the non-deterministic nature of generative AI, but also for legal reasons (European-Commission, 2020), AI-assisted workflows must offer end-to-end traceability of all generated artefacts *including* all AI interactions.

[1]https://goosvc.com

While conventional version control systems (VCS) such as Git track file changes and allow branching and merging, they are not designed to handle dynamic AI-generated outputs or iterative dialogue histories. Existing generative AI platforms (e.g., ChatGPT, Microsoft Co-Pilot, Google Gemini) store chat logs and generated files, however, they lack robust mechanisms to manage branching workflows, merge parallel conversation threads, or revert selectively to earlier states. Versioning of AI models has been discussed in prior work (Vadlapati, 2024), but to our knowledge no system explicitly supports full-versioning of generative AI interactions alongside the content they create—particularly when multiple AI assistants are used in parallel. This gap often forces users to adopt fragmented workflows, where they manually copy AI outputs, store them in external tools, and struggle to piece together a coherent project history.

In this paper, we address this gap by introducing a novel version control approach that captures both AI-driven conversations and their resulting artefacts within a single, integrated framework. Specifically, our work makes the following contributions.

- We identify the core challenges for an integration of AI assistants into iterative, multimodal workflows. From these observations, we derive the requirements for an AI-focused VCS.
- We propose a new VCS that treats every user prompt, AI response, and generated artefact as part of a unified version history, enabling branching, merging, and reverting at both the project and conversation levels while offering end-to-end traceability for each artefact created.
- We demonstrate the system's practicality through a data science application that generates synthetic datasets using GOOSVC as underlying VCS.

The remainder of this paper is organized as follows. Section 2 introduces a detailed use case to motivate the requirements for versioning generative AI workflows. We then review related work in generative AI interfaces and traditional VCSs in Section 3, highlighting their limitations. Section 4 describes our proposed system's architecture, data model, and merging strategies. Section 5 showcases a real-world demonstration of our approach in synthetic dataset creation, and finally Section 6 provides concluding remarks and outlines directions for future work.

2 Content Creation Use Case

Consider a creative director tasked with producing an advertisement clip for a new, innovative product. The process begins by defining a target persona and mapping out their journey, capturing key emotional touchpoints and decision-making moments. Next, the director articulates the product's value proposition and envisions how it can transform the persona's experience, weaving these elements into a compelling narrative. From that, multiple iterations of storylines are developed and refined—from initial concepts to detailed scene descriptions. The final storyboard emerges as a composite artefact that combines descriptive texts with illustrative images and may also include audio or video elements for more immersive storytelling.

We assume that the creative director uses an authoring tool for this task. In principle, it is possible to use AI assistants for each of these work steps. For example, a large language model can help brainstorm ideas for the persona and the journey, propose story lines and a multimodal AI can generate images and text for the storyboard. Additionally, specialized generative models can produce audio or video prototypes. Based on this use case, we will now identify typical procedures and derive the requirements that are placed on an underlying VCS used by the authoring tool.

2.1 Iterative Development of Artefacts

In creative workflows, it is common to iteratively develop artefacts. For instance, the creative director may want to refine persona sketches in multiple iterations. When using an authoring tool offering AI assistance for this task, the director may need to adjust the prompts to elicit more detailed responses or to clear up misunderstandings. When the authoring tool wants to send these prompt to a chat-based assistant, such as ChatGPT via API, any request must include the complete chat history. Therefore, the chat history for every generated artefact must be stored. To support this, a VCS must:

R1 Provide a mechanism to version artefacts together with their chat history.

2.2 Using Multiple Assistants

In creative workflows, it is common to use multiple AI assistants. These assistants may be used independently of each other or in a collaborative way. For instance, the creative director may use a large language model to develop a storyline in collaboration with a multimodal AI to generate images for the storyboard. Or they may use multiple instances of the same AI model with different roles to investigate different perspectives like the view of the customer and the view of the service provider onto the product.

To support iterative development with multiple assistants, the authoring tool must store the chat history and the generated artefacts for every assistant separately. This is necessary to keep the chat context clean for every assistant and to prevents unintended cross-contamination of different chat contexts. Furthermore, as long as there is no interaction between the assistants, using separate contexts allows to revert one chat to a previous state without affecting the others. However, to allow collaborative use of assistants, it must be possible to share artefacts between the contexts. To support this, a VCS must:

R2 Provide a mechanism to create and manage chat contexts for multiple assistants that contain artefacts and chat histories.

R3 Provide a mechanism to share artefacts between chat contexts.

2.3 Reverting to Previous Versions

In creative workflows, it is common to revert to previous versions to revise decisions or to correct mistakes. For instance, the creative director may want to revert to an earlier persona sketch or revisit a previous storyboard to incorporate a discarded scene. For this task, an AI-assisted authoring tool should support two types of revert: reverting the complete project to a previous state and reverting a single chat to a previous state. When reverting a project, all artifacts and the related chats must be reverted. When reverting a single chat, only artifacts

generated in this chat must be reverted. However, the causality must be maintained between the chat and the project. For instance, if an artefact is reverted that has been used elsewhere in the project. To support this, a VCS must:

R4 Provide a mechanism to revert a project to any previous version including all chats and their artefacts.

R5 Provide a mechanism to revert a chat including the generated artifacts to any previous version while preserving the causality between the chat and the project.

2.4 Creating Variants of Workflows

In creative workflows, it is common to create multiple variants to explore different ideas or evaluate the impact of changes. For instance, the creative director may want to generate several versions of a storyboard to compare different visual styles or experiment with alternative personas. When using chat-based assistants via the provided API, the authoring tool must keep track of alternative paths because any request must include the complete chat history. To support this, a VCS must:

R6 Provide a mechanism to start alternative paths from any previous version including all artefacts and their associated chat histories without losing the progress made so far.

2.5 Combining Parallel Workflows

It is common in creative workflows to parallelize work on different parts of a project to increase efficiency. For instance, the creative director may want to distribute the work on different parts of the storyboard among several team members. To get the final storyboard, the authoring tool needs to combine the results of parallel workflows. During this merge, two types of conflicts may arise: Conflicts on artefacts occur when the same artefact is changed in multiple branches. Conflicts on chat histories occur when the same chat was continued in multiple branches. Furthermore, these merges will often include more than two branches. To support this, a VCS must:

R7 Provide a mechanism to merge any number of parallel workflows including their artefacts and the associated chat histories offering methods to resolve conflicts on artefacts and on chats.

2.6 Defining Stages in Workflows

In creative workflows, it is common to define stages to structure the creative process. For example, the creative director may want to define stages for the definition of the personas or the definition of their journeys. Stages help simplify to revert to defined milestones, to create variants and to parallelize workflows. To achieve that, an authoring tool must guarantee that the stages are unique at any time within the project history. This must be ensured when stages are added but also when parallel workflows are merged. To support this, a VCS must:

R8 Provide a mechanism to define stages in the version history and to keep these stages unique with the project history also when merging parallel workflows.

2.7 Summary

These procedures derived from our use case reveal several critical challenges for a version control system that is used by an authoring tool for creative workflows. First, there is an urgent need for **dual versioning with contexts** (**R1**, **R2**, **R3**) that maintains a direct link between evolving artefacts and the underlying AI-driven conversations that produce them. Second, the ability to **revert projects and chats** (**R4**, **R5**) is crucial for revising decisions and recovering from mistakes. When reverting single chats, the system must maintain causality between the chat and the project. Third, the iterative nature of creative work requires robust **automatic branching** (**R6**) to explore alternatives without losing previous progress. Fourth, while parallelization can enhance efficiency in creative workflows, **merging parallel paths** (**R7**) requires a mechanism to combine multiple branches and resolve conflicts on artefacts and chats. Finally, the ability to define and maintain **project stages** (**R8**) in the creative process is essential for structuring complex workflows and project planning.

Every requirement above targets a distinct dimension of AI-assisted content creation, ensuring coverage of iterative development, collaboration among multiple assistants, safe reversion, parallel exploration, and structured milestone definition. Together, they form an orthogonal set that comprehensively addresses the challenges identified in this use case. Overall, these requirements address the nonlinear, multimodal, and iterative nature of generative AI use cases introduced in Sect. 1.

3 Background

In this section, we first assess how existing generative AI tools and interfaces manage iterative creative workflows and highlight their current limitations in terms of traceability and prompt reuse. We then turn to VCS as a potential source for more advanced branching and merging concepts. We go on to evaluate how well these approaches fulfil the requirements described in section 2, and finally identify the key gaps that motivate our solution.

3.1 Generative AI Interfaces

Contemporary generative AI front-ends such as ChatGPT, Microsoft Co-Pilot, Google Gemini and Anthropic Claude have transformed content creation by delivering multimodal outputs and allowing users to refine prompts on the fly. Despite these strengths, they offer limited support for complex, iterative workflows that require branching, merging and robust versioning. While some interfaces allow users to revisit or modify previous prompts, each output is still treated as an isolated event, and complex artefacts or composite data sets are not intrinsically linked to the conversation history that produced them. This missing link makes end-to-end traceability difficult: although users can see the final product, they have no systematic way of exploring the creative process that led to that outcome.

Popular AI front-ends often confine interactions to a single chat context, making it difficult to collaborate across multiple AI models or to run parallel explorations of the same artefact. In particular, if AI models from different providers are involved, their interaction cannot be documented. The functions available within the web interface, such as storing or editing a chat history especially from different models, are vendor-specific, manual processes. When these models are invoked programmatically via an API, the application developer must transmit the entire conversation history again for every request and in addition manage the branching logic. As a result, key requirements such as **R6** (starting alternative paths), **R7** (merging parallel paths), **R8** (defining stages in workflow), and **R1** (versioning artefacts with their chat history) remain uncovered. Consequently, creators either do without the possibility of branching and merging or resort to inefficient workarounds such as copying intermediate outputs, duplicating prompts, and manually tracking versions outside the AI tool.

In summary, while modern generative AI systems excel at generating rich content, they offer minimal native support for iterative, branching workflows. This limitation hinders the kind of controlled exploration and traceability that creators increasingly need when integrating AI into complex projects.

3.2 Version Control Systems

VCSs have long been essential for tracking and managing changes across software projects and other text-based repositories. Traditional systems, such as Subversion and Git, typically provide several core capabilities. First, they maintain a chronological sequence of changes, known as linear versioning, which preserves a historical record of modifications. Second, they allow branching, so that work can proceed in parallel lines of development, making it possible to explore experimental features or maintain distinct configurations. Third, these systems include merging functionality, enabling divergent branches to be reconciled into a unified project state. Finally, they allow projects to maintain different variants through mechanisms that can track concurrent releases or alternate product lines.

While these mechanisms provide a solid blueprint for managing project histories, they were never designed to track interactive conversations or dynamically generated content from AI models. Such conversations must be versioned in a manner similar to code commits, yet cannot be handled by line-based diffs. Conventional VCSs distinguish between text and binary files, both of which are inadequate for storing conversational histories. This is because requirements such as **R1** (versioning artefacts with their chat history), **R2** (managing multiple chat contexts), **R3** (sharing artefacts between chat contexts), and **R5** (reverting selected chats while preserving causality) are not met.

In light of these gaps, where generative AI tools lack integrated branching, merging, and revert capabilities, and where conventional VCS fail to capture conversational histories, we propose a specialised versioning framework tailored to AI-driven creative workflows. Our approach unifies conversation and artefacts tracking, addresses branching across multiple assistants, and integrates robust merging mechanisms, laying the groundwork for end-to-end traceability and prompt reuse in generative AI projects.

4 Versioning Approach

In this section, we present our new versioning approach and show how the requirements listed in 2.7 are addressed. Our implementation GOOSVC (Grünert, 2025) is designed to be used in a production environment considering operational aspects such as performance, scalability, and security. As shown in Figure 1, users will typically not interact directly with GOOSVC. The goal is to simplify the integration of AI assistants into workflows, enabeling seamless branching, reverting, and merging, and to offer traceability across both the prompts and the generated content.

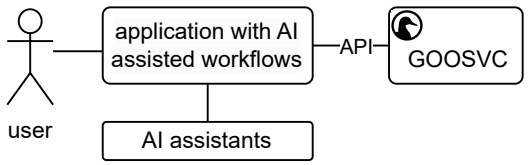


Figure 1: System overview

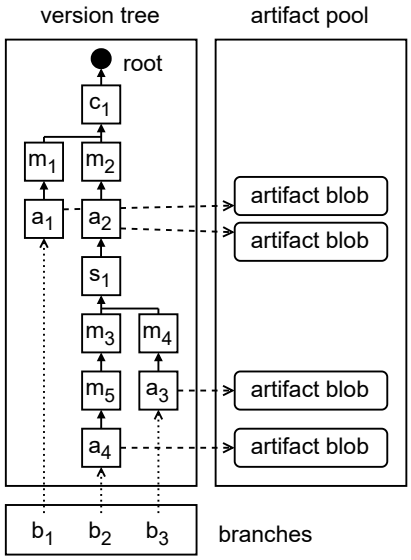


Figure 2: Data model example showing a version tree with three branches (b_1 to b_3) containing chat- (c), message- (m), stage- (s) and artifact-nodes (a) with references to the artifact pool.

Figure 2 presents the core elements of the data model. It consist of an immutable node-based version tree, where every interaction is recorded as a distinct node with associated metadata (e.g., parent node, time stamp, author, committer, and type specific data), an immutable pool for storing generated artifacts, and branches to keep track of all available paths. Nodes have one of the following types:

- **Chat Nodes**: Add a new chat context to the project for interactions with an AI. Every chat context has a unique ID later used by messages and artifacts to declare their affiliation. Chats can be started either from scratch or with a parent by referring to a message node. Chats with parents are interpreted as a continuation from the referenced message.
- **Message Nodes**: Store prompt-response pairs that capture the conversational exchange with an AI. All messages must be associated with a chat context.
- **Artifact Nodes**: Store artifacts (text, images, audio) by referencing the actual data in the artifact pool and defining metadata such as the filename. All artifacts must be associated with a chat context.
- **Stage Nodes**: Define named project milestones. Stage names must be unique within any path of the version tree.
- **Merge Nodes**: Document a merge operation of parallel branches. Merge nodes are not shown in Figure 2. Details are described in section 4.4.

4.1 Dual Versioning with Contexts

When inserting nodes into the version tree, the position of the new node must be defined. This position can either be a branch or an existing node. When a branch is used, the new node is appended to the branch. When a node is used, the system inserts the new node as a child of the given node. Depending on the parent's position, this will either create a new branch or continue a branch (see 4.3).

The system establishes a link between AI interactions and their generated artifacts by capturing both within the same version tree. All prompt-response pairs are stored as message nodes in the version tree, referencing a chat context and thereby forming a complete lineage of the conversational history. Similarly, every artifact is represented as an artifact node, referencing both the chat context used to create it and the artifact in the pool. When storing or changing artifacts, the system adds a new artifact node containing all metadata such as file name, path, scope (chat, global) and the operation (add, update, rename, delete). Artifacts in the pool are immutable, therefore, when artifacts are added or updated, a new artifact is added to the pool.

Every node of the version tree represents a version of the project. This version is defined as the union of all nodes along the path from that node back to the root. When retrieving a selected project version, the system can either return the complete project or only data from a specific chat context. If the path contains multiple artifacts with the same combination of path and filename, the latest node masks all older ones. For any version of a project, only one artifact for a given path-filename combination is visible. If the last node for a given path-filename combination has the operation delete, there is no such artifact in the respective version. Path and filename are both defined by the workflow application. Similar to popular VSC, The complete set of all artifacts can be checked out to the local file system for any version of the project.

While messages are always limited to one chat context, the scope of artifacts can be set. If scope is set to *chat*, the artifact is only visible within the chat context. If set to *global*, the artifact is visible in all chats. Setting the scope of an artifact to global allows to share artifacts between chats. This dual versioning approach with contexts ensures that every creative decision is fully traceable, enabling users to audit the entire workflow and understand the context behind every artifact generated within the project.

4.2 Revert Projects and Chats

Our system supports two types of revert: reverting the entire project to a previous state and reverting a single chat to a previous state. When reverting a single chat, only the artifacts within the chat's context and their associated chat history are reverted. In contrast, reverting a project to a previous state will revert all chats and their artifacts.

Reverting a single chat is achieved by inserting a new chat node that references the previous state as parent. New messages added to the chat after the revert must then be associated with this new chat. Figure 3 shows an example of reverting a chat. Subfigure a) shows the project before the revert containing two chats: c_1 (m_1, m_3) and c_2 (m_2, m_4). Subfigure b) shows the project after reverting the chat c_1 to message m_1 and adding an additional message m_5. Reverting is achieved by adding c_3, referencing m_1 as parent. The chat is then extended with message m_5. After this revert, the project contains three chats: c_1 (m_1, m_3), c_2 (m_2, m_4), and c_3 (m_1, m_5).

Reverting a project is achieved by branching off from the previous state. The new branch is then used to continue the project. Figure 3 Subfigure c) shows the project after reverting the entire project to message m_1 and adding the additional message m_5 to chat c_1. The project is branched off from m_1, and c_1 is extended with m_5. The project in branch b_2 contains chat c_1 with the messages m_1 and m_5.

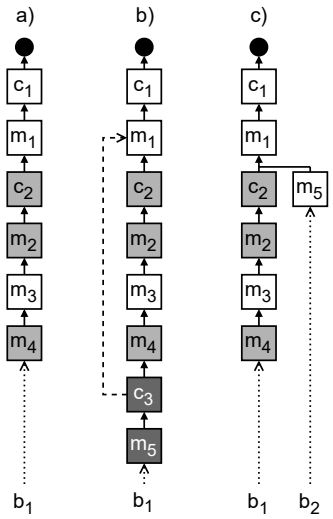


Figure 3: a) Original project with two chats (c_1, c_2). b) Project after reverting **chat** c_1 to m_1 and adding additional message m_5. c) Project after reverting **project** to m_1 and adding additional message m_5. Nodes with the same background color belong to the same chat context.

4.3 Automatic Branching

In our system, branching is used for three different purposes: Variants, reverting, and parallelization. Variants are used to explore different ideas or evaluate the impact of changes. Finally, one variant is selected to continue with. Reverting is used to go back to a previous state and continue working from there by branching off. Parallelization is used to increase efficiency by working simultaneously on different parts of a project. The results of parallel workflows are merged to create the final output (see 4.4).

The creation of these branches is not always a conscious decision. Often, they emerge naturally as the creative process unfolds. To capture this organic branching, our system automatically creates new branches whenever a node diverges from an existing path. Instead of using names for branches, the system uses unique identifiers. These identifiers are used to reference a branch when appending nodes or merging branches.

4.4 Merging Parallel Paths

Merging creative workflows is used to combine the work of a parallelized sections of a project. Such a merge may include more than two branches. Furthermore, the merge does not necessarily include the complete branch up to its head. In the example shown in Figure 4, branch b_4 is merged with n_8 from b_2 and n_6 from b_3.

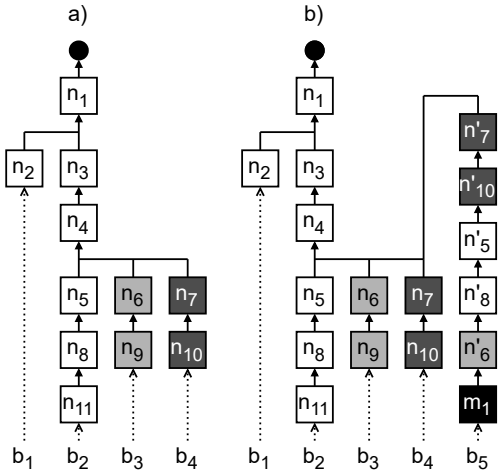


Figure 4: Example of a conflict-free merge with nodes n of unspecified type. a) Original project before the merge. b) Project after merging b_4, n_8 and n_6. Nodes with the same background color belong to the same branch before the merge.

Merging is achieved by replaying all nodes that follow the first common ancestor into a new branch to create a unified, sequential history. At the end of this sequence, an additional merge node is added to document the merge. In the example shown in Figure 4, branch b_5 contains this sequence and the additional merge node (m_1).

As introduced in 2.5, two types of conflicts my arise when merging parallel paths: Chat conflicts, when the same chat was continued in multiple branches, and artifact conflicts, when the same artifact was changed in multiple branches. Chat conflicts are resolved as follows: the system splits the dialogue automatically into two chats with a shared history before the first common ancestor. This approach ensures that the context of every conversation remains intact, even when the content diverges. In the example shown in Figure 5, branches b_2 and b_3 both continued the chat c_1. To resolve the conflict when merging b_2 and b_3, the system creates a new chat (c_2) with the common history of c_1 before the divergence. Message m'_4 and artifact a'_3 are then both added to c_2.

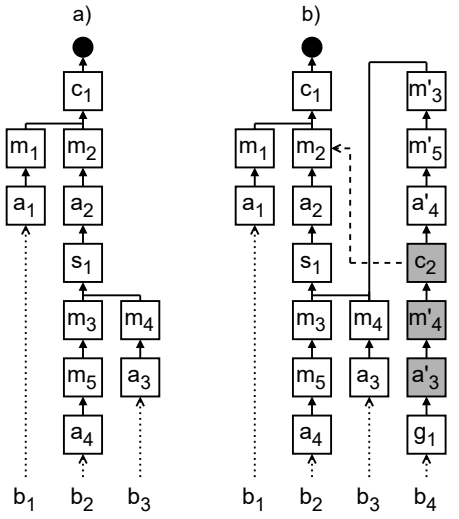


Figure 5: Example of a merge with conflicting chats a) Version tree before the merge. b) Version tree after merging b_2 and b_3. Nodes with the same background color belong to the same chat context.

The resolution of artifact conflicts depends on the file type. For text files, the system may automatically merge the changes, if independent sections were modified. For binary files, the system will rename the files. Both cases may require manual review of the result. In general, merge conflicts on artifacts should be avoided. For most situations, parallelized work only makes sense if the work is independent.

Thus, the merging mechanism effectively integrates parallel creative paths while resolving conflicts automatically. Instead of relying on standard diff-based methods, our approach tailors conflict resolution to the nature of the content, ensuring that the creative process remains fluid and efficient.

4.5 Project Stages

Stages function as immutable checkpoints within the project history. Therefore, stages are not associated with a chat context. Every stage marks a milestone that remains unchanged regardless of subsequent iterations, offering stable reference for reverting or branching the project. Stages are implemented as stage nodes in the version tree. The system ensures that stages are unique within any path of the version tree. The system refuses to add a stage if the name already exists. Also when merging parallel workflows, stages must be kept unique. To achieve this, the system refuses to merge branches containing any stages. Stages represent milestones for the entire project. Adding a

stage via a merge would contradict this concept. In summary, stages provide fixed anchors in the creative process, ensuring that pivotal moments remain preserved and clearly defined.

5 Demonstration

To demonstrate the flexibility and real-world utility of our version control approach, we applied it to a complex workflow that generates synthetic crime datasets for research. Data recorded during criminal investigations is often confidential and therefore unavailable for research. Existing datasets from other domains do not share the characteristics of crime-related data, which typically include telephone recordings, audio surveillance with varying quality, multilingual and emotional speech, and background noise containing relevant information. Moreover, higher-level analyses such as communication structure detection require the spoken content to match the context of actual criminal cases.

To address these challenges, we presented a workflow in (Grünert et al., 2024) that generates synthetic datasets from a case outline (see Figure 6). Specifically, it uses LLMs to produce transcripts of conversations and messages related to a hypothetical criminal case. This involves 22 different prompt templates and over 400 individual requests to LLMs. Next, these transcripts are annotated with emotions and timing aspects and then converted to audio files. Background noise and signal processing are subsequently applied to create realistic acoustic variations. The final dataset comprises text messages, audio files, and annotations (RTTM, TextGrid), making it suitable for research on speech analysis or communication structures.

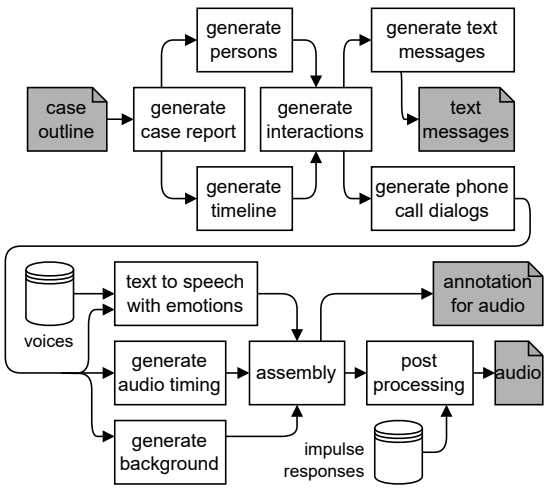


Figure 6: Case generation pipeline

Using GOOSVC, we developed an interactive web application that orchestrates every phase of this workflow while providing robust version control capabilities. One key advantage is the ability to manage distinct stages in the dataset generation process. Users can revert to any prior milestone (**R4**, **R5**) and make adjustments without having to restart the entire pipeline. For instance, if a user wants to revise how two suspects interact in the transcripts or modify the background noise level, they can branch off from the relevant stage, edit just the targeted prompts and parameters, and then regenerate only the affected outputs—preserving all other completed work.

Furthermore, for every artifact created, the system automatically stores the associated AI interactions. This provides end-to-end traceability (**R1**), allowing users to see which prompts and responses led to a specific audio track, transcript, or annotation. The same approach also supports branching out (**R6**) into parallel workflows—such as exploring different emotional tones for conversation—before merging them (**R7**), if needed. As a result, workflows deemed as successfull can be conveniently reused and adapted for new case outlines, emphasizing the flexible and iterative nature of GOOSVC.

6 Conclusion

In conclusion, we have introduced a novel version control approach tailored for generative AI-driven content creation. Our VCS captures both AI interactions and the resulting artifacts in a unified system offering branching, merging, and stable milestones. By addressing the key challenges of iterative creative workflows—such as maintaining traceability, managing parallel explorations, and resolving content-specific conflicts—our approach offers a robust framework that enhances reproducibility and flexibility. This work not only streamlines the creative process but also lays the groundwork for future enhancements in collaborative AI-driven design and content management.

References

Coca-Cola. 2023. Coca-cola invites digital artists to 'create real magic' using new ai platform.

Cygnis. 2024. Best practices for implementing ai workflow automation in enterprises.

Thomas H. Davenport and Nitin Mittal. 2022. How generative ai is changing creative work.

European-Commission. 2020. White paper on artificial intelligence-a european approach to excellence and trust.

David Grünert. 2025. Goosvc github repository. https://github.com/goosvc/goosvc.

David Grünert, Dominic Pfister, Alexandre de Spindler, and Volker Dellwo. 2024. Generating synthetic datasets for the validation and training of automatic speech analysis systems in the context of organized crime. 2nd VoiceID conference, Marbug, Germany.

Dinesh Kumar and Nidhi Suthar. 2024. Ethical and legal challenges of ai in marketing: an exploration of solutions. *Journal of Information, Communication and Ethics in Society*, 22(1):124–144.

Jack McGuire, David De Cremer, and Tim Van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports*, 14(1):18525.

Jeba Rezwana and Mary Lou Maher. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*, 30(5):1–28.

Praneeth Vadlapati. 2024. Updagent: Ai agent version control framework for real-time updation of tools. *International Journal of Science and Research (IJSR*, 13(11):628–632.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

LLM-based Translation for Latin: Summaries Improve Machine Translation

Dominic P. Fischer, Martin Volk
University of Zurich
Department of Computational Linguistics
dominicphilipp.fischer@uzh.ch

Abstract

Recent studies demonstrated that modern Large Language Models set a new state-of-the-art in translating historical Latin texts into English and German. Building upon this foundation, we investigate the impact of incorporating text summaries into prompts for LLM-based translation tasks. Having both the historical text and a modern-language summary is a typical setup for classical editions. Our findings reveal that integrating summaries significantly enhances translation accuracy and coherence.

Keywords: Large Language Models, Machine Translation, Summaries, Latin, Historical Letters

1 Introduction

Summarizing is an essential task for editors of historical texts. Editors create summaries in modern languages to distill the complex and extensive information found in historical documents, ensuring that the core message and significant details are preserved. This editorial practice not only aids in efficient information retrieval but also maintains the integrity and context of historical records.

Historical texts are therefore often accompanied by summaries. This provides a valuable opportunity to exploit the modern language summary when processing the historical text with an AI system. We suspect that LLMs may profit from the expert-distilled information in the summaries.

This paper explores the impact of manual summaries in LLM-based machine translation for Latin to English and German translation of 16th century letters, following up on (Volk et al., 2024a). We hypothesize that providing LLMs with well-crafted summaries will enrich the translation process, yielding superior quality text in the target language. The translation of the full text provides the complete rendering of the original content and thus allows for a more comprehensive analysis than only relying on the summaries.

Pairing the letter text with its summary as input to machine translation not only highlights the practical applications of LLMs in historical research but also underscores the value of editorial practices in the digital age. By combining the strengths of human expertise and advanced AI, we push the boundaries of what can be achieved in the translation of historical texts.

Our research is in the line of research on prompting strategies for LLM-based machine translation (Zhang et al., 2023; He et al., 2024) which focuses on the impact of providing translation examples. We are the first to test the impact of providing a target language summary together with the source text.

Adding the summary is a form of knowledge injection through the prompt. Similar to the integration of domain-specific terminology to a prompt (as in Bogoychev and Chen (2023)), and similar to adding translation suggestions from lexical footnotes (Volk et al., 2024b) or from bilingual dictionaries to the prompt (Ghazvininejad et al., 2023). The latter show that LLM prompting provides an effective solution for rare word translation, by using knowledge from bilingual dictionaries. Yao et al. (2024) introduce various strategies to incorporate external and internal cultural knowledge into the prompt. Strategies include self-explanation and self-ranking to activate the relevant knowledge of the LLM.

2 Corpus and Methodology

For our evaluation, we utilized the test set from (Fischer et al., 2022), which includes eight Latin letters manually translated into German by a domain expert. This test set comprises 121 Latin sentences, ranging from short greetings to sentences as long as 47 words, totaling 1240 words in Latin and 1768 words in the corresponding German translations. To adapt this test set for translation into English, we

used GPT-4 to automatically translate the German texts into English.

The letters are taken from the 16th century letter exchange of the Zurich Reformer Heinrich Bullinger. 3200 of the letters have been published in 21 printed volumes over the last 50 years by the Institute for Swiss Reformation Studies[1], professionally edited, summarized and extensively commented.[2] Depending on the volume, the summary length and format varies. The German summaries are as short as a few sentences in the first volumes (published in the 1970s) and then increase in length to being more comprehensive.

For example, the letter from Berchtold Haller to Heinrich Bullinger (February 1532, published in volume 2; not part of the test set) consists of 609 tokens in Latin (5 lengthy paragraphs plus initial greetings and letter closing). But the editors summarized it with only one paragraph (68 tokens) in German:

- *Berichtet von der Lage nach der Berner Synode, deren Akten bald im Druck erscheinen werden und worüber er Bullingers Meinung erfahren möchte. Bittet um Antwort auf verschiedene Fragen, um die Zusendung von Bullingers und Pellikans Werken, macht Vorschläge für eine Annäherung zwischen Bern und Zürich und betont, daß Zwinglis Sohn Wilhelm in Bern unter den besten Voraussetzungen erzogen wird. Grüße.*
  (Reports on the situation after the Bern Synod, whose records will soon be published, and wishes to hear Bullinger's opinion on the matter. Requests answers to various questions, the sending of works by Bullinger and Pellikan, and makes suggestions for a rapprochement between Bern and Zurich. Emphasizes that Zwingli's son, Wilhelm, is being educated in Bern under the best conditions. Sends greetings.)

Starting from volume 16 (published in 2014), the summaries are done paragraph by paragraph, covering the entire letter. These summaries can be seen as shortened German paraphrases of the letter. Still, as from the first volume of the edition, the summaries are written as a description of the letters ("The author X reports on the situation after the Bern synod, ...") in contrast to the letters themselves that are written from a personal perspective ("After the synod was concluded, ...").

The eight letters in our test set are taken from volumes 14, 15 and 16; three of them have paragraph-wise summaries. The summary lengths range from 54 to 428 tokens with the ratios of summary length to letter length ranging from 0.43 to 0.96 (cf. Table 5).

For the LLM-based translation of the test set without and with the summary, we employed the following two prompts:

- Without summary: Translate the following Latin text into German/English while keeping the formatting as it is: *Latin text*.
- With summary: Translate the following Latin text into German/English: *Latin text*. Keep the formatting as it is. As a help for your translation, consult this summary: *summary*.

Additionally, we tested whether GPT-4 performs better at translating a letter when it is aligned with the sentences of the corresponding summary. For this purpose, we manually aligned the sentences of the summary with the letters, inserting them in brackets after the sentence they refer to, like in Table 1.

[...] Nihil certi ex comitiis audio. Expectatur adhuc Ferdinandus rex. **(The Reichstag [in Speyer] is still waiting for King Ferdinand.)** [...]

Table 1: Letter with aligned summary sentence (Johannes Gast to Heinrich Bullinger on 1.4.1544)

Automatic alignment with GPT-4 provided results with only minor discrepancies with regards to the human alignment, indicating a promising avenue for automatic text-summary alignment. For the purpose of the experiment, however, we used human alignment to avoid inducing any errors.

For the translation with aligned summaries, another two prompts - one as short as possible, one more descriptive - were used:

- Translate the following Latin sentences into German/English. Use the sentences in brackets to guide your translation. Preserve the formatting: *Latin text with aligned summary*
- Translate the following Latin letter into German/English. The lines in brackets are from a

[1] https://www.irg.uzh.ch/

[2] The complete preserved Bullinger correspondence consists of 12,000 letters.

summary of the letter and have been aligned, so that they explain the preceding lines. Take them into account, but do not output them in your translation. Keep the line breaks as they are: *Latin text with aligned summary*

3 Main Findings

The analysis of translation quality revealed notable improvements when summaries were included, as indicated by both BLEU (SacreBLEU) and ChrF scores[3]. However, this only applies if the summary is in the same language as the target text, as is illustrated in Tables 2 and 3. When translating the test set into English, the BLEU score increased only marginally from 32.1 to 32.5 when given the German summary, but increased significantly by 2.3 points to 34.4 with an English summary (which we automatically translated from German). When translating into German, the increase in BLEU is 2.0 points when given the German summary.

Table 2: Translation Quality Scores (BLEU) on the test set with and without summaries.

Testset	No Summary	W/ Summary
DE	25.8	**27.8** (DE)
EN (GPT-4)	32.1	32.5 (DE)
		34.4 (EN)

Table 3: Translation Quality Scores (ChrF) on the test set with and without summaries.

Testset	No Summary	W/ Summary
DE	51.6	**53.3** (DE)
EN (GPT-4)	52.6	53.4 (DE)
		54.7 (EN)

While these BLEU score increases of 2.3 for English and 2.0 for German were similar, the absolute values of the BLEU scores are higher for translations into English. We will discuss the reasons for this in section 4. With regards to the ChrF scores, we see the same trend - an increase of about 2 points when summaries are included, yet interestingly, the difference in absolute values between English and German is only marginal (cf. Table 3).

[3]BLEU evaluates translation quality by measuring the overlap of sequences of n words (so-called n-grams) between the machine-generated and a reference translation, while ChrF uses overlapping *character* n-grams.

The experiments with the aligned summaries showed interesting results. With the simple prompt, the results were the same or only slightly better (~1 BLEU/ChrF point) than the translation *without* summary.

The longer, more descriptive prompt yielded different results in German and English. In German, the results were worse than with the simple prompt, with almost the same scores as without summary. For English, this resulted in the best translation yet, surpassing the translation scores with target language summary by 0.9 BLEU points and 0.7 ChrF scores (cf. Table 4). Nevertheless, this approach did not yield consistent improvements, as illustrated by the wrong translation in the last row of Table 6.

Table 4: BLEU and ChrF scores for translation with aligned summary in the target language.

	With Aligned Summary	
	BLEU	**chrF**
P1: DE	**26.8**	**52.2**
P2: DE	25.7	51.8
P1: EN	32.4	53.2
P2: EN	**35.3**	**55.4**

Table 5 shows that longer summaries, or summaries that cover more of a given letter do not necessarily lead to greater improvements in translation. At the same time, even short summaries (as short as a single sentence) can lead to significant quality increases. It therefore stands to reason that situating the letter, outlining its content and the actors therein is an efficient way of injecting crucial information for translation quality gains.

letter id	#tok.L	#tok.S	ratio	ΔBLEU
12151	244	105	0.43	4.3
11916	180	96	0.53	5.31
11898	98	54	0.55	3.33
12838	98	54	0.55	1.31
11930	179	109	0.61	-0.39
12378	106	67	0.63	0.02
12154	172	157	0.91	2.62
12509	444	428	0.96	0.11

Table 5: Comparison of letter (L) and summary (S) token counts, ratio, and BLEU improvement measured between without and with summary. The entries are ordered ascendingly by ratio. (11898 and 12838 happen to have the same counts for summaries and letter texts.)

Latin original	Commissum habeas adolescentulum; polliceor et ego me non ingratum fore.
EN reference	I recommend the young boy to you; I assure you that I too will not be ungrateful.
EN without summary	You may have committed the young man; I also promise that I will not be ungrateful.
EN with summary in DE	You have the young man in your care; I promise that I will not be ungrateful.
EN with summary in EN	**You may have the young man in your care**; I promise that I will not be ungrateful.
EN with aligned summary in EN	You have a committed young man; I promise that I will not be ungrateful.

Table 6: Translations of the Latin sentence without summary and with summary in German and English. The Latin conjunctive 'commissum habeas' only gets correctly translated with the English summary in the prompt to 'You may have'.

This is supported by our qualitative analysis of the summary-induced effects in the German translations (cf. Table 7). To that end, we manually compared the 121 test set sentences when translated with and without summary. It results that 51 stayed the exact same, while 70 contained changes. Out of these 70, 36 contained minor neutral (word choice) changes, 23 minor positive changes, and only 4 minor negative changes. 7 sentences contained major positive changes, including changes crucial to the correct understanding of the sentence and major changes in the sentence structure.

	amount	**percentage**
the same	51	42
different	70	58
... *of which*		
neut. (≈)	36	30
pos. (+)	**23**	**19**
neg. (-)	4	3
major pos. (++)	**7**	**6**

Table 7: Overview of changes induced by including the summaries in the prompt.

Classified as "minor" were changes of often one, sometimes multiple (max. 3) words. Minor positive changes contained predominantly name corrections/normalizations (*Marcus* → *Markus*, *Caesar* → *Kaiser Karl V.*), clarifications of pronouns (*these* → *these news*, *he* → *it*), and previously missed precisions (*an answer* → *any answer*). Negatives included wrongful reversals of such things, like *these questions* → *these*, *pray to the Lord* → *pray*. The major positive changes greatly affected the understanding of the sentence, including changes of modus (imperative → conjunctive) or of an entire (part of a) sentence, such as in table 6. Major negative or neutral changes were not present.

4 Discussion

The observed improvements in translation quality with the inclusion of summaries can be attributed to the additional context provided by the summaries. This context helps the LLMs generate more accurate and coherent translations by offering clear guidance on the essential points and context of the text.

The better performance of English translations with regards to BLEU could be linked to two main factors. Firstly, the gold standard translation of the letters in English are a GPT-4 translation of the German gold standard, which might have introduced a bias towards higher scores due to the model's own translation capabilities. This could mean that the English summaries were inherently more aligned with the model's strengths.

Secondly, GPT-4 and similar LLMs are extensively trained on English language texts, leading to inherently better performance in English due to the abundance of training data and resources. This extensive training allows the model to produce English text with higher accuracy and fluency, as has been observed in other studies.

The first above point implies that the quality of the English translation is not actually significantly better than the German translation, it merely appears to be because of the skewed English transla-

tion. As the ChrF scores are very close between translations into German and English, ChrF seems to balance this skewness.

A reason might be ChrF's indifference to the structural differences of the two languages. For example, German has a more flexible word order and often requires reordering phrases in translation to maintain grammatical correctness. This can result in lower n-gram overlap in BLEU scores because SacreBLEU heavily relies on exact matches of words and phrases. Similarly, the morphological complexity of German works against the exact matching of word n-grams that BLEU measures, and is better suited to character-level comparisons like ChrF.

In other words: BLEU amplifies the skewness, since it looks for exact matches of n-grams, which are more likely to be present if the reference itself is a product of GPT-translation.

Our findings suggest that including a sentence-aligned summary in the prompt for translation does not lead to significant improvements in the translation quality over feeding the summary as a block of text. While the fleshed-out prompt did lead to the best results for English, the improvement compared to the inclusion of the unaligned summary is not high enough to be significant. Furthermore, the same prompt did not lead to increased, but to clearly decreased translation quality in German, as the translation with aligned summary gets basically the same scores as translation without any summary at all.

5 Conclusion

Incorporating human-made summaries into LLM-based translation tasks significantly enhances translation quality, when the summary language and the target language are equal. This is evidenced by the improved BLEU and ChrF scores of 2+ points when summaries are included in the prompt. Splitting the summaries into sentences and aligning them with the sentences in the letter does not lead to significant improvements and is highly dependent on the prompt and the language. These findings underscore the usefulness of language-specific summaries in improving LLM performance for the translation of historical texts.

This study invites many avenues for further investigation. A baseline experiment could be to regard the summaries as translations and to measure their BLEU scores. For short summaries that are less than half the length of the letter texts, this will inevitably lead to low scores. But for the longer summaries, this might give an interesting lower bound.

Another experiment to investigate the impact of the summary would involve the use of some arbitrary text instead of the summary. This will help us understand the impact of the summary in the automatic translation.

In future work we will also test whether the addition of summaries helps in translating from Early New High German to modern languages, as a follow-up of the work in (Volk et al., 2024b).

Another option is the combination of two LLMs, one that produces a summary (or a draft translation) for the letter in the target language, and another LLM that uses the summary in combination with the letter text for the translation.

Acknowledgments

We would like to thank the two anonymous reviewers for insightful comments. This research is part of the project "Bullinger Digital" funded by the UZH Foundation.

References

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 890–896, Singapore. Association for Computational Linguistics.

Lukas Fischer, Patricia Scheurer, Raphael Schwitter, and Martin Volk. 2022. Machine translation of 16th century letters from Latin to German. In *Proceedings of 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) at LREC-2022*, pages 43–50, Marseille.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arxiv.org/abs/2302.07856*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Martin Volk, Dominic P. Fischer, Lukas Fischer, Patricia Scheurer, and Phillip B. Ströbel. 2024a. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language*

Technologies for Historical and Ancient Languages – LT4HALA (at LREC/COLING), Torino.

Martin Volk, Dominic P. Fischer, Patricia Scheurer, Raphael Schwitter, and Phillip B. Ströbel. 2024b. LLM-based translation across 500 years. The case for Early New High German. In *Konferenz zur Verarbeitung natürlicher Sprache 2024 (KONVENS)*, Wien.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking LLM-based machine translation on cultural awareness. *arXiv:2305.14328v2*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A case study. In *Proceedings of the 40 th International Conference on Machine Learning*, Honolulu.

Probing BERT for German Compound Semantics

Filip Miletić **Aaron Schmid** **Sabine Schulte im Walde**
Institute for Natural Language Processing, University of Stuttgart, Germany
{filip.miletic, aaron.schmid, schulte}@ims.uni-stuttgart.de

Abstract

This paper investigates the extent to which pretrained German BERT encodes knowledge of noun compound semantics. We comprehensively vary combinations of target tokens, layers, and cased vs. uncased models, and evaluate them by predicting the compositionality of 868 gold standard compounds. Looking at representational patterns within the transformer architecture, we observe trends comparable to equivalent prior work on English, with compositionality information most easily recoverable in the early layers. However, our strongest results clearly lag behind those reported for English, suggesting an inherently more difficult task in German. This may be due to the higher productivity of compounding in German than in English and the associated increase in constituent-level ambiguity, including in our target compound set.

1 Introduction

Noun compounds – such as *music festival* and *ivory tower* in English; *Obstsaft* 'fruit juice' and *Sündenbock* 'scapegoat' (lit. 'sin buck') in German – comprise a productive class of expressions characterized by variable degrees of compositionality, i.e., relatedness of the individual constituents to the overall meaning of the compound. The ubiquitousness of noun compounds has motivated a long line of research modeling different aspects of their meanings (Ó Séaghdha, 2007; Mitchell and Lapata, 2008; Reddy et al., 2011; Schulte im Walde et al., 2016; Cordeiro et al., 2019, i.a.), while more recent work has specifically drawn on their semantically challenging nature to examine the linguistic knowledge encoded in transformer-based language models.

For instance, Garcia et al. (2021a) question BERT's ability to represent compositionality similarly to humans based on comparisons of compounds in context vs. in isolation. On a more specific level, Garcia et al. (2021b) find a lower quality of BERT representations for non-compositional compounds. Focusing on semantic relations of noun compounds, Rambelli et al. (2024) highlight strong performance variability across large language models as well as difficulties in generalizing to novel compounds, also noted in other related work (Li et al., 2022; Coil and Shwartz, 2023).

Further studies have attempted to explain these patterns by zooming into the model architecture, but without always reaching a consensus. As an example, Miletić and Schulte im Walde (2023) predict the compositionality of open (space-separated) English noun compounds, achieving the best results with embeddings from early transformer layers. Buijtelaar and Pezzelle (2023) similarly predict the semantic transparency of closed (orthographically joined) English noun compounds, but their best results use embeddings from later layers. Contradictory findings such as these still preclude broader generalizations; they are compounded by a limited understanding of cross-lingual trends given a near-exclusive focus on English in prior work (Miletić and Schulte im Walde, 2024).

Moving beyond that focus, we probe BERT via compositionality prediction of 868 German noun-noun compounds (Schulte im Walde et al., 2016). We replicate our setup from Miletić and Schulte im Walde (2023) for strict comparability with our prior results on English, but we introduce a scenario which is more challenging in several key respects: the closed spelling of German compounds limits constituent-level information in pretraining; the higher productivity of compounding in German (Berg et al., 2012) entails more diverse usage contexts and thereby may hinder learning; and as a result of the higher productivity, the ambiguity of individual constituents may also increase – including within constituent family sets (i.e., compounds which share one constituent) included in our gold standard data – and further challenge the models.

We provide a two-fold contribution. (i) Comparing all configurations, we broadly find that **representational patterns generalize cross-lingually**, in particular the relevance of constituent–context comparisons and the recoverability of compositionality information in early transformer layers. (ii) Looking at the best configurations, we find that **BERT's performance on German clearly lags behind English**, which may indicate an inherently more challenging task in German. On a more specific level, this trend may reflect the higher productivity of compounding in German and the related distinctiveness of gold standard information in the two languages. More generally, our study extends a very limited body of prior work (Falk et al., 2021; Jenkins et al., 2023) on German multiword expressions in transformer models.

2 Data

Gold standard compounds. We rely on the G_hoSt-NN dataset of 868 German noun-noun compounds annotated for compositionality, i.e., meaning contributions of the constituents to the overall compound meaning (Schulte im Walde et al., 2016). The targets in the dataset were selected starting from a seed set of 45 compounds balanced for modifier productivity and head ambiguity, and then adding further compounds which contain a modifier or a head already present in the seed set. By design, the dataset therefore includes constituent family sets, i.e., groups of compounds sharing a constituent. For example, it contains 15 compounds with the head *Kette* 'chain', such as *Bergkette* 'mountain chain', *Hotelkette* 'hotel chain', and *Halskette* 'necklace' (lit. 'neck chain'). Overall, the dataset contains 550 unique modifiers, of which 129 appear more than once; and 279 unique heads, of which 70 appear more than once.

For a given compound–constituent pair, expert annotators were asked to provide a rating from 1 (definitely semantically opaque) to 6 (definitely semantically transparent). The averaged final ratings subsume between 5 and 13 individual judgments. Sample items are shown in Table 1.

Corpus. We use the well-established DECOW corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015) with ≈ 11.6 billion tokens of web-crawled text. For each compound from the gold standard, we extract all occurrences from the corpus. In preprocessing, we deterministically split the compound into its constituents by replacing it with the modifier and head provided in the gold standard. This is done to constrain the output of the pretrained tokenizer used by the BERT models we deploy: it could otherwise split target compounds into subword fragments which are not morphologically motivated (cf. Jenkins et al., 2023), which would preclude us from analyzing the model's ability to represent the actual constituents.

Compound	Modif.	Head	M	H
Erbsensuppe *pea soup*	*Erbse* *pea*	*Suppe* *soup*	5.3	5.3
Kirchspiel *parish*	*Kirche* *church*	*Spiel* *game*	4.4	3.1
Eifersucht *jealousy*	*Eifer* *zeal*	*Sucht* *addiction*	2.0	2.1

Table 1: Sample compounds and compositionality ratings for the modifier (M) and the head (H).

3 Experimental Setup

We assess the compositionality information encoded in pretrained BERT via the task of unsupervised compositionality prediction. We follow the well-established framing of this problem as a ranking task, where a model's ability to represent compound semantics is evaluated by predicting the degrees of compositionality for a set of compounds and correlating those predictions with gold-standard compositionality ratings. Replicating the experimental setup we introduced in Miletić and Schulte im Walde (2023) for English, we experiment with a wide range of BERT-derived compositionality estimates. We evaluate each experimental configuration by calculating Spearman's rank correlation coefficient between the predicted degrees of compositionality (based on the cosine score, see below) and the gold-standard compositionality ratings for both modifiers and heads.

BERT models. We use the base German BERT model released by DBMDZ[1] (12 layers, 768 dimensions). We expand the English setup by comparing the cased and uncased versions of the model given the strong relevance of capitalization for German (nouns are systematically capitalized). We do not fine-tune the model since our primary aim is to assess the linguistic knowledge it inherently encodes rather than optimize it on the target task.

[1] https://huggingface.co/dbmdz/bert-base-german-cased

For a given compound, we feed each corpus example into the model individually. For each token in the sentence, this yields an embedding corresponding to each layer in the model architecture; we retain all these embeddings. We then estimate compositionality by comparing pairs of target embeddings in different ways.

Target embeddings. We use the following target embeddings: `modif`, corresponding to the modifier token; `head`, corresponding to the head token; `comp`, the average of `modif` and `head`; `cont`, corresponding to the sentence context, i.e., the average of all tokens except for `modif`, `head`, `[CLS]` and `[SEP]`; `cls`, corresponding to the `[CLS]` token which we assume to capture the meaning of the whole sentence. If the modifier or the head token is split into subwords by BERT's tokenizer, we average over those subwords.

Layers. We investigate all available layers, i.e., the input embedding layer and 12 hidden state outputs. We experiment with all spans of adjacent layers, ranging from a single layer in isolation to the full range of 13 layers, for a total of 91 unique combinations. When combining embeddings from multiple layers, we average over them.

Compositionality estimates. We predict compositionality in two ways. (i) Direct estimates correspond to the cosine score for a pair of target embeddings (e.g., `modif` and `comp`) from a given layer span. We test all pairs of target embeddings. (ii) Composite estimates use previously proposed composition functions (Reddy et al., 2011) to combine `head` and `modif` predictions obtained with one of the three other target embeddings: `comp`, `cont`, and `cls`. For example, starting from the cosines for (`modif`, `comp`) and (`head`, `comp`), we calculate `ADD` as the sum of the two; `MULT` as the product of the two; and `COMB` as the sum of `ADD` and `MULT`.

Other settings. In order to constrain the experimental space, we only vary the parameters discussed thus far, which we previously found to have a strong effect on model performance in English. We fix the remaining parameters from our setup in Miletić and Schulte im Walde (2023): as pooling function, we use averaging over vectors; we model 100 sentences per compound without controlling for sentence length; and we use token-level estimates, i.e., we compute compositionality estimates for each sentence individually and then average those estimates to obtain a compound-level value.

	Model	Layer	Emb.	ρ
Modif.	uncased	4–4	mod, cont	**0.332**
	uncased	3–4	mod, cont	0.319
	uncased	3–5	mod, cont	0.317
	uncased	4–5	mod, cont	0.313
	uncased	3–3	mod, cont	0.309
Head	cased	1–1	head, cont	**0.433**
	cased	1–2	head, cont	0.411
	cased	0–3	head, cont	0.402
	cased	1–3	head, cont	0.397
	cased	0–2	head, cont	0.393

Table 2: Best-performing experimental configurations for modifier and head compositionality predictions.

Prior approach	Modif.	Head
Schulte im Walde et al. (2016) LMI vectors; same data	0.490	0.590
Miletić and Schulte im Walde (2023) same method; English data	0.553	0.645

Table 3: Best results reported in prior work.

4 Results

4.1 Best parameter constellations

We begin by identifying the best-performing constellations of experimental parameters (Table 2). Our strongest results are weak-to-moderate correlations with gold standard compositionality ratings: $\rho = 0.332$ for modifiers and 0.433 for heads. But the full set of experimental configurations covers a very broad performance range, reaching negative correlations in the weakest cases ($\rho = -0.159$ for modifiers and -0.234 for heads), which confirms that compositionality information is not equally accessible across the BERT architecture. Furthermore, modifier and head predictions are only weakly correlated with one another ($\rho = 0.334$), i.e., the two constituents' respective contributions to the compound meaning are best captured by rather different representational information.

Looking at prior work (Table 3), the higher performance for head than modifier predictions aligns with previously reported trends. However, our highest results are around ≈ 0.2 ρ behind the count-based cooccurrence approach deployed by Schulte im Walde et al. (2016) on the same German dataset. We also observed a comparable lag of BERT behind simpler vector space approaches for English (Miletić and Schulte im Walde, 2023) with a setup that we replicate here. But our performance on

		mod	head	comp	cont	cls
Modif.	mod		**0.170**	**0.174**	**0.332**	**0.266**
	head	0.170		0.130	0.019	0.024
	comp	0.174	0.130		0.154	0.113
	cont	**0.332**	0.019	0.154		0.123
	cls	0.266	0.024	0.113	0.123	
Head	mod		0.327	0.202	0.178	0.084
	head	**0.327**		0.290	**0.433**	**0.246**
	comp	0.202	0.290		0.318	0.149
	cont	0.178	**0.433**	**0.318**		0.096
	cls	0.084	0.246	0.149	0.096	

Table 4: Best individual results obtained using direct comparisons of pairs of embeddings for modifier predictions (top) and head predictions (bottom). Bold values are best in a column; shaded values are best overall.

German also lags behind our prior results for English despite a strictly comparable experiment. As suggested above, this trend is consistent with an inherently higher difficulty of compositionality prediction in German. Its more challenging nature could be more specifically due to the higher productivity of compounding in German than in English, which may exacerbate constituent-level ambiguity, including within the G$_h$oSt-NN dataset given its reliance on constituent family sets.

As for the effect of individual experimental parameters, Table 2 indicates differences between modifier and head predictions regarding the strongest models (uncased vs. cased, respectively) and layers (mid-range vs. early layers, respectively). In both cases, the use of embeddings corresponding to the target structure (modifier and head, respectively) in combination with the embedding of the context yields the highest results. Taking a closer look at the interdependency of modifier/head representations and the corresponding predictions, we additionally break down the results across all pairs of target embeddings (Table 4). This further confirms the central importance of representational information corresponding to the constituent of interest, closely reflecting prior findings for English (Miletić and Schulte im Walde, 2023).

4.2 Cased vs. uncased models

Regarding differences between BERT models, modifier predictions are better under experimental configurations using the uncased version (median $\rho = 0.060$ vs. 0.073); by contrast, head predictions benefit from the cased version (median $\rho = 0.201$ vs. 0.165). Looking at the predictions obtained with the cased and uncased model across *all* experimental settings, we find that they are themselves strongly correlated with one another, for modifiers ($\rho = 0.768$) as well as heads ($\rho = 0.901$). In other words, the patterns captured by the two model versions are affected by the underlying properties of representational information (embeddings and layers) in a similar – but not identical – way.

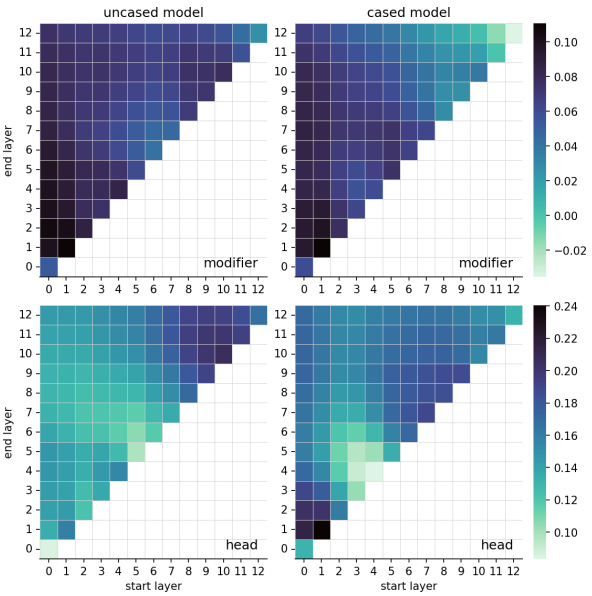


Figure 1: Mean performance across contiguous spans of layers, defined by the start layer (x-axis) and end layer (y-axis). Left: uncased model; right: cased model. Top: modifier predictions; bottom: head predictions.

To further understand these interactions, we compute correlations between predictions obtained by the cased vs. uncased model in *subsets* of experimental settings. (i) We first do this for each of 91 layer combinations. By keeping the layers fixed, we assess model sensitivity to compositionality estimates. We find strong mean correlations for modifiers (0.736 ± 0.111) and heads (0.915 ± 0.055). (ii) We then compute the correlations for each of 19 compositionality estimates. By keeping the estimates fixed, we assess model sensitivity to layer combinations. We find moderate mean correlations for modifiers (0.432 ± 0.264) and heads (0.494 ± 0.213). These results indicate that compositionality information captured by the uncased vs. cased model is rather similar across compositionality estimates; and rather different across layers.

4.3 Layers

We now examine mean prediction performance for different layer spans to gain further insight into the transformer architecture (Figure 1). For modifier predictions, the best results are on average obtained

in the lower range of layers, with the single highest mean result on layer 1 in isolation ($\rho = 0.110$). Similar performance is obtained by other layer spans – including very broad ones – which start from the earliest layers. By contrast, later layers yield clearly lower results, including layers that are often used in lexical semantic tasks (e.g., layers 9–12, $\rho = 0.036$; and layer 12 in isolation, which obtains the single lowest mean $\rho = -0.004$).

Head predictions exhibit comparatively more variance and rather different trends. Like for modifiers, the single best mean results is on layer 1 in isolation ($\rho = 0.199$). However, the next range of performance is occupied by spans limited to the very early layers (0–2) and the later layers (7–12), and quite distinctly *not* very broad spans starting from the earliest layers. The lowest mean result is obtained by input layer 0 in isolation ($\rho = 0.108$).

These findings broadly align with good performance of lower layers we reported for English in Miletić and Schulte im Walde (2023), but we find stronger differences between head and modifier predictions. Some of these interact with the choice of uncased vs. cased model; we report the differences between the two models for each layer span in Appendix A, and summarize the trends below. The uncased model obtains better performance (by ≈ 0.05 ρ) in the early-to-mid range (layers 3–5, especially for heads) and in the later range (layers 9–12, especially for modifiers). Put differently, it benefits from stronger contextualization (i.e., processing in more layers), whose disambiguating effect may be relevant given the loss of information inherent in case folding. The cased model yields gains especially for head predictions, and most clearly in the very early layers (up to 0.08 ρ). We hypothesize that capturing the nominal nature of a constituent – reflected by capitalization in German, which is preserved by the cased model – is more important for heads given their dominant role in the morphosyntactic constituency of compounds.

5 Conclusion

We investigated the extent to which pretrained German BERT encodes the knowledge of noun compound semantics. We systematically varied representational information (across target tokens, layers, and cased vs. uncased models) to predict the degrees of compositionality of 868 noun-noun compounds. Our best result ($\rho = 0.433$) lags behind equivalent prior work on English – suggesting a more challenging nature of the task in German – but we also confirm previously reported patterns of model processing such as the importance of early layers. Our insights more generally illustrate the key importance of cross-lingual extensions of probing studies to languages other than English.

Limitations

We note several limitations of our work. (i) Our study provides a direct cross-lingual comparison with prior results obtained on English, but it is limited to only one other language – German – which also belongs the Germanic family and exhibits relatively similar patterns of multiword expression formation. Typologically more distant languages with stronger structural differences (e.g., Romance languages with a preference for N–Prep–N rather than N–N structures) could provide a further cross-lingual validation of the reported patterns. (ii) We only consider noun compounds, but other categories of multiword expressions (e.g., particle verbs) may exhibit different processing patterns in the transformer architecture. (iii) We compare the cased and uncased versions of a single German BERT model. Other variations such as German models using different pretraining data or parameter sizes, as well as comparisons with multilingual models, could provide further insights. (iv) We compare BERT performance on English and German based on a strictly comparable experimental setup for both languages. However, we use language-specific (and therefore different) gold standard compositionality ratings. While the datasets on which we rely are well-established for each language and define the annotation task in a comparable way, they follow different strategies of selecting target items, which may affect some of the reported trends. For a recent discussion of such effects, see Schulte im Walde (2024).

Acknowledgments

The research presented here was supported by DFG Research Grant SCHU 2580/5-1 (*Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*).

References

Thomas Berg, Sabine Helmer, Marion Neubauer, and Arne Lohmann. 2012. Determinants of the extent of compound use: A contrastive analysis. *Linguistics*, 50(2).

Lars Buijtelaar and Sandro Pezzelle. 2023. A psycholinguistic analysis of BERT's representations of compounds. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2230–2241, Dubrovnik, Croatia. Association for Computational Linguistics.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Neele Falk, Yana Strakatova, Eva Huber, and Erhard Hinrichs. 2021. Automatic classification of attributes in German adjective-noun phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 239–249, Groningen, The Netherlands (online). Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2023. To split or not to split: Composing compounds in contextual vector spaces. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16131–16136, Singapore. Association for Computational Linguistics.

Siyan Li, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's interpretation of novel English noun compounds. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 717–728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2007. Annotating and learning compound noun semantics. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 73–78, Prague, Czech Republic. Association for Computational Linguistics.

Giulia Rambelli, Emmanuele Chersoni, Claudia Collacciani, and Marianna Bolognesi. 2024. Can large language models interpret noun-noun compounds? a linguistically-motivated study on lexicalized and novel compounds. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11823–11835, Bangkok, Thailand. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 486–493, Istanbul, Turkey. European Language Resources Association.

Sabine Schulte im Walde. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword Expressions in Lexical Resources. Linguistic, Lexicographic and Computational Perspectives*, Phraseology and Multiword Expressions. Language Science Press, Berlin.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective.

In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portorož, Slovenia. European Language Resources Association (ELRA).

A Layer performance

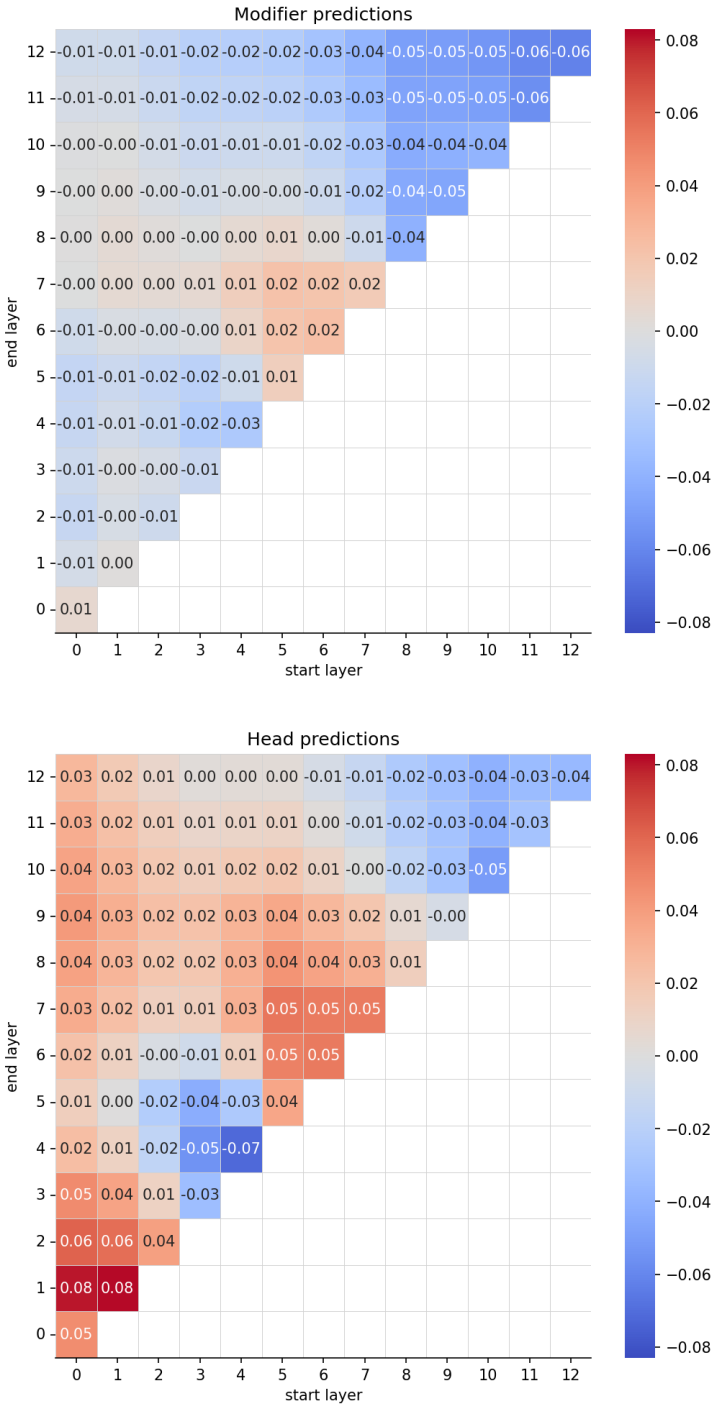


Figure 2: Layer-wise difference in cased vs. uncased model performance. **Positive values:** better performance of the cased model. **Negative values:** better performance of the uncased model.

SLANet-1M: A Lightweight and Efficient Model for Table Recognition with Minimal Computational Cost

Nguinwa Mbakop Dimitri Romaric
University of Florence
iCoSys - HEIA-FR, HES-SO
Fribourg, Switzerland
nguinwa.dimitri@edu.unifi.it

Andrea Petrucci
iCoSys - HEIA-FR, HES-SO
Fribourg, Switzerland
andrea.petrucci@hefr.ch

Simone Marinai
DINFO
University of Florence
Florence, Italy
simone.marinai@unifi.it

Jean Hennebert
iCoSys - HEIA-FR, HES-SO
Fribourg, Switzerland
jean.hennebert@hefr.ch

Abstract

Modern approaches for table recognition consist of an encoder for feature extraction and one or more decoders for structure recognition and cell box detection. Recent advancements in this field have introduced Transformers, initially in the decoders and more recently in the encoder as well. While these improvements have enhanced performance, they have also increased model complexity, requiring larger datasets for training, a pre-training step, and higher inference time.

In this paper, we explore SLANet, a lightweight transformer-free model originally trained on PubTabNet. To train a more robust version, we combined two publicly available datasets (PubTabNet and SynthTabNet) into one dataset of 1 million of images table, which led us to name the resulting model **SLANet-1M**. On PubTabNet, SLANet-1M improves the original SLANet's S-TEDS score by **0.35%**. It also scores only **0.53%** below the state-of-the-art UniTable Large, while using nearly **14 times fewer parameters**. SLANet*—a variant trained on PubTabNet and a quarter of SynthTabNet— achieves a 0.47% improvement. On SynthTabNet, SLANet-1M performs exceptionally well, with an S-TEDS score just **0.03%** lower than UniTable Large. Additionally, SLANet-1M outperforms major large vision-language models (VLMs) like GPT-4o, Granite Vision, and Llama Vision on this specific table recognition task. **SLANet-1M is also more efficient during inference**, offering faster processing and CPU-friendly execution, eliminating the need for a GPU.

1 Introduction

Tables contain a wealth of information in a concise format and are prevalent in documents. Extracting table information accurately is crucial for many applications (data analysis, finance, health, and so on). The table recognition task focuses on detecting tables in image-based documents and extracting their structure and contents in HTML format. However, due to the complexity of tables—such as rowspan, colspan, and multi-header layouts—table recognition remains a challenging task, even for advanced large vision-language models (VLMs) like GPT-4o (OpenAI, 2024), GPT-4-turbo (Yang et al., 2023), Granite Vision 3.2 (Team et al., 2025), and Llama Vision 3.2 (AI, 2025).

This paper presents a solution for companies that require high-performance table recognition without extensive computational resources. We enhanced SLANet (Li et al., 2022) quantitatively and qualitatively by training it on additional data, demonstrating that a lightweight model without Transformers can achieve performance comparable to more complex transformer-based models. Furthermore, we show that our improved SLANet is faster than state-of-the-art (SOTA) models while maintaining high accuracy.

We name this enhanced model **SLANet-1M** as it is trained on 1 million images by combining PubTabNet[1] and SynthTabNet[2] datasets.

2 Related works

Many recent models on table recognition task have demonstrated great performance. Here we explore some of them, in particular models that follow the encoder-decoder architecture. We show that with the introduction of Transformers, their structure has adopted this technology firstly in the decoders and subsequently in the encoder as well.

2.1 EDD

The Encoder-Dual-Decoder (EDD) model was introduced in the PubTabNet paper (Zhong et al., 2020). EDD consists of an encoder, an attention-based structure decoder, and an attention-based cell decoder. The use of two decoders stems from the

[1] https://github.com/ibm-aur-nlp/PubTabNet
[2] https://github.com/IBM/SynthTabNet

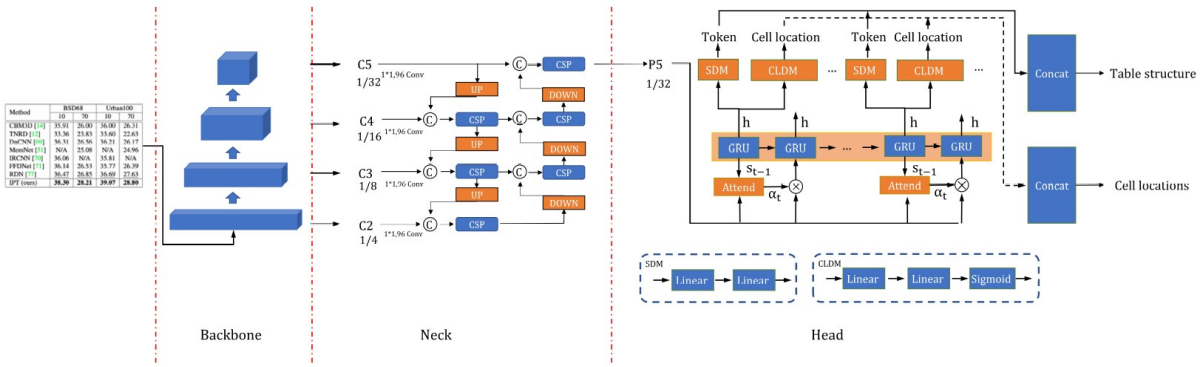


Figure 1: Architecture of SLANet.

observation that table structure recognition and cell content recognition are distinct tasks that are inefficient to solve with a single attention-based decoder.

EDD's encoder is a convolutional neural network (CNN) that captures visual features from input table images. The structure decoder and the cell decoder are recurrent neural networks (RNNs) equipped with an attention mechanism to process and reconstruct the structure and content of the table.

2.2 Table Master

Table Master (Ye et al., 2021) was introduced as a solution for the ICDAR 2021 competition on scientific literature parsing (Task B: table recognition to HTML). Inspired by MASTER (Lu et al., 2021), its decoder is composed of Transformer decoder layers.

Table Master employs two decoder branches, each consisting of three Transformer decoder layers, with the first layer shared between both branches. One branch is responsible for predicting the HTML sequence, while the other conducts box regression. Unlike other models that split tasks at the final layer, Table Master decouples sequence prediction and box regression immediately after the first Transformer decoder layer.

2.3 TableFormer

Introduced in the SynthTabNet paper (Nassar et al., 2022), TableFormer employs an hybrid CNN-Transformer architecture as encoder. The encoder consists of a ResNet-18 CNN and a Transformer encoder with two encoder layers, extracting features from input images into a fixed-length feature vector. TableFormer has two decoders: a structure decoder, modeled as a Transformer decoder with four decoder layers, incorporating multi-head attention and feed-forward networks (FFNs), and a cell box decoder, which utilizes the same Transformer encoder and decoder but introduces an additional attention-based FFN block to refine cell-level predictions.

2.4 VAST

The Visual-Alignment Sequential Coordinate Table Recognizer (VAST) (Huang et al., 2023) consists of three primary components: a modified ResNet enhanced with multi-aspect global content attention as the CNN-based image encoder, a transformer-based HTML sequence decoder, and a Transformer block for coordinate sequence decoding, allowing precise localization of table structures.

2.5 UniTable

UniTable (Peng et al., 2024) is the most recent model in table recognition, introducing a transformer-based encoder alongside a Transformer decoder. Initially, in an earlier attempt (Huang et al., 2023), replacing the CNN encoder with a vanilla Transformer with linear projection led to a performance drop compared to models using CNN or hybrid CNN-Transformer encoders.

To address this issue, UniTable implements self-supervised pre-training for the visual encoder.

2.6 SLANet

SLANet stands for Structure Location Alignment Network, presented in PP-StructureV2 (Li et al., 2022) as an efficient Table Recognition algorithm. In Figure 1 we show the network architecture of the model, composed of a backbone, a neck, and a head. we provide a detailed description of the architecture in Section 4.2.

3 Contribution

Our main contribution lies in adapting and evaluating the SLANet model (Li et al., 2022) on an additional dataset to assess its generalization capabilities and performance relative to state-of-the-art (SOTA) methods. Detailed information on the implementation and training procedure is provided in Sections 6.1.1 and 6.1.2.

In addition, we extend prior work by evaluating and comparing the **inference time on CPU** of some of the models discussed in the previous section—an aspect that has not been systematically analyzed in their original studies.

4 Formulation and SLANet's details

In this section, we define the table structure recognition task and provide a detailed description of the model we adopt for our experiments. We also outline the loss functions used during training.

4.1 Task Definition

The objective of **Table Recognition (TR)** is to convert a tabular image I into a structured, machine-readable format T, capturing both its *logical* and *physical* structure. The logical structure is often represented in HTML format, denoted as a tokenized sequence $S = [s_1, ..., s_T]$, where each s corresponds to an HTML tag. The physical structure consists of the bounding box coordinates of non-empty cells, represented as $B = [b_1, ..., b_N]$, where each bounding box is defined as $b = (x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$, with integer values. Additionally, $C = [c_1, ..., c_N]$ represents the textual content inside each cell, following a reading order. While the number of elements in B and C are the same, they are typically fewer than those in S ($N < T$), since the HTML sequence includes both filled and empty cells. Each cell is associated with a single bounding box and may contain either a single line or multiple lines of text.

4.2 SLANet's Architecture

4.2.1 Backbone

SLANet employs PP-LCNet (Cui et al., 2021) as its backbone, a lightweight, CPU-friendly convolutional neural network architecture. PP-LCNet introduced several novel ideas to improve the accuracy without increasing the inference time. These techniques can be summarize as follows:

- Better activation function; from ReLU to H-Swish.

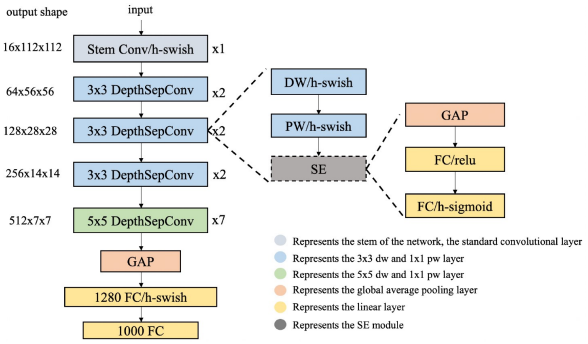


Figure 2: PP-LCNet. PP-LCNet includes optional modules, indicated by the dotted box. The stem section utilizes a standard 3×3 convolution. DepthSepConv refers to depth-wise separable convolutions, where DW stands for depth-wise convolution, PW denotes point-wise convolution, and GAP represents Global Average Pooling.

- SE (squeeze-and-excitation) modules (Hu et al., 2018) at appropriate positions.
- Larger convolution kernels; replacing the 3x3 convolutional kernels with the 5x5 convolutional kernels only at the tail of the network.
- Larger dimensional 1x1 conv layer after GAP; to give the network a stronger fitting ability and allow for more storage of the model with little increase of inference time. PP-LCNet appended a 1280-dimensional size 1x1 conv (equivalent to FC layer) after the final GAP layer.

PP-LCNet uses DepthSepConv (Howard et al., 2017) as its basic block, the architecture is shown in Figure 2. Depthwise Separable Convolution is a good alternative to the classic convolution, as it can reduce the complexity and improve the inference speed of the operation while maintaining the accuracy. With all these improvements, PP-LCNet achieves better performance on multiple tasks with respect to lightweight models such as ShuffleNetV2 (Ma et al., 2018), MobileNetV3 (Howard et al., 2019), and GhostNet (Han et al., 2020).

4.2.2 Neck

SLANet enhances feature fusion to effectively address challenges caused by scale variations in complex scenes. To achieve this efficiently, it utilizes CSP-PAN (Yu et al., 2021), which integrates the PAN (Path Aggregation Network) structure for multi-level feature extraction and the CSP (Cross Stage Partial) structure for feature concatenation and fusion between adjacent feature maps.

Path Aggregation Network (PAN) (Liu et al., 2018) improves the feature pyramid by enhancing localization accuracy and optimizing information flow. It introduces:

- Bottom-up path augmentation, which shortens the information path and strengthens low-level features with precise localization signals.
- Adaptive feature pooling, which aggregates features across all levels for each proposal, ensuring a more structured and efficient feature propagation while avoiding arbitrary assignments.

These enhancements create more efficient and structured feature pathways, improving feature fusion and ultimately boosting detection performance.

Cross Stage Partial (CSP) Structure (Wang et al., 2020) is designed to enhance gradient flow while reducing computational cost. It achieves this by splitting the base layer's feature map into two parts and merging them through a cross-stage hierarchy. By dividing the gradient flow into separate network paths, CSP ensures that the propagated gradient information exhibits a greater correlation difference, improving learning efficiency through alternating concatenation and transition steps.

To optimize efficiency further, SLANet reduces the output channels of CSP-PAN from 128 to 96, effectively decreasing the model size without compromising performance.

4.2.3 Head

In its head module, SLANet employs a GRU along with two key components: the **Structure Decode Module** (SDM) and the **Cell Location Decode Module** (CLDM). The result of the feature fusion is passed in the GRU, and at each step, the GRU's output is concatenated and passed to both SDM and CLDM, generating cell tokens and their corresponding bounding box coordinates.

SLANet ensures one-to-one alignment between cell tokens and their coordinates, with SLAHead responsible for maintaining this correspondence. The tokens and coordinates from all decoding steps are concatenated to construct the HTML table representation along with the precise coordinates of all cells.

Inspired by TableMaster (Ye et al., 2021), SLANet treats <td> and </td> as a single token (<td></td>), simplifying the tokenization process for table structure generation.

	Weaning	Week 15	Off-test
Weaning	–	–	–
Week 15	–	0.17 ± 0.08	0.16 ± 0.03
Off-test	–	0.80 ± 0.24	0.19 ± 0.09

Figure 3: An example image from PubTabNet.

4.3 Loss Functions

The total loss function consists of two components: *structure loss* and *localization loss*, combined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{structure}}\mathcal{L}_{\text{structure}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}}$$

This combined loss ensures the model effectively learns both table structure and bounding box localization.

4.3.1 Structure Loss

The structure loss measures the accuracy of table structure predictions using the cross-entropy loss:

$$\mathcal{L}_{\text{structure}} = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{T} y_{i,j}\log(\hat{y}_{i,j})$$

where K is the batch size, T is the sequence length, $y_{i,j}$ is the ground truth token, and $\hat{y}_{i,j}$ is the predicted probability.

4.3.2 Localization Loss

The localization loss evaluates bounding box accuracy using the *SmoothL1* loss:

$$\textit{SmoothL1}(x) = \begin{cases} 0.5\,x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

where $x = \mathbf{b}_{i,j} - \hat{\mathbf{b}}_{i,j}$.
The localization loss is normalized as:

$$\mathcal{L}_{\text{loc}} = \frac{\sum_{i,j}\textit{SmoothL1}(\mathbf{b}_{i,j} - \hat{\mathbf{b}}_{i,j}) \cdot m_{i,j}}{\sum_{i,j} m_{i,j} + \epsilon}$$

where $\epsilon > 0$ prevents division by zero, $\mathbf{b}_{i,j} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ is the ground truth bounding box, $\hat{\mathbf{b}}_{i,j}$ is the predicted bounding box, and $m_{i,j}$ is a mask for valid bounding boxes.

5 Datasets and Metrics

5.1 Datasets

In this paper, we explore two publicly available table structure recognition benchmark datasets: PubTabNet and SynthTabNet.

5.1.1 PubTabNet

The PubTabNet (Zhong et al., 2020) dataset consists of 509,892 annotated PNG images (500,777 for training and 9,115 for validation). Each table is annotated with its structure in HTML format, along with tokenized text and bounding boxes for each cell. As shown in Figure 3, the dataset primarily contains simpler table structures with relatively few rows and columns. Additionally, the dataset exhibits limited variation in table styles, which hinders model generalization to unseen table formats. Recognizing these limitations, the authors of TableFormer introduced SynthTabNet to address these issues.

5.1.2 SynthTabNet

SynthTabNet (Nassar et al., 2022) is a large-scale synthetically generated dataset designed to offer control over dataset size, table structures, table styles, and content types.

The dataset aims to overcome the shortcomings of PubTabNet and FinTabNet, which suffer from skewed distributions toward simpler tables, limited stylistic diversity, and restricted cell content types. SynthTabNet consists of 600,000 tables, divided into four 150,000-table subsets:

Finance (1) and *PubTabNet* (3), which mimic FinTabNet[3] and PubTabNet while incorporating more complex structures. *Marketing* (2), which features high-contrast, colorful tables that resemble real-world marketing documents as shown in Figure 4. *Sparse* (4), which contains tables with minimal content, testing model performance on incomplete or sparsely populated tables. All parts are divided into Train, Val, and Test splits (80%, 10%, 10%). Because SynthTabNet provides a comprehensive evaluation of table recognition models across diverse table structures, we use it for ablation studies and present results separately for each subset.

5.2 PubTables-1M

Although we did not use PubTables-1M (Smock et al., 2022) in our experiments, we include it here as it is one of the largest table recognition (TR) datasets. PubTables-1M comprises nearly one million tables extracted from scientific articles, supports multiple input modalities, and provides detailed header and location information for table

[3] https://developer.ibm.com/data/fintabnet/

Balance at beginning of period	Expected return on plan assets	
For the years ended December 31,	Item 7.	
Americas	Expired	Fair value
	Total consumer	Current liabilities
Fixed maturity securities	Repurchase of common stock	Gross
Years Ended December 31	Derivative instruments	Liabilities
	Total comprehensive income	
Total equity	Fixed rate	

Figure 4: An example image from SynthTabNet (Marketing subset).

structures. These features make it a valuable resource for various modeling approaches.

However, as noted by UniTable (Peng et al., 2024), PubTables-1M suffers from several inconsistencies, particularly in its annotation method. The dataset uses word-wise bounding box (bbox) annotations, whereas PubTabNet and SynthTabNet follows a cell-wise annotation approach.

- Cell-wise annotation assigns a single bbox per table cell, allowing for a direct mapping between non-empty cells and their corresponding HTML structure.
- Word-wise annotation, used in PubTables-1M, assigns a bbox to each individual word, making it challenging to integrate with the table structure as effectively as cell-wise annotation.

This fundamental difference limits the general applicability of PubTables-1M for certain table recognition tasks.

5.3 Metrics

5.3.1 Accuracy

Used during the training the accuracy refers to the proportion of correctly identified table elements (such as structure, cells, or text) compared to the total number of ground truth elements. It measures the effectiveness of a table recognition system in correctly detecting and extracting tables from documents.

It is defined as:

$$\text{Acc.} = \frac{\text{Numb. of Correctly Recognized Elements}}{\text{Total Numb. of Ground Truth Elements}}$$

5.3.2 TEDS

TEDS (Tree-edit-distance-based Similarity), introduced by PubTabNet (Zhong et al., 2020), converts the table into a tree structure in HTML format

and measures the edit distance between the prediction T_{pred} and the groundtruth T_{gt}. A shorter edit distance indicates a higher degree of similarity, leading to a higher TEDS score. TEDS measures both the table structure and table cell content. We also use S-TEDS as metric where only the table structure is considered. For comparison we consider more S-TEDS because for the content of cells some models rely on external text detection and text recognition models, which can differ from model to model and so can compromise the comparison.

TEDS between two trees is computed as:

$$TEDS = 1 - \frac{\mathit{EditDist}(T_{gt}, T_{pred})}{\max(|T_{gt}|, |T_{pred}|)} \quad (1)$$

where *EditDist* denotes tree-edit distance (Pawlik and Augsten, 2016), and $|T|$ is the number of nodes in T.

Datasets	**Records**		**Size (GB)**	
Name	**Train**	**Val**	**Train**	**Val**
PubTabNet	500,777	9,115	11.6	0.2
SynthTabNet	480,347	59,618	24.2	3.0
Merged	**981,124**	**68,733**	**35.8**	**3.2**

Table 1: Dataset details including records and sizes for training and validation for SLANet-1M.

Models	**Datasets**	
	PubTabNet	**SynthTabNet 3**
SLANet	76.35	17.21
SLANet*	**77.07**	**81.72**

Table 2: Results (accuracy) of the first experiment, SLANet is the original model trained on PubTabNet, and SLANet* is the model trained on both PubTabNet and SynthTabNet part 3.

Models	**Datasets**			
	PubTabNet		**SynthTabNet 3**	
	TEDS	**S-TEDS**	**TEDS**	**S-TEDS**
SLANet	**95.89**	97.01	89.01	95.65
SLANet*	95.83	**97.48**	**92.87**	**99.47**

Table 3: Results (TEDS and S-TEDS) of the first experiment, SLANet is the original model trained on PubTabNet, and SLANet* is the model trained on both PubTabNet and SynthTabNet part 3.

6 Experiments and Results

6.1 Experiments

6.1.1 Implementations

We conducted two setup experiments, both on a 48G A40 GPU device, during 50 epochs using Adam as optimizer, the initial learning rate is set to 0.001 and adjusted to 0.0001 and 0.00005 after 29 and 39 epochs. The batch size is set to 48 for the first experiment and to 72 for the second.

6.1.2 Training

For the first experiment we trained SLANet from scratch on the PubTabNet and the third part of SynThTabNet for a total of 620,772 images for the training set, validate on the validation set of PubTabNet and tested on the same set because there is no the groundthruth fot the test set of PubTabNet.

For the second experiment we merged both datasets (PubTabNet and SynthTabNet) as detailed in Table 1. The validation set is obtained by merging all the validation sets of subsets of SynthTabNet with the validation set of PubTabNet. The tests are made on the test sets of SynthTabNet subsets.

6.2 Results

6.2.1 First Experiment

The model obtained with the first training setup is named **SLANet*** and the Table 2 and Table 3 summarize the performance of SLANet and SLANet* across PubTabNet and SynthTabNet (Part 3). SLANet*, trained on both datasets, consistently outperforms the original SLANet. On PubTabNet, SLANet* achieves a slight **0.72%** improvement in accuracy while maintaining comparable TEDS performance and a **0.47%** increase in S-TEDS.

The performance boost is more pronounced on SynthTabNet (Part 3), where SLANet* significantly surpasses SLANet, improving accuracy from 17.21% to 81.72%. Additionally, it demonstrates a substantial increase in TEDS (**+3.86%**) and S-TEDS (**+3.82%**), confirming its enhanced adaptability when trained on a more diverse dataset.

Table 4 compares SLANet* to state-of-the-art models on PubTabNet. Despite having **significantly fewer parameters** —nearly 14 times fewer than the strongest models — SLANet* achieves competitive performance. It is only **0.41%** on S-TEDS score behind the SOTA UniTable Large, demonstrating its efficiency and effectiveness in the table recognition task.

Models	TEDS	S-TEDS	SIZE (M)
EDD (Zhong et al., 2020)	88.30	89.90	-
TableMaster (Ye et al., 2021)	96.12	97.56	253
TableFormer (Nassar et al., 2022)	93.60	96.75	53.2
VAST (Huang et al., 2023)	96.31	97.23	-
UniTable Base (Peng et al., 2024)	94.78	95.63	30
UniTable Large (Peng et al., 2024)	**96.50**	**97.89**	125
SLANet (Li et al., 2022)	95.89	97.01	**9.2**
SLANet* (ours)	95.83	97.48	**9.2**

Table 4: Comparison on PubTabNet of models based on TEDS, S-TEDS, and SIZE.

Models	S-TEDS	Size (M)
TableFormer	96.70	53.2
UniTable Base	98.97	30
UniTable Large	**99.39**	125
SLANet-1M	99.36	**9.2**

Table 5: Comparison of performance on SynthTabNet.

6.2.2 Second Experiment

In the second experiment, we trained SLANet on the consolidated dataset detailed in Table 1. The resulting model, referred to as SLANet-1M, demonstrates strong performance on the SynthTabNet benchmark, as illustrated in Table 5. In particular, SLANet-1M lags behind UniTable Large by a mere **0.03%**, despite possessing approximately 14 times fewer parameters. It is important to highlight that UniTable Large benefits from a significantly broader training regimen—having been trained on PubTabNet, SynthTabNet, and FinTabNet for table recognition, in addition to undergoing a pre-training phase on PubTabNet, SynthTabNet, FinTabNet, and PubTables-1M.

6.2.3 Ablation Study

Table 6 presents an ablation study comparing the S-TEDS scores of UniTable, SLANet, SLANet*, and SLANet-1M across the four subsets of the SynthTabNet dataset. As expected, SLANet-1M outperforms both SLANet and SLANet* on all the three other subsets, given that it was explicitly trained on these data partitions. Notably, SLANet-1M also demonstrates a modest improvement of **0.05%** on the PubTabNet subset of SynthTabNet.

When compared to UniTable Large, SLANet-1M achieves superior performance on the Marketing subset with a **0.14%** lead and matches UniTable's score on the Sparse subset. On the PubTabNet subset, it trails slightly by only **0.04%**. The most pronounced difference is observed on the Finance subset, where SLANet-1M falls behind UniTable Large by **0.23%**—this being the only subset where UniTable Base also surpasses SLANet-1M, albeit by a smaller margin of **0.06%**. This performance gap can likely be attributed to UniTable's broader training scope, as it was trained on a more diverse set of datasets, including FinTabNet, which may contribute to its enhanced generalization on financial tables.

7 Qualitative Results and Inference Time

7.1 Qualitative Results

In this section, we present a qualitative analysis by first comparing SLANet-1M with the original SLANet, followed by a comparison with several large vision-language models (VLMs). One representative sample per configuration was retained. Additional examples can be found in the appendix.

7.1.1 SLANet vs SLANet-1M

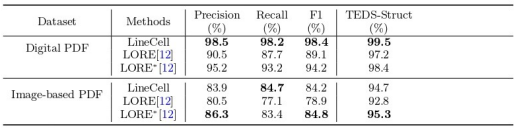
Figure 5 illustrates the inputs provided to both SLANet and SLANet-1M, along with the corresponding HTML tables generated by each model. As shown, SLANet encounters difficulties in accurately identifying and separating the correct number of rows. In contrast, SLANet-1M successfully overcomes this limitation, generating a well-structured HTML table that clearly delineates rows, even in cases where they are not explicitly wired in the input.

7.1.2 SLANet vs VLMs

Following the approach of UniTable (Peng et al., 2024), we conduct a qualitative comparison between our model and several state-of-the-art large vision-language models (VLMs). Figure 6 presents the input image alongside the outputs generated by SLANet-1M, GPT-4o (OpenAI, 2024), Granite

Models	Finance	Marketing	PubTabNet	Sparse
UniTable Base (Peng et al., 2024)	99.41	98.35	99.44	98.69
UniTable Large (Peng et al., 2024)	**99.58**	99.08	**99.56**	99.34
SLANet (Li et al., 2022)	89.83	80.83	95.65	86.10
SLANet* (ours)	91.26	82.99	99.47	91.33
SLANet-1M (ours)	99.35	**99.22**	99.52	**99.34**

Table 6: Comparison across different subsets of SynthTabNet dataset.



Dataset	Methods	Precision (%)	Recall (%)	F1 (%)	TEDS-Struct (%)
Digital PDF	LineCell	**98.5**	**98.2**	**98.4**	**99.5**
	LORE[12]	90.5	87.7	89.1	97.2
	LORE*[12]	95.2	93.2	94.2	98.4
Image-based PDF	LineCell	83.9	**84.7**	84.2	94.7
	LORE[12]	80.5	77.1	78.9	92.8
	LORE*[12]	**86.3**	83.4	**84.8**	**95.3**

(a) Input table image extracted from PdfTable (Sheng and Xu, 2024).

Dataset	Methods	Precision (%)	Recall (%)	F1 (%)	TEDS-Struct (%)
Digital PDF	LineCell	98.5	98.2	98.4	99.5
	LORE[12]	90.5	87.7	89.1	97.2
	LORE* [12]	95.2	93.2	94.2	98.4
Image-based PDF	LineCell	83.9	84.7	84.2	94.7
	LORE[12]	80.5	77.1	78.9	92.8
	LORE* [12]	86.3	83.4	84.8	95.3

(b) SLANet-1M's output.

Dataset	Methods	Precision (%)	Recall (%)	F1 (%)	TEDS-Struct (%)
Digital PDF	LineCell LORE[12] LORE*[12]	98.5 90.5 95.2	98.2 87.7 93.2	98.4 89.1 94.2	99.5 97.2 98.4
Image-based PDF	LineCell LORE[12] LORE*[12]	83.9 80.5 86.3	84.7 77.1 83.4	84.2 78.9 84.8	94.7 92.8 95.3

(c) SLANet's output.

Figure 5: Qualitative comparison between SLANet and SLANet-1M.

Vision 3.2 (Team et al., 2025), and Llama Vision 3.2 (AI, 2025).

We adopt the same prompt used in UniTable (Peng et al., 2024) and in the evaluation of the Optical Character Recognition (OCR) capabilities of GPT-4V (Shi et al., 2023): *"Please read the table in this image and return an HTML-style reconstructed table in text. Do not omit anything."*

The results show that SLANet-1M outperforms GPT-4o, which fails to preserve the correct number of rows and introduces unnecessary blank spaces and empty cells. In contrast, SLANet-1M more faithfully maintains the table's structural integrity. Among the baseline VLMs, Granite Vision 3.2 performs the best, although it misplaces the content of the first cell by rendering it in the last cell of the first row. Llama Vision 3.2 simplifies the output by reducing the table to just two columns, revealing its limitations in handling complex table structures.

One qualitative result is shown here; more quantitative and qualitative results are in Appendices A and B, respectively.

7.2 Inference Time

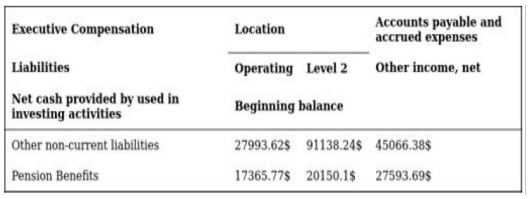
One of the main objectives of this research was to provide an alternative to transformer-based table recognition models—one that achieves similar performance while remaining efficient enough to run on a CPU with a satisfactory inference time.

All the models cited in this paper overlook this aspect. To address this, we compared the inference time of SLANet-1M (which is essentially the same as SLANet) against two state-of-the-art models: TableFormer and UniTable Large. The evaluation was conducted on a CPU-powered system with the following specifications:

- Processor: 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz, 2496 MHz, 8 Cores, 16 Logical Processors.
- Memory: 32.0 GB RAM.
- System Type: x64-based PC.
- Dataset: 200 images (50 images per subset).

The Docling technical report (Auer et al., 2024) highlights that TableFormer suffers from high inference time on CPU due to its reliance on EasyOCR[4], a finding that our experiments confirmed. Specifically, TableFormer exhibited an average inference time of 10,020 milliseconds, while UniTable Large was even slower, likely due to its fully transformer-based architecture, with an average inference time of 118,729 milliseconds. In contrast, SLANet-1M

[4] https://github.com/JaidedAI/EasyOCR



Executive Compensation	Location		Accounts payable and accrued expenses
Liabilities	Operating	Level 2	Other income, net
Net cash provided by used in investing activities	Beginning balance		
Other non-current liabilities	27993.62$	91138.24$	45066.38$
Pension Benefits	17365.77$	20150.1$	27593.69$

(a) Input table image.

Executive Compensation	Location		Accounts payable and accrued expenses
Liabilities	Operating	Level 2	Other income,net
Net cash provided by used in investing activities	Beginning balance		
Other non-current liabilities	27993.62$	91138.24$	45066.38$
Pension Benefits	17365.77$	20150.1$	27593.69$

(b) SLANet-1M's output.

Executive Compensation	Location	Accounts payable and accrued expenses	
Liabilities	Operating	Other income, net	
Net cash provided by used in investing activities	Level 2		
Beginning balance			
Other non-current liabilities	27993.62$	91138.24$	45066.38$
Pension Benefits	17365.77$	20150.1$	27593.69$

(c) GPT-4o's output.

	Location		Executive Compensation Accounts payable and accrued expenses
Liabilities	Operating	Level 2	Other income, net
Net cash provided by used in investing activities	Beginning balance		
Other non-current liabilities	27993.62$	91138.24$	45066.38$
Pension Benefits	17365.77$	20150.1$	27593.69$

(d) Granite Vision's output.

Executive Compensation	Location
Liabilities	Operating Level 2
Net cash provided by used in investing activities	Beginning balance
Other non-current liabilities	$27993.62
Pension Benefits	$17365.77

(e) Llama Vision's output.

Figure 6: Qualitative comparison between GPT-4o, Granite Vision, Llama Vision and SLANet-1M.

significantly outperformed both models, achieving an average inference time of less than **500 milliseconds**. The inference time refers to the time required to process the table, generate the HTML code, and save the result in Excel or CSV format.

8 Conclusion

In this paper, we evaluate SLANet on a new dataset and introduce SLANet-1M, a model trained on one million table images. We demonstrate both quantitatively and qualitatively that SLANet-1M outperforms SLANet and competes effectively with transformer-based architectures, and VLMs.

When trained on PubTabNet and the third subset of SynthTabNet, SLANet* achieves an S-TEDS score on PubTabNet that is only **0.41%** lower than the state-of-the-art (SOTA), despite using 14 times fewer parameters. When trained on PubTabNet and all subsets of SynthTabNet, its S-TEDS score on SynthTabNet is just **0.03%** below SOTA, maintaining the same efficiency.

Models	Inf. Time (ms)	Size (M)
TableFormer	10,020	53.2
UniTable Large	118,729	125
SLANet-1M	**463**	**9.2**

Table 7: Comparison of inference time on CPU.

Additionally, SLANet-1M offers faster inference time while being CPU-friendly, with only 9.2 million parameters. This makes it an ideal solution for users seeking a high-performance model without significant computational demands. Finally, we deployed SLANet-1M in the core engine of the Swiss AI center, making it accessible for those interested in testing it, it can be accessed here.

Limitations

Despite its many strengths, SLANet-1M does exhibit certain limitations. The most prominent among these is its dependence on external models for text detection and recognition. Additionally, due to its use of lightweight components, the quality of its predicted bounding boxes falls short compared to some state-of-the-art models in table recognition. Furthermore, since the majority of the training data comprises wireless tables, SLANet-1M encounters minor challenges in accurately interpreting the structure of fully wired tables. Notably, the latter limitation could be effectively mitigated through training on a more diverse and representative dataset.

Acknowledgements

This work was carried out within the framework of the Swiss-European Mobility Programme (SEMP), during a semester exchange from the University of Florence to the College of Engineering and Architecture of Fribourg. This work was also supported financially by the research grant Swiss AI Center 135788/IA-RECHERCHE24-11 and the institute iCoSys from the Swiss Universities of Applied Science HES-SO. We also thank the anonymous reviewers for their valuable feedback during the preparation of this paper.

References

Meta AI. 2025. Llama 3.2-vision: Large language and vision model. https://ollama.com/library/llama3.2-vision. Accessed: 2025-04-12.

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, et al. 2024. Docling technical report. *arXiv preprint arXiv:2408.09869*.

Cheng Cui, Tingquan Gao, Shengyu Wei, Yuning Du, Ruoyu Guo, Shuilong Dong, Bin Lu, Ying Zhou, Xueying Lv, Qiwen Liu, et al. 2021. Pp-lcnet: A lightweight cpu convolutional neural network. *arXiv preprint arXiv:2109.15099*.

Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. 2020. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. 2023. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11134–11143.

Chenxia Li, Ruoyu Guo, Jun Zhou, Mengtao An, Yuning Du, Lingfeng Zhu, Yi Liu, Xiaoguang Hu, and Dianhai Yu. 2022. Pp-structurev2: A stronger document analysis system. *arXiv preprint arXiv:2210.05391*.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768.

Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. 2021. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623.

OpenAI. 2024. Gpt-4o technical report. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-04-12.

Mateusz Pawlik and Nikolaus Augsten. 2016. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173.

ShengYun Peng, Aishwarya Chakravarthy, Seongmin Lee, Xiaojing Wang, Rajarajeswari Balasubramaniyan, and Duen Horng Chau. 2024. Unitable: Towards a unified framework for table recognition via self-supervised pretraining. *arXiv preprint arXiv:2403.04822*.

Lei Sheng and Shuai-Shuai Xu. 2024. Pdftable: A unified toolkit for deep learning-based table extraction. *arXiv preprint arXiv:2409.05125*.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, et al. 2025. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *arXiv preprint arXiv:2502.09927*.

Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. 2020. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v:

A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. 2021. Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: table recognition to html. *arXiv preprint arXiv:2105.01848*.

Guanghua Yu, Qinyao Chang, Wenyu Lv, Chang Xu, Cheng Cui, Wei Ji, Qingqing Dang, Kaipeng Deng, Guanzhong Wang, Yuning Du, et al. 2021. Pp-picodet: A better real-time object detector on mobile devices. *arXiv preprint arXiv:2111.00902*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

A Quantitative comparison with VLMs

Model	Finance		Marketing		PubTabNet		Sparse	
	Number of Samples							
	10	50	10	50	10	50	10	50
Llama Vision 3.2	53.80	43.17	37.02	41.22	49.83	46.31	23.23	30.66
Granite Vision 3.2	76.30	72.40	58.54	58.82	81.04	80.04	46.04	40.10
SLANet-1M (ours)	**99.50**	**99.48**	**99.78**	**99.16**	**99.92**	**99.55**	**97.69**	**99.20**

Table 8: Quantitative results (S-TEDS) comparison between Llama Vision, Granite Vision and SLANet-1M.

We selected two newly available large vision-language models (VLMs), Granite Vision 3.2 (Team et al., 2025) and Llama Vision 3.2 (AI, 2025), to compare quantitatively against SLANet-1M. We also evaluated MiniCPM-v (Yao et al., 2024), but its performance was insufficient for inclusion in the final comparison.

Following the methodology from (Peng et al., 2024), we randomly sampled a few images from each subset of the SynthTabNet dataset and conducted two experiments. In the first, we selected 10 images per subset; in the second, 50 images per subset. For each image, the VLMs were prompted with: *"Based on the table in the image, please generate the corresponding HTML code. Output only the HTML code."* We then computed the S-TEDS score for each output.

The results, shown in Table 8, clearly demonstrate that SLANet-1M significantly outperforms both Llama Vision and Granite Vision. Notably, while Granite Vision exhibited the strongest performance among the tested VLMs, it struggled considerably when processing large, information-dense tables.

B More Qualitative Results

Figures 7 and 8 present a qualitative comparison between SLANet-1M and Granite Vision. Since Granite Vision showed the best quantitative performance among the VLMs we evaluated, we chose it for a more in-depth qualitative analysis.

In Figure 7, panel (a) shows the input image, which comes from the PubTabNet subset of the SynthTabNet test set. Panel (b) displays the output of SLANet-1M, which achieves a perfect S-TEDS score of 1.00. While a few minor content errors are visible in some cells, these are attributable to limitations in the external models used for text detection and recognition, not SLANet-1M itself. Panel (c) shows Granite Vision's output, with a significantly lower S-TEDS score of 0.7658. The model incorrectly merges some cells, produces the wrong number of columns, and introduces an excess of blank cells.

In Figure 8, panel (a) shows an input image taken from the Finance subset of the SynthTabNet test set. Once again, SLANet-1M achieves a perfect S-TEDS score of 1.00, as shown in panel (b). In this case, Granite Vision in panel (c) performs noticeably better than in the previous example, though still not at SLANet-1M's level. This improvement can be attributed to the simpler and less structured layout of the input table.

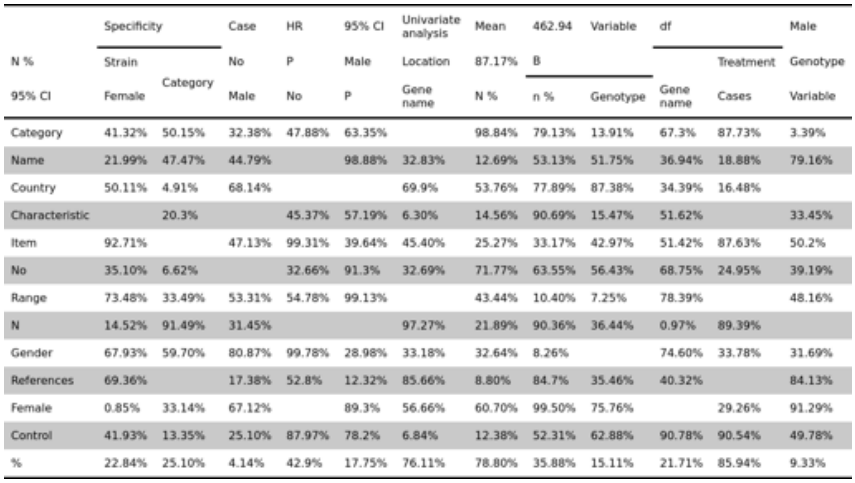
These qualitative results further support the superiority of SLANet-1M over some of the most recent VLMs in handling complex table understanding tasks.

C SLANet vs SLANet* vs SLANet-1M on PubTabNet

Models	TEDS	S-TEDS
SLANet (Li et al., 2022)	**95.89**	97.01
SLANet* (ours)	95.83	**97.48**
SLANet-1M (ours)	95.77	97.36

Table 9: Comparison on PubTabNet of models based on TEDS, S-TEDS.

Table 9 shows that SLANet-1M underperforms SLANet* on PubTabNet, likely due to SLANet* overfitting on the PubTabNet validation set, which was the only validation set used during its training.



	Specificity		Case	HR	95% CI	Univariate analysis	Mean	462.94	Variable	df		Male
N %	Strain		No	P	Male	Location	87.17%	B			Treatment	Genotype
95% CI	Female	Category	Male	No	P	Gene name	N %	n %	Genotype	Gene name	Cases	Variable
Category	41.32%	50.15%	32.38%	47.88%	63.35%		98.84%	79.13%	13.91%	67.3%	87.73%	3.39%
Name	21.99%	47.47%	44.79%		98.88%	32.83%	12.69%	53.13%	51.75%	36.94%	18.88%	79.16%
Country	50.11%	4.91%	68.14%			69.9%	53.76%	77.89%	87.38%	34.39%	16.48%	
Characteristic		20.3%		45.37%	57.19%	6.30%	14.56%	90.69%	15.47%	51.62%		33.45%
Item	92.71%		47.13%	99.31%	39.64%	45.40%	25.27%	33.17%	42.97%	51.42%	87.63%	50.2%
No	35.10%	6.62%		32.66%	91.3%	32.69%	71.77%	63.55%	56.43%	68.75%	24.95%	39.19%
Range	73.48%	33.49%	53.31%	54.78%	99.13%		43.44%	10.40%	7.25%	78.39%		48.16%
N	14.52%	91.49%	31.45%			97.27%	21.89%	90.36%	36.44%	0.97%	89.39%	
Gender	67.93%	59.70%	80.87%	99.78%	28.98%	33.18%	32.64%	8.26%		74.60%	33.78%	31.69%
References	69.36%		17.38%	52.8%	12.32%	85.66%	8.80%	84.7%	35.46%	40.32%		84.13%
Female	0.85%	33.14%	67.12%		89.3%	56.66%	60.70%	99.50%	75.76%		29.26%	91.29%
Control	41.93%	13.35%	25.10%	87.97%	78.2%	6.84%	12.38%	52.31%	62.88%	90.78%	90.54%	49.78%
%	22.84%	25.10%	4.14%	42.9%	17.75%	76.11%	78.80%	35.88%	15.11%	21.71%	85.94%	9.33%

(a) Input table image (From PubTabNet subset of SynthTabNet).

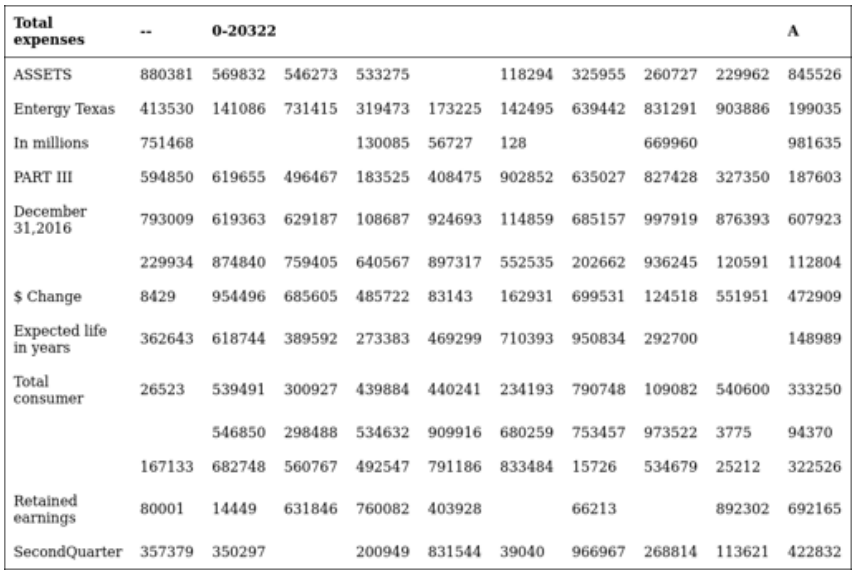
	Specificity		**Cose**	**HR**	**95% CI**	Univariate analysis	Mean	462.94	Variable	**df**		Male
N %	Strein		No	p	Male	**Location**	87.17%	B			Trewtment	**Genotype**
95% CI	Female	**Category**	Mal e	No	P	Gene name	**N %**	**n %**	**Genotype**	name Gene	**Cases**	**Variable**
Category	41.32%	50.15%	32.38%	47.88%	63.35%		98.84%	79.13%	13.91%	67.3%	87.73%	3.39%
Nome	21.99%	47.47%	44.79%		98.88%	32.83%	12.69%	53.13%	51.75%	36.94%	18.88%	79.16%
Country	50.11%	4.91%	68.14%			69.9%	53.76%	77.89%	87.38%	34.39%	16.48%	
Characteristic		20.3%		45.37%	57.19%	6.30%	14.56%	90.69%	15.47%	51.62%		33.45%
Item	92.71%		47.13%	99.31%	39.6 4%	45.40%	25.27%	33.17%	42.97%	51.42%	87.63%	50.2%
No	35.10%	6.62%		32.66 %	93.3%	32.69%	73.77%	63.55%	56.43%	68.75%	24.95%	39.19%
Range	73.48%	33.49%	53.31%	54.78%	99.13%		43.44%	10.40%	7.25%	78.39%		48.16%
N	14.52%	91.49%	31.45%			97.27%	21.89%	90.36%	36.44%	0.97%	89.39%	
Gender	67.93%	59.70%	80.87%	99.78%	28.98%	33.18%	32.64%	8.26%		74.60%	33.78%	31.69%
References	69.36%		17.38%	52.8%	12.32%	85.66%	8.80%	84.7%	35.46%	40.32%		84.13 %
Female	0.85%	33.14%	67.12%		89.3%	56.66%	60.70%	99.50%	75.76%		29.26%	91.29%
Control	43.93%	13.35%	25.10%	87.97%	78.2%	6.84%	12.38%	52.31%	62.88%	90.78%	90.54%	49.78%
%	22.84%	25.10%	4.14%	42.9%	17.75%	76.11%	78.80%	35.88%	15.11%	21.71%	85.94%	9.33%

(b) SLANet-1M's output (S-TEDS = 1.00).

Category	Specificity		Case No	HR P	95% CI		Univariate analysis Location	Mean 87.17%	462.94 B	Variable	df	Male			
	Strain Female	Category Male			No P	Genotype						Treatment	Genotype name	Cases	Variable
Name	21.99%	47.47%	44.79%				98.84%	79.13%	13.91%	67.3%	87.73%	3.39%			
Country	50.11%	4.91%	68.14%				69.0%	53.76%	77.89%	87.38%	34.39%	16.48%			
Characteristic		20.3%					14.56%	90.69%	15.47%	51.62%		33.45%			
Item	92.71%		47.13%	99.31%	39.64%	45.0%	25.27%	33.17%	42.97%	51.42%	87.63%	50.2%			
No	35.10%	6.62%		32.66%	91.3%	32.69%	71.77%	63.55%	56.43%	68.75%	24.95%	39.19%			
Range	73.48%	33.49%	53.31%	54.76%	99.13%		43.44%	10.40%	7.25%	78.39%		48.16%			
N	14.52%	91.49%	31.45%				97.27%	21.89%	90.36%	36.44%	0.97%	89.39%			
Gender		67.03%	59.70%	80.87%	99.78%	28.98%	33.18%	8.26%			74.60%	33.78%	31.69%		
References	69.36%		17.38%	52.8%	12.32%	85.66%	8.80%	84.7%	35.46%	40.32%		84.13%			
Female	0.85%	33.14%	67.12%			89.3%	56.66%	60.70%	99.50%	75.76%		29.26%	91.29%		
Control	41.93%	13.35%	25.10%	87.97%	78.2%	6.84%	12.38%	52.31%	62.88%	90.78%	90.54%	49.78%			
%	22.84%	25.10%	4.14%	42.9%	17.75%	76.11%	78.80%	35.88%	15.11%	21.71%	85.94%	9.33%			

(c) Granite Vision's output (S-TEDS = 0.7658).

Figure 7: Qualitative comparison between Granite Vision and SLANet-1M.



Total expenses	--	**0-20322**								**A**
ASSETS	880381	569832	546273	533275		118294	325955	260727	229962	845526
Entergy Texas	413530	141086	731415	319473	173225	142495	639442	831291	903886	199035
In millions	751468			130085	56727	128		669960		981635
PART III	594850	619655	496467	183525	408475	902852	635027	827428	327350	187603
December 31,2016	793009	619363	629187	108687	924693	114859	685157	997919	876393	607923
	229934	874840	759405	640567	897317	552535	202662	936245	120591	112804
$ Change	8429	954496	685605	485722	83143	162931	699531	124518	551951	472909
Expected life in years	362643	618744	389592	273383	469299	710393	950834	292700		148989
Total consumer	26523	539491	300927	439884	440241	234193	790748	109082	540600	333250
		546850	298488	534632	909916	680259	753457	973522	3775	94370
	167133	682748	560767	492547	791186	833484	15726	534679	25212	322526
Retained earnings	80001	14449	631846	760082	403928		66213		892302	692165
SecondQuarter	357379	350297		200949	831544	39040	966967	268814	113621	422832

(a) Input table image (From Finance subset of SynthTabNet).

Total expenses	--	**0-20322**								**A**
ASSETS	880381	569832	546273	533275		118294	325955	260727	229962	845526
Entergy Texas	413530	14 1086	731415	319473	173225	142495	639442	831291	903886	199035
In millions	751468			130085	56727	128		669960		981635
PART III	594850	619655	496467	1 83525	408475	902852	635027	827428	327350	187603
December 31,2016	793009	619363	629187	108687	924693	114859	685157	997919	876393	607923
	229934	874840	759405	64056 7	897317	552535	202662	936245	120591	112804
$ Change	8429	954496	685605	48 5722	83143	162931	699531	124518	551951	472909
Expected life in years	362643	618744	389592	273383	469299	710393	950834	292700		148989
Total Consumer	26523	539491	300927	439884	440241	234193	790748	109082	540600	333250
		546850	298488	534632	909916	680259	753457	973522	3775	94370
	167133	682748	560767	492547	791186	833484	15726	534679	25212	322526
Retained earnings	80001	14449	631846	760082	403928		66213		892302	692165
SecondQuarter	357379	350297		200949	831544	39040	966967	268814	113621	422832

(b) SLANet-1M's output (S-TEDS = 1.00).

	Total expenses				- 0-20322					A
ASSETS	880381	569832	546273	533275	118294	325955	260727	229962	845526	
Entergy Texas	413530	141086	731415	319473	173225	142495	639442	831291	903886	199035
In millions	751468			130085	56727	128		669660		981635
PART III	594850	619655	496467	183525	408475	902852	635027	827428	327350	187603
December 31, 2016	793009	619363	629187	108687	924693	114859	685157	997919	876393	607923
	229934	874840	759405	640567	897317	552535	202662	936245	120591	112804
$ Change	8429	954496	685605	485722	83143	162931	699531	124518	551951	472909
Expected life in years	362643	618744	389592	273383	469299	710393	950834	292700		148889
Total consumer	26523	539491	300927	439884	440241	234193	790748	109082	540600	333250
	546850	298488	534632	909916	80259	753457	973522	3775	94370	
	167133	682748	560767	492547	791186	833484	15726	534679	25212	322526
Retained earnings	80001	14449	631846	760082	403928		66213		892302	692165
Second-Quarter	357379	350297		200949	831544	39040	966967	268814	113621	422832

(c) Granite Vision's output (S-TEDS = 0.9070).

Figure 8: Qualitative comparison between Granite Vision and SLANet-1M.

Simulating Human Interactions for Social Behaviour Coaching

Daniela Komenda, Livio Bürgisser, Noémie Käser, and Alexandre de Spindler
Zurich University of Applied Sciences, Winterthur, Switzerland
{komendan,buergli1,kaeseno1,desa}@zhaw.ch

Abstract

Many individuals struggle with informal interactions like small-talk, which are vital in daily and professional settings. We introduce a conversational agent that combines a state-based interaction model with a social behaviour regulation (SBR) layer to provide structured coaching and real-time conversational modulation. The agent dynamically addresses issues such as oversharing or topic divergence and triggers coaching interventions based on user disengagement or inappropriateness. An exploratory study with neurodivergent-focused educators suggests the system's potential to foster socially appropriate communication. Our work shows how modular prompt orchestration can enhance both adaptability and the pedagogical value of conversational agents.

1 Introduction

Many individuals face challenges when navigating everyday verbal interactions, particularly in social or professional settings where informal conversations or spontaneous exchanges are expected. This is especially true for those with social communication difficulties, such as individuals on the autism spectrum or others who experience anxiety, cultural dissonance, or uncertainty around conversational norms.

Traditional approaches to improving these skills include self-directed learning through books, online courses, or video resources, as well as formal interventions such as coaching or workshops. However, learners lack opportunities to repeatedly engage in realistic, simulated interactions that provide immediate and contextualised feedback. Without iterative practice in conversation scenarios, the ability to generalise and apply new strategies in real-world settings may remain limited.

Conversational agents, powered by modern language models, offer an opportunity to bridge this gap. However, designing such agents for small-talk training demands more than simply role-playing a dialogue partner. For example, a training may require that the agent can realistically simulate a small-talk counterpart, regulate its conversational style to help uncover user deficiencies, detect those deficiencies in real time, deliver targeted coaching interventions, and return the user to the simulated scenario to apply what they have learnt. Achieving such complex agent behaviour necessitates multiple layers of adaptability, where the agent dynamically adjusts its behaviour based on various user cues and interaction patterns. Additionally, controlling a powerful language model across these layers must be done with precision to ensure consistent, interpretable, and reliable outputs.

We present a chatbot-based coaching system that demonstrates how two layers of adaptation can be specified independently yet operate in tandem. The system combines state-based adaptability, where structured prompts define interaction flows and transitions such as from simulated small-talk to coaching, with social behaviour regulation, which dynamically modulates in-state behaviour based on real-time conversational cues detections. By dissecting instructions into modular, minimal prompts, we retain tight control over the language model's behaviour while leveraging its natural ability to engage in open, human-like conversations. This layered and modular approach to agent adaptability enables more realistic and pedagogically effective coaching scenarios for social communication training, creating robust and responsive verbal interactions for improving small-talk skills, and serving as a building block for more immersive agentic systems such as avatars or social robots.

2 Related Work

Large language models (LLMs) have transformed conversational AI by enabling agents to engage in open-ended, context-aware dialogue. This progress has opened new opportunities for coaching appli-

cations, where agents must sustain natural conversations while guiding users toward learning goals (Aymerich-Franch and Ferrer, 2022). However, to effectively enhance learning, timely and contextualised feedback is essential (Hattie, 2008; Ajogbeje, 2023). Therefore, coaching systems must go beyond the conversational fluency LLMs inherently provide to detect user behaviours and deliver pedagogically meaningful interventions (Liu et al., 2025).

One key layer is the structuring of interaction flows to implement pedagogic coaching strategies. While fine-tuning LLMs to adopt specific coaching behaviours is possible, such methods are often resource-intensive, inflexible, and impractical when designing agents that must support a variety of coaching or training methods (Hadi et al., 2023). As a result, prompt engineering has emerged as a more feasible alternative for dynamically shaping agent behaviour (White et al., 2023). Yet, complex coaching scenarios often demand multi-step and layered prompts, which can increase the risk of inconsistent or unreliable outputs if merged into one large prompt (Long et al., 2024).

To address these challenges, the PROMISE framework (Wu et al., 2024a,b) introduced a state-based prompt orchestration approach. By decomposing complex instructions into modular and precise prompts tied to conversational states and transitions, PROMISE enhances LLM controllability (Helland et al., 2023) and supports the creation of structured, coherent coaching dialogues while leveraging the model's generative capabilities.

However, structuring the dialogue alone is insufficient for effective coaching, which also relies on real-time social adaptability. Persuasion techniques, conversational tone modulation, and user-tailored feedback must be dynamically selected based on in-the-moment user behaviour and conversational cues (Woolf et al., 2009). Such factors that cannot be fully predefined at design time.

We therefore extend PROMISE with a Social Behaviour Regulation (SBR) layer, which enables conversational agents to detect and respond to verbal cues such as oversharing, awkwardness, or deep talk divergence. This two-layered approach separates dialogue management via PROMISE from fine-grained behavioural modulation via SBR, allowing for orthogonal and layered adaptability. Together, these mechanisms enhance the agent's ability to deliver personalised and impactful coaching interventions, while maintaining reliable control over the language model's behaviour.

3 Use Case

We present a conversational agent designed to support users in practising small-talk while receiving adaptive, real-time coaching as exemplified in Fig. 1. The agent applies two layers of adaptability, a state-based interaction model that transitions the user between small-talk (light green) and coaching (dark green), and a SBR layer that modulates the simulated colleague's behaviour within the small-talk state.

In the first user utterance, the agent detects that the user introduces a philosophical topic, shifting the conversation into deep talk. In response, the agent's SBR adapts the behaviour by redirecting the conversation to light and casual topics. As the user exhibits signs of introversion by providing a mismatched and disengaged response in their second utterance, the agent transitions to a coaching state. As a result, the user receives feedback, suggestions for alternative responses, and the option to obtain more advice or return to the conversation.

While this scenario is highly simplified to convey the key idea within limited space, it demonstrates how in-state behaviour modulation and state-based transitions complement each other. The SBR layer helps surface conversational challenges through subtle adjustments, while the state model delivers targeted coaching interventions when necessary. Together, these layers create a responsive and iterative learning environment.

4 Approach

The agent architecture integrates two complementary layers of adaptability: state-based interaction management and SBR. Both layers rely on the PROMISE framework to structure and control language model prompts through modular and precise instructions.

4.1 State-Based Adaptability

PROMISE coordinates both state prompts for response generation and transition prompts for managing flow between states. At the core of an agent system, the state-machine-based interaction model depicted in Fig. 2 orchestrates the overall conversation flow. This model defines an outer state (grey) that ensures the agent consistently alternates be-

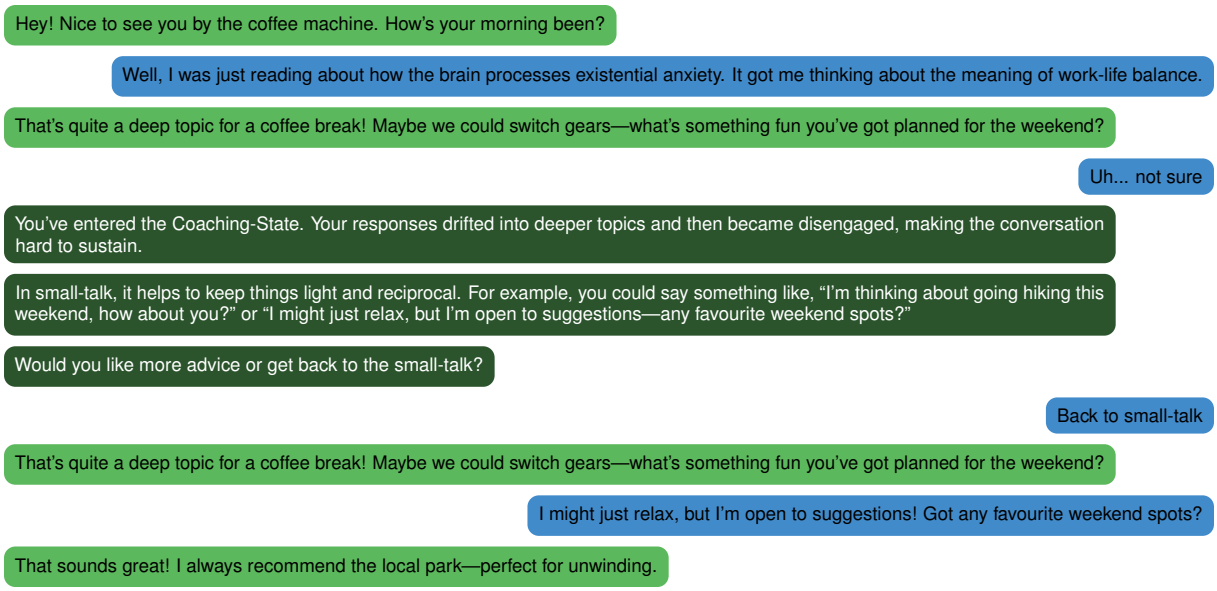


Figure 1: Example of how SBR and state transitions combine to redirect and coach the user during small-talk.

tween two roles: simulating a small-talking colleague or delivering coaching interventions.

The simulation state (green) models casual workplace interactions, allowing users to engage in small-talk with a fictitious co-worker. From this state, transitions occur when conversational issues are detected, leading into issue-specific coaching states that provide targeted feedback. Offensiveness coaching (blue) is activated when user utterances are perceived as offensive or inappropriate. Introversion coaching (purple) is activated when users display minimal engagement or withdrawal. After receiving feedback in a coaching state, users return to the small-talk simulation to apply newly acquired strategies.

4.2 Social Behaviour Regulation (SBR)

The SBR layer runs in parallel with the state machine and continuously analyses conversational cues during small-talk simulations. It performs real-time detections of conversational aspects.

For each detection, the system applies behaviour adaptations by appending specific prompt instructions to the original state prompt, as illustrated in Fig. 3. This helps to modulate the baseline behaviour of the simulated person as illustrated with the following examples.

Deep topic divergence → Redirect to light, casual topics
Awkwardness → Select familiar, casual topics
Oversharing → Limit to safe and neutral topics

These adaptations help surface conversational deficiencies and prepare the ground for targeted coaching interventions.

4.3 Integrated Adaptation

The dual-layer architecture of our system integrates a state-based interaction model (Baseline Behaviour) with either a smalltalk state or a coaching state, layered on top of an SBR module. The baseline behaviour, implemented via the PROMISE framework, manages the overall conversation flow through a state machine. Upon receiving each user utterance, PROMISE appends it to the conversation history, determines the current state, and checks transition conditions to decide whether to move into a new state. Transition conditions are encoded as prompts used to instruct the underlying language model to analyse the conversation history and make a decision. The active state points at the prompt to be used to generate the agent's response to the user. As a result, different prompts are used depending on the state currently active.

Running in parallel with the state machine, the SBR layer processes the same interaction history to detect conversational issues using dedicated prompts. Based on these detections, it selects behaviour regulation strategies, encoded as prompt elements, and appends them to the active state prompt provided by the baseline behaviour. This combined prompt enables the agent to produce socially adaptive and context-sensitive responses. Thus, while the state model governs high-level conversation flow and role-switching between simulation and coaching, the SBR layer fine-tunes in-state

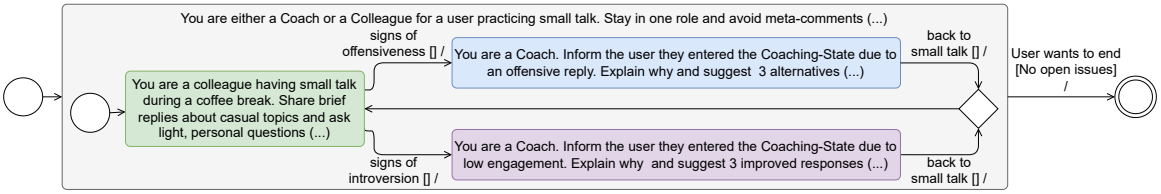


Figure 2: State model with transitions between small-talk and coaching based on detected conversational issues.

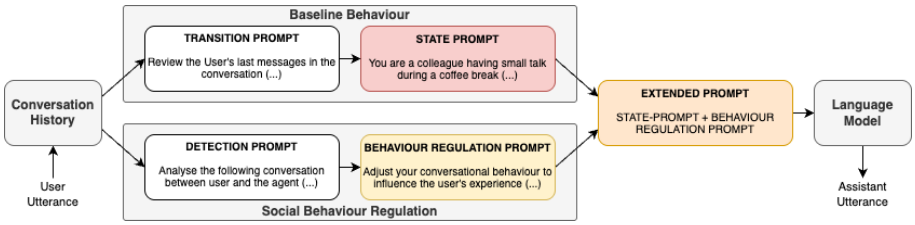


Figure 3: Social behaviour regulation flow combining state prompt and behaviour regulation prompt

responses, ensuring that conversations remain both structured and dynamically responsive to evolving user inputs.

5 Validation

To assess the chatbot's ability to support users in improving small-talk skills, we conducted exploratory user testing with eight special needs teachers specialised in supporting neurodivergent learners. The objective was to evaluate whether the agent could effectively simulate workplace small-talk scenarios and deliver timely coaching interventions.

Participants engaged in five to ten minute sessions with the chatbot, alternating between two predefined colleague personas (male, female). The agent successfully maintained realistic small-talk while triggering coaching interventions in cases of conversational disengagement (introversion) or inappropriate responses (offensiveness). Notably, only three users entered a coaching state, all cases activating the introversion coaching, while no occurrences of offensiveness were observed.

User evaluations were given by responses to an adapted chatbot usability questionnaire (Holmes et al., 2019). The responses summarised in Fig. 4 highlight positive perceptions of the chatbot's conversational realism, appropriateness, and perceived understanding. The three users who entered coaching states also rated the provided feedback as helpful and relevant to small-talk scenarios.

While the small sample limits generalisability, the results suggest that the dual-layered adaptability approach creates a plausible and responsive environment for practising small-talk.

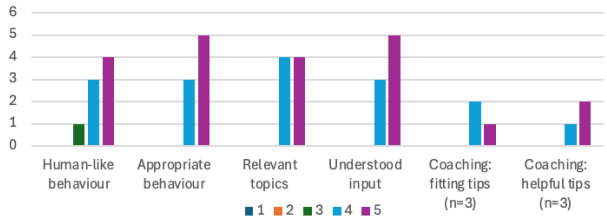


Figure 4: Chatbot usability ratings (1 = low, 5 = high)

6 Conclusion

We introduced a conversational coaching agent designed to help users improve their small-talk skills through a layered approach to adaptability. By combining a state-based interaction model with SBR, the system delivers both structured conversation flows and real-time behavioural adjustments. This dual mechanism enables the agent to simulate realistic social interactions, detect conversational challenges, and deliver targeted coaching interventions. The modularity of the approach ensures precise control over the language model's behaviour, while maintaining the flexibility needed to support dynamic, user-centred training scenarios.

Initial user testing suggests that the agent can provide a usable and engaging environment for practising small-talk. However, due to the small number of test participants, further studies are necessary to validate these findings.

In future work, we aim to extend this framework to support richer interaction modalities, including non-verbal cues. Additionally, the layered design can be adapted to other socially complex domains such as customer support, negotiation training, or conflict resolution.

References

Oke James Ajogbeje. 2023. Enhancing classroom learning outcomes: The power of immediate feedback strategy. *International Journal of Disabilities Sports and Health Sciences*, 6(3):453–465.

Laura Aymerich-Franch and Iliana Ferrer. 2022. Investigating the use of speech-based conversational agents for life coaching. *International Journal of Human-Computer Studies*, 159:102745.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

John Hattie. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

Solveig Helland, Elena Gavagnin, and Alexandre de Spindler. 2023. Divide et impera: Multi-transformer architectures for complex NLP-tasks. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 70–75, Neuchatel, Switzerland. Association for Computational Linguistics.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics (ECCE 2019)*, pages 207–214, New York, NY, USA. ACM.

Vincent Liu, Ehsan Latif, and Xiaoming Zhai. 2025. Advancing education through tutoring systems: A systematic literature review. *arXiv preprint arXiv:2503.09748*.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F Chen. 2024. Multi-expert prompting improves reliability, safety, and usefulness of large language models. *arXiv preprint arXiv:2411.00492*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. 2009. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3/4):129–164.

Wenyuan Wu, Jasmin Heierli, Max Meisterhans, Adrian Moser, Andri Färber, Mateusz Dolata, Elena Gavagnin, Alexandre de Spindler, and Gerhard Schwabe. 2024a. Promise: A framework for model-driven stateful prompt orchestration. In *Intelligent Information Systems: CAiSE Forum 2024, Limassol, Cyprus, June 3–7, 2024, Proceedings*, volume 520 of *Lecture Notes in Business Information Processing*. Springer International Publishing.

Wenyuan Wu, Jasmin Heierli, Max Meisterhans, Adrian Moser, Andri Färber, Mateusz Dolata, Elena Gavagnin, Alexandre de Spindler, and Gerhard Schwabe. 2024b. Promise: Model-driven stateful prompt orchestration for persuasive conversational interactions. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 185–185, Chur, Switzerland. Association for Computational Linguistics.

Soft Skills in the Wild: Challenges in Multilingual Classification

Laura Vásquez-Rodríguez[1], Bertrand Audrin[2], Samuel Michel[1], Samuele Galli[3], Julneth Rogenhofer[2], Jacopo Negro Cusa[3], Lonneke van der Plas[4,1]

[1]Idiap Research Institute, Switzerland
[2]EHL Hospitality Business School, HES-SO,
University of Applied Sciences and Arts Western Switzerland, Switzerland
[3]Arca24.com SA, Switzerland
[4]Università della Svizzera italiana, Switzerland

Correspondence: laura.vasquez@idiap.ch, bertrand.audrin@ehl.ch, lonneke.vanderplas@usi.ch

Abstract

Soft skills are a crucial factor in candidate selection for recruitment. However, they are often overlooked due to the challenges in their identification. In this study, we compare soft and hard skills as well as occupations, both in terms of surface and semantic properties of the annotations and as part of an automatic extraction task, showing clear differences between the types of skills. Soft skills can be easily limited to a small number of categories, as we show in our annotation framework, which is based on well-known taxonomies. However, the way they are expressed in texts varies more widely than other entity types. These insights help to understand possible causes for the large variation in performance we see when using a multilingual BERT-based classifier for the identification of soft skills compared to other entities, which can help the community to develop more reliable algorithms for recruitment.[1]

1 Introduction

Applicant Tracking Systems (ATS) have often focused on hard skills only and neglected soft skills in their matchmaking of candidates with job openings. The notion of soft skills refers to behavioral and social abilities that people tend to possess or develop through social interactions (Heckman and Kautz, 2012). There is a debate about what these skills entail and how to label them, which, of course, makes their identification very complex, as the way they are expressed may vary from person to person.

Research on skill extraction has predominantly focused on identifying occupations or hard skills (Senger et al., 2024). Occupations are very straightforward and refer to a large set of clearly identifiable positions (e.g., "plumber" or "architect"). Hard skills are also quite straightforward to ground in existing knowledge and categorized according to employment, but many different hard skills can be relevant for each specific position (e.g., a plumber needs to have a specific skill set: blueprint comprehension, pipe installation, drilling, etc.). In that respect, there are many more hard skills than occupations. Soft skills, for their part, also known as behavioral skills (Tamburri et al., 2020), are primarily acquired in social contexts and may be independent of technical knowledge (Sayfullina et al., 2018). There is a limited set of soft skills that exist, and these skills can be relevant for a variety of jobs (e.g., "collaboration" can be useful for a plumber or an architect). Despite their importance, their identification and impact remain challenging to assess due to their abstract nature, which can explain why less attention has been given to soft skills. Specific findings remain scarce and domain-specific, often differing significantly in data, methodology, and language (Sayfullina et al., 2018; Beauchemin et al., 2022; Zhang et al., 2022).

One of the main challenges is that these approaches have mostly focused on extracting soft skills from job advertisements (also referred to as job offers). However, soft skills in job advertisements are more likely to be standardized and are an opportunity for the company to develop its employer branding (Elving et al., 2013). In contrast, soft skills may appear very differently in resumes, sometimes less explicitly or straightforwardly, if at all, especially for certain roles for which it is not common for candidates to emphasize this part of their skill set. Moreover, soft skills are often identified in later stages of recruitment, typically through interviews or work simulations.

Our study is developed within the context of SEM24 project, which is supported by the Innosuisse (Swiss Innovation Agency). This project aims to enhance multilingual, multidomain competency detection in the European job market, with a particular focus on developing explainable algorithms for fairer recruitment processes. In this study, we

[1]We will release our results on GitHub: https://github.com/idiap/multilingual_skill_extraction

Surface Properties →		Total Entities			Unique Entities			Unique/Total Ratio			Avg. Len		
Text ↓	Language	Hard	Occ	Soft	Hard	Occ	Soft	Hard	Occ	Soft	Hard	Occ	Soft
Jobs (our annotations)	EN	878	322	326	777	109	255	0.88	0.34	0.78	37.53	22.19	30.36
	FR	1119	126	777	886	107	282	0.79	0.85	0.36	30.67	29.13	19.61
	IT	704	116	285	651	101	216	0.92	0.87	0.76	53.67	22.52	34.30
	PT	667	81	257	560	68	89	0.84	0.84	0.35	30.87	21.37	15.42
Jobs (Sayfullina)	EN	-	-	7403	-	-	1140	-	-	0.15	-	-	13.87
Jobs (Fijo)	FR	-	-	932	-	-	702	-	-	0.75	-	-	58.67
Jobs (Green)	EN	12573	2571	-	10079	1591	-	0.80	0.62	-	32.67	17.49	-
Resumes (our annotations)	EN	4024	692	520	3020	565	351	0.75	0.82	0.68	20.01	24.85	16.52
	FR	2063	645	441	1700	466	294	0.82	0.72	0.67	25.76	21.66	18.63
	IT	1985	645	464	1613	435	318	0.81	0.67	0.69	26.77	21.91	22.92
	PT	3312	729	200	2439	447	126	0.74	0.61	0.63	28.79	24.78	19.60

Table 1: We report the number of total and unique entities and its ratio, and average character length of hard skills, occupations, and soft skills in resumes and job offers.

specifically examine the detection of soft skills in multilingual environments. The key contributions of this paper are:

1. A discussion of the varying nature of soft skills across different document types and languages.
2. A comparative evaluation of multilingual soft skills identification from job offers and resumes.
3. Development of an annotation framework for identifying soft skills in multilingual job offers and resumes. To the best of our knowledge, this is the first time the O*NET resource has been leveraged for guidelines design in the soft skill extraction task, extending its use beyond a taxonomy of occupational terms.

2 Methodology

The main objective of this research is to characterize and understand the variable nature of soft skill annotations in job offers and resumes in a multilingual setting. As a first step, we collected multilingual resumes and job offers (Section 2.1) and annotated relevant entities following in-house annotation guidelines (Section 2.2). We then analyzed the extracted entities, characterizing the datasets and conducting a semantic analysis to explore the main differences between job offers and resumes across different languages (Section 2.3). Finally, we assessed the impact of annotation variability on performance through experiments on an entity-based classification task (Section 2.4).

2.1 Data

We collected a total of 800 resumes and job offers in 4 different languages: English (EN), French (FR), Italian (IT), and Portuguese (PT), and multiple domains such as engineering, administration, and management from our industrial partner of the SEM24 project. All documents were annotated by HR specialists according to the annotation guidelines (See Section 2.2) using a span-based Named Entity Recognition (NER) approach, where relevant entities are sequences of multiple tokens that are explicitly mentioned in the text. The annotation task was performed using the Docanno (Nakayama et al., 2018) tool. Native speakers annotated texts in French and Italian, whereas annotators with C2 proficiency annotated the other languages.

2.2 Annotations Guidelines

We extended the annotation guidelines proposed by Vásquez-Rodríguez et al. (2024) to include soft skills. Two HR researchers elaborated the proposed guide based on the O*NET taxonomy (Peterson et al., 2001) and the HEXACO personality inventory (Ashton and Lee, 2007). The categorization and identification of soft skills followed three predefined categories, then divided into a total of 21 subcategories as follows:[2]

- **Social Skills:** Coordination, Instructing, Negotiation, Persuasion, Service Orientation, and Social Perceptiveness.
- **Thinking Skills:** Active Learning, Active Listening, Complex Problem Solving, Critical Thinking, Judgment and Decision Making, Learning Strategies, Monitoring, and Time Management.

[2]These categorizations were used as a guide for the annotators to define a clear criterion that could define more precisely the concept of soft skills. However, the final labeling of entities for training was unified into the "Soft Skill" label.

Evaluation →			Exact (F1-score)			Partial (F1-score)		
Skills Type	Dataset ↓	Language	Precision	Recall	F1	Precision	Recall	F1
Jobs (Our annotations)	Soft Skills	EN	0.320	0.348	0.333	0.440	0.478	0.458
		FR	0.774	0.818	**0.796**	0.823	0.869	**0.845**
		IT	0.393	0.393	0.393	0.625	0.625	0.625
		PT	0.750	0.581	0.655	0.812	0.629	0.709
Jobs (Sayfullina)		EN	0.879	0.887	**0.883**	0.921	0.93	**0.926**
Jobs (Fijo)		FR	0.354	0.429	0.388	0.54	0.655	0.592
Jobs (Our annotations)	Hard Skills	EN	0.286	0.306	0.294	0.460	0.507	0.480
		FR	0.399	0.465	**0.427**	0.551	0.641	**0.589**
		IT	0.241	0.263	0.251	0.478	0.523	0.499
		PT	0.395	0.495	**0.438**	0.512	0.644	**0.569**
Resumes (Our annotations)	Soft Skills	EN	0.415	0.347	0.378	0.537	0.449	0.489
		FR	0.575	0.455	**0.508**	0.675	0.535	0.597
		IT	0.525	0.544	**0.534**	0.636	0.658	**0.647**
		PT	0.158	0.200	0.176	0.263	0.333	0.294
	Hard Skills	EN	0.354	0.356	0.355	0.490	0.491	0.490
		FR	0.352	0.389	0.369	0.501	0.555	0.526
		IT	0.411	0.426	0.418	0.577	0.600	0.588
		PT	0.449	0.562	0.497	0.548	0.688	**0.608**

Table 2: We report the exact (i.e., the entire entity was detected) and partial (i.e., entity was detected partially) scores for the soft and hard skills detection of the multilingual BERT model.

- **Personality Traits:** Achievement Orientation, Adjustment, Conscientiousness, Independence, Interpersonal Orientation, Practical Intelligence, and Social Influence.

Before the annotation process, all annotators were trained during an in-person workshop to discuss the final annotation guidelines and solve any disagreements between the participants. The strengths of our annotation approach lie in the fact that the annotation guidelines were developed by HR researchers following clear guidelines based on reliable frameworks. We further differ from previous work because we do not follow a static taxonomy of concepts that are expected to be explicit in the text (Sayfullina et al., 2018), and annotators are not limited to any particular domain (Beauchemin et al., 2022).

2.3 Surface Properties and Semantic Analysis

As for the surface properties of texts, we report the number of tokens and unique types (including the ratio between these metrics) of hard skills, soft skills, and occupations in the annotations in Table 1. Also, we calculated the average length of the soft skills measured by the number of characters. For all the metrics, we report the results by document type (i.e., resumes vs job offers) and language.

Another relevant aspect is the semantic similarity between soft skills in different contexts. For this analysis, we compare the extracted soft skills using the t-SNE algorithm (van der Maaten and Hinton, 2008). This algorithm reduces high-dimensional embeddings into a lower-dimensional space, where similar data points are grouped based on local similarities in their original space. The visualization of clusters shows skills that are potentially equivalent within the selected samples of documents. We highlight our results based on three different scenarios: job offers vs resumes, by language, and both resumes/job offers by language. We encoded the extracted soft skills using multilingual[3] SBERT pretrained embeddings (Reimers and Gurevych, 2020). These embeddings were then input into the t-SNE algorithm[4] using the scikit-learn Python library (Pedregosa et al., 2011). To visualize potential clusters in the data, we experimented with various perplexity levels (i.e., 10, 30, 50, 70, 100), with 50 yielding the best results across all our experiments.

2.4 Skill extraction experiments

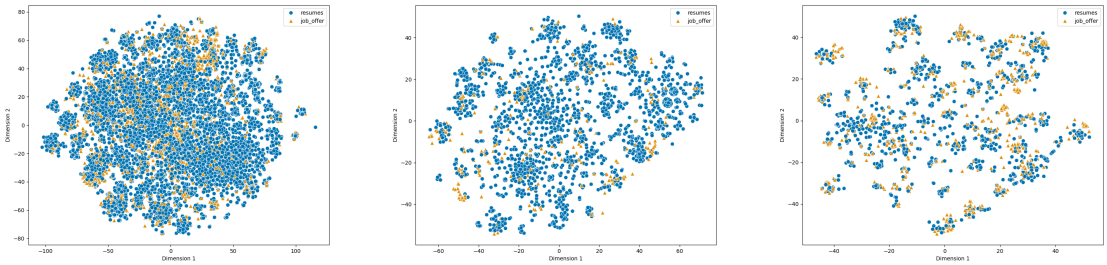
To explore the impact of soft skills annotation variability on performance (measured by F1-score) in the skill extraction task, we trained a supervised system for token classification using a span-based approach.[5] We employed a BERT-based multilingual model[6] and fine-tuned it using our manually annotated skill datasets. The corpus was split into

[3] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[4] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[5] We acknowledge that skills can sometimes be inferred from text, and the span-based approach may not always be the most suitable. However, we chose to leverage existing online resources and tools mostly designed for a span-based approach.

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



(a) Hard Skills (b) Occupations (c) Soft Skills

Figure 1: t-SNE visualization of all entities embeddings, comparing job offers and resumes.

train, validation, and test subsets (80/10/10) within each language and document type. In addition, we trained the model on datasets available from previous work, Sayfullina (Sayfullina et al., 2018) for soft skills in English, Fijo (Beauchemin et al., 2022) for French,[7] and the Green dataset (Green et al., 2022) for hard skills in English. We differ from previous work in soft skills, so that we rely on the knowledge of the HR researchers and detailed guidelines for the identification of entities in the text, rather than following a limited taxonomy of concepts (Sayfullina et al., 2018). Similarly, our proposed categories are based on updated resources that are more specific and still relevant, and available today. For evaluation, we post-processed the results using the IOB format (Ramshaw and Marcus, 1999) and then calculated precision, recall, and F1-score using the *nervaluate* Python library.[8]

3 Results

We present the surface-level statistics of soft skills, hard skills, and occupations across all datasets in Table 1. In addition, Figure 1 shows a semantic analysis comparing soft skills to hard skills and occupations. This visualization highlights how some concepts, such as hard skills, tend to form many clusters more densely scattered over the space compared to soft skills, which show a more clearly delineated representation of fewer clusters.

Furthermore, we include the main results of our evaluation in Table 2. These results represent a comparative evaluation of the skill extraction task for soft skills for both existing and in-house datasets. Also, we add the extraction results for hard skills as a reference, not only across document types but also across languages. Finally, in Table 3, we include examples of system outputs (i.e., soft skills) for a closer look at the variability between the datasets and their impact on prediction.

4 Discussion

The quality of automatic extractions of soft skills has high variability. In Table 2, we observe how performance in soft skills (measured by exact F1-scores) varies in ranges of 0.1-0.8, while in hard skills, values are more stable (between 0.2-0.4).

For the detection of soft skills in job advertisements, it is often hard to clearly distinguish between those soft skills that pertain to the specific job at hand and others that are more used for employer branding purposes (e.g., "working in a diverse team" might refer to open-mindedness as a requirement for candidates, or as an employer branding signal that the team is diverse). In resumes, soft skills are often less standardized, and applicants are likely to use a broader variety of terms to refer to them. Moreover, soft skills are often less clearly defined, and some statements might refer to several soft skills at the same time (e.g., "ability to adapt to customers' needs" could refer to flexibility, but also to customer-centricity). This aspect of subjectivity in soft skills expression makes it a challenging task.

We observe the same variability in Table 1. There are fewer underlying categories of soft skills than occupations or hard skills, but the way they are expressed varies widely. We see more entities (both in total and unique) for soft skills in several languages and document types than the underlying 21 subcategories. The ratio between unique and total number of annotated skills shows that although soft skills are generally expressed with shorter phrases, the variability is relatively high. In Figure 1c, we see that hard skills are scattered

[7] Details of our training procedure are provided in Appendix A.1.

[8] https://pypi.org/project/nervaluate/

Dataset	Example
Jobs (Sayfullina)	R: you will have excellent communication and leadership abilities P: you will have excellent communication and leadership abilities R: be a national provider of independent sector complex healthcare P: be a national provider of independent sector complex healthcare
Jobs (Fijo)	R: anglais intermediaire connaissances excel intermediaire - avance attitude positive et aimant travailler en equipe [9] P: anglais intermediaire connaissances excel intermediaire avance attitude positive et aimant travailler en equipe R: service a la clientele traiter, analyser et gerer les correspondances aupres de la clientele interne[10] P: service a la clientele traiter, analyser et gerer les correspondances aupres de la clientele interne
Jobs (Our Annotations)	R: customer service skills great attention to detail P: customer service skills great attention to detail R: follow verbal and written instructions be able to work quickly and concisely under pressure P: follow verbal and written instructions be able to work quickly and concisely under pressure
Resumes (Our Annotations)	R: curious and thoughtful person with good inventive P: curious and thoughtful person with good inventive R: analysis of feasibility problems and ability to problem solving design P: analysis of feasibility problems and ability to problem solving design

Table 3: Soft skill detection examples for the multilingual BERT model in all datasets. For each example, we include the model's prediction (P) and the annotators' reference labels (R).

over the entire space, whereas occupations and soft skills are more clearly grouped in a smaller number of clusters. For soft skills, the clusters are more distinctly separated, supporting the idea of having fewer categories.

We cannot draw firm conclusions from the difference in performance between the languages, because here datasets also differ a lot, both in how annotations are done and what type of data is selected. For example, there is a significant gap between Sayfullina (0.883) and Fijo (0.33), not only because of the size of the dataset but also because of how soft skills are defined (e.g., "team working", "independent" vs "Être orienté vers l'action").[11] Overall it is difficult to pose strict boundaries to soft skills as in a span-based approach, which results in variable average lengths as demonstrated in Table 1. The verbosity of the Fijo datasets is also evident in Table 3, where large portions of the sentence are highlighted in both the predictions and the references. Whether a concept like "curious" or a phrase such as "ability to problem solving" is identified as a soft skill depends on multiple factors including the annotator's previous knowledge, model's learned patterns, the taxonomy design, and less evident influences such as language, writing style, document type, and context. The same variability can be found in our datasets because it is based on real-world data, where clients in one language are typically less varied than in another language.

We chose the span-based approach as it is convenient to compare against existing literature and to leverage existing annotation tools. However, conversational LLMs could help to categorize into broader, more general categories (e.g., thinking skills, social skills, personality, etc.). It could also mitigate the difficulties of extracting non-explicit skills, while making sure they align with human judgements, so that algorithms are trustworthy and reliable. We leave this avenue for future work.

5 Conclusions

In this paper, we have shown an analysis of the nature of soft skills and have provided experiments that test the performance of automatic soft skill identification, as compared to the extraction of hard skills and occupations. We demonstrated differences across skill types and across resumes as well as job offers in a number of languages, while resumes are absent in most previous work.

Although soft skills can be summarized in a small number of well-known categories, the variability in human expression is more pronounced than for hard skills. We show that this variability poses additional challenges for the extraction of soft skills when compared to hard skills. These results underline the importance of considering approaches that move away from a span-based approach. Similarly, resumes often present hard skills in lists or conjoined ways that significantly limit the ability to extract them precisely. In that respect, working on soft skills has been incredibly helpful in revealing challenges that are also faced in hard skill extraction but tend to be dismissed for simplification purposes.

[9] In English, "customer service process, analyze and manage correspondence with internal customers."

[10] In English, "intermediate english, intermediate to advanced excel skills, positive attitude and enjoys working in a team."

[11] In English, "being action-oriented."

Limitations

We recognize the importance of releasing both models and data to support the research community. However, due to privacy concerns and the intellectual property policies of the company, we are unable to release proprietary job offers and resumes. To support the reproducibility of our work, we instead provide our models and datasets based on publicly available data.

Acknowledgments

We would like to thank Arca24 HR specialists for their great effort in completing all annotations for our paper. Finally, we gratefully acknowledge the support from Innosuisse (grant 104.069 IP-ICT).

References

Michael C. Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

David Beauchemin, Julien Laumonier, Yvan Ster, and Marouane Yassine. 2022. "fijo": a french insurance soft skill detection dataset. *arXiv*.

Wim J L Elving, Jorinde J C Westhoff, Kelta Meeusen, and Jan-Willem Schoonderbeek. 2013. The war for talent? The relevance of employer branding in job advertisements for becoming an employer of choice. *Journal of Brand Management*, 20(5):355–373.

Thomas Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208, Marseille, France. European Language Resources Association.

James J. Heckman and Tim Kautz. 2012. Hard evidence on soft skills. *Labour Economics*, 19(4):451–464. European Association of Labour Economists 23rd annual conference, Paphos, Cyprus, 22-24th September 2011.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Norman G Peterson, Michael D Mumford, Walter C Borman, P Richard Jeanneret, Edwin A Fleishman, Kerry Y Levin, Michael A Campion, Melinda S Mayfield, Frederick P Morgeson, Kenneth Pearlman, et al. 2001. Understanding work using the occupational information network (o* net): Implications for practice and research. *Personnel psychology*, 54(2):451–492.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Luiza Sayfullina, Eric Malmi, and Juho Kannala. 2018. Learning representations for soft skill matching. In *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7*, pages 141–152. Springer.

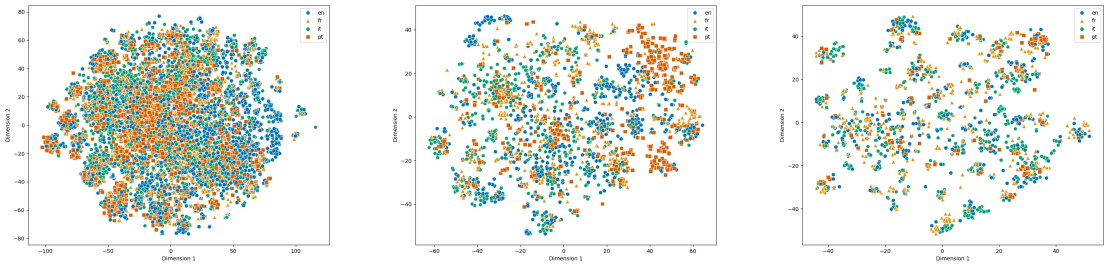
Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 1–15, St. Julian's, Malta. Association for Computational Linguistics.

Damian A. Tamburri, Willem-Jan Van Den Heuvel, and Martin Garriga. 2020. Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 391–394.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

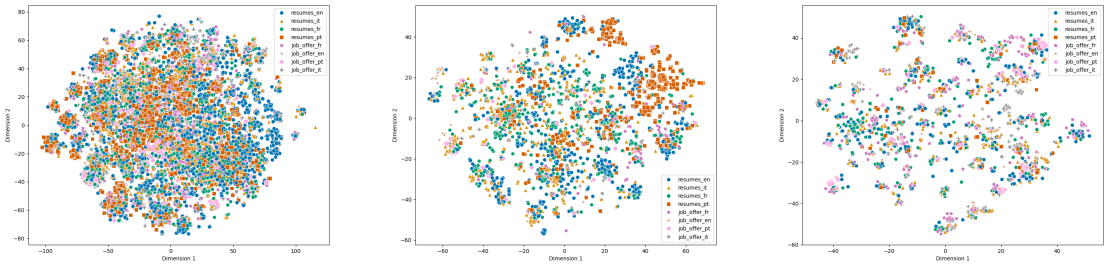
Laura Vásquez-Rodríguez, Bertrand Audrin, Samuel Michel, Samuele Galli, Julneth Rogenhofer, Jacopo Negro Cusa, and Lonneke Van Der Plas. 2024. Hardware-effective approaches for skill extraction in job offers and resumes. In *RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems.*, pages 1–12. CEUR Workshop Proceedings.

Mike Zhang, Kristian Jensen, Sif Sonniks, and Barbara Plank. 2022. SkillSpan: Hard and soft skill extraction from English job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984, Seattle, United States. Association for Computational Linguistics.



(a) Hard Skills (b) Occupations (c) Soft Skills

Figure 2: t-SNE visualization of all entities embeddings, comparing all languages.



(a) Hard Skills (b) Occupations (c) Soft Skills

Figure 3: t-SNE visualization of all entities embeddings, comparing all languages per resume and job offer.

A Appendix

A.1 Training details

As for the training parameters, we ran all experiments using 3 different seeds, then we reported the average results across all runs. The selected hyperparameters include a batch size of 16, a learning rate of 5.00×10^{-5}, and a maximum of 10 epochs. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of RAM.

A.2 Results

Further, in our analysis presented in Figure 1, we include the t-SNE visualization highlighting each language only (Figure 2) and also, considering the existing pairs of resumes-job offers and languages (Figure 3).

Using Phonemes in cascaded S2S translation pipeline

Rene Pilz and **Johannes Schneider**
Department of Computer Science & Information Systems
University of Liechtenstein, Liechtenstein
rene.pilz@uni.li, johannes.schneider@uni.li

Abstract

This paper explores the idea of using phonemes as a textual representation within a conventional multilingual simultaneous speech-to-speech translation pipeline, as opposed to the traditional reliance on text-based language representations. To investigate this, we trained an open-source sequence-to-sequence model on the WMT17 dataset in two formats: one using standard textual representation and the other employing phonemic representation. The performance of both approaches was assessed using the BLEU metric. Our findings shows that the phonemic approach provides comparable quality but offers several advantages, including lower resource requirements or better suitability for low-resource languages.

1 Introduction

According to Wang et al. (2022) simultaneous speech-to-speech (S2S) translation systems play a crucial role in enabling real-time multilingual communication. Conventionally, these systems employ a pipeline consisting of Automatic Speech Recognition (ASR), text-to-text translation, and Text-to-Speech (TTS) synthesis, each step utilizing standardized textual representations of language. However, this traditional methodology inherently depends on the existence of official written language forms and requires extensive speech data for training ASR systems, posing significant challenges for under-resourced or endangered languages. Jiang, Ahmed, Carson-Berndsen, Cahill, and Way (2011) and Do, Coler, Dijkstra, and Klabbers (2022) successfully utilized phonetic representations for under-resourced source languages. In particular, spoken dialects prevalent in many countries often lack an official written form.

Furthermore, recent progress in Text-to-Speech (TTS) models has shown a preference for phoneme-based embeddings, attributed to their improved performance. This typically involves an extra computational process where the text input is first converted to phonemes before being transformed into audio output.

To address these gaps, our study investigates whether adopting phonemic representations throughout the entire multilingual S2S pipeline provides distinct advantages. We trained an open-source sequence-to-sequence (seq2seq) model using the WMT17 dataset, comparing translation quality for English-to-German translations conducted at both textual and phonemic levels. Figure 2 shows the processes used to train and validate both models. By evaluating model outputs using standard BLEU scores we demonstrate that phoneme-based representations can offer comparable translation quality. Consequently, the phonemic approach might emerge as an advantageous alternative, particularly beneficial for under-resourced languages, while simultaneously streamlining the translation process by reducing computational complexity within the TTS step.

Figure 1: Pipeline of the cascaded S2S system (Wang, Wu, He, Huang, & Church, 2022)

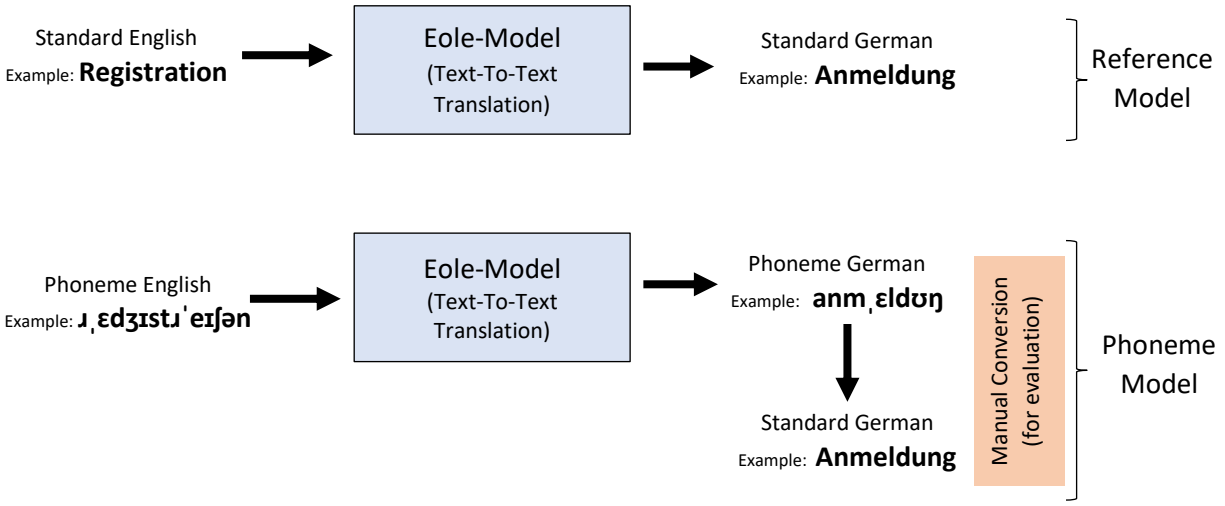


Figure 2: Reference and Phoneme Model

Contribution This study demonstrates that a sequence-to-sequence (seq2seq) model is capable of translating from English to German operating at the phonemic level with comparable quality measured using the BLEU score. Thus, phoneme-based translation might be preferable to conventional translation in particular in case of data scarcity and limited computational resources.

2 Related work

Wang et al. (2022) describe a prevalent methodology in AI-driven multilingual translation, which employs a simultaneous speech-to-speech (S2S) translation pipeline as shown in Figure 1. The process involves the use of standardized text representations of languages, adhering to the common strategy of decomposing complex problems into manageable sub-tasks:

1. Transcription of spoken language into written text.
2. Translation of text across languages.
3. Generation of spoken language from written text.

This segmentation facilitates individual analysis of each step, contingent upon the availability of standardized language text but also has limitations, i.e., the need for written language forms and extensive speech data.

The literature explored the substitution of the source language with phonetic representations but no approach replaced source and destination language by their phonetic representations.

Investigations into the latest TTS models reveal a predominant use of phoneme embedding, particularly in models based on advancements such as FastSpeech 2 (Ren et al., 2020), models based on StyleTTS 2 (Y. A. Li, Han, Raghavan, Mischler, & Mesgarani, 2023), models based on Transformer-TTS (N. Li et al., 2018), and models based on VITS (Kim, Kong, & Son, 2021). While models like Tacotron 2 (Shen et al., 2017) can operate on standard text representations, they generally exhibit improved performance with phoneme representation.

In essence, employing a TTS model reliant on phonemes necessitates the conversion of standard text into phoneme representation—an additional computational step that incurs extra processing time.

Gupta and Kumar (2021) demonstrated that seq2seq models are capable of translating between languages in text format, showcasing the versatility of these models in handling text-based translations. Therefore, this study also utilizes a seq2seq model for the translation process.

3 Materials and Methods

3.1 Model and Tools

The Eole model[1], a derivative of the OpenNMT Toolkit as developed by Klein, Kim, Deng, Senellart, and Rush (2017), was employed as the sequence-to-sequence (Seq2Seq) framework for translating English text or phonemes into German. To convert language text in German and English to phenom representation the espeak-ng[2] Framework

[1] https://github.com/eole-nlp/eole

[2] https://github.com/espeak-ng/espeak-ng

Model	Source sentence	Target sentence
Reference	Registration for the event can be submitted.	Die Anmeldung zur Veranstaltung kann vorgenommen werden.
Phoneme	ɹˌɛdʒɪstɹˈeɪʃən fəðɪ ɪvˈɛnt kan biː səbmˈɪtɪd	diː ˈanmˌɛldʊŋ tsuːɾ fɛrˈanʃtˌaltʊŋ kˌan fˈoːɾgənˌɔmən vˌɛɾdən

Table 1: Example (source, target) pair for training of the reference and phoneme model

was used. This framework is also used by the previously stated TTS models Tacotron 2 and FastSpeech 2.

3.2 Dataset

We performed a German-English news translation task, using the WMT17 dataset from the official website. The validation phase utilized the newstest2016 files for assessing performance[3].
For the development of our Phoneme Model, both the English source and the German target components of the WMT17 dataset were transformed into their phonetic equivalents (see examples in Table 1).

3.3 Measures

The BLEU score, introduced by Papineni, Roukos, Ward, and Zhu (2001), serves as a conventional tool for evaluating the quality of language translations. But this method is not suitable for comparing sentences written in phoneme format as words may exhibit acoustic similarity and multiple similar phonemic representations, as illustrated in Table 2. It is non-trivial and beyond the scope of this research to automatically create a standardized phonemic representation or match reliably similar phonemic representations. We decided to manually convert the output of the Phoneme Model to Standard German for evaluation (as shown in Figure 2). That is, we converted 100 sentences with a total amount of 2155 words from phonemic representation back to standard German, which are publicly available for reproducibility at https://github.com/fungus75/Phonemes_S2S_Pipeline . Due to the absence of phonemic representations for punctuation marks such as commas and full stops, these characters were eliminated prior to validation.

4 Experimental Setup

In our study, we juxtapose our phoneme-based methodology with a reference translation that adheres to the unaltered WMT17 recipe[4]. The configuration of this comparison is depicted in Figure 2. Initially, we trained the reference model and computed its BLEU score. Subsequently, we employed the identical recipe to train the Phoneme Model, albeit exclusively utilizing the phoneme dataset. Both models were trained from scratch on the same hardware (a Nvidia 3090 GPU) in about 24 hours. No modifications nor training parameter adjustments to the default WMT17 recipe of the Eole model were made.

5 Results

Table 3 presents a comparison between the two models. Both models are of comparable quality as measured by the BLEU score. A higher score correlates with a better translation outcome. Lavie (2011) posits that a BLEU score of 30 and above signifies that a translation is "understandable".

Model	BLEU score
Reference Model	39
Phoneme Model	38

Table 3: Benchmarking both models

	Word 1	Word 2	Word 3
Reference	ʊmzɛtsʊŋ	nɪçt	yːbɜ
Phoneme Model Output	ˈʊmzˌɛtsʊnɡ	nˈɪçt or nˈiçt	ˌyːbɜ or ˌyːba
Standard German	Umsetzung	nicht	über

Table 2: Different phonemic variants.

[3] https://www.statmt.org/wmt17/translation-task.html

[4] https://github.com/eole-nlp/eole/tree/main/recipes/wmt17

We observed that both models tend to exhibit enhanced performance on shorter sentences, as evidenced by higher scores. For instance, the sentence "She and her mother were absolutely best friends" was translated by both models perfectly as "Sie und ihre Mutter waren absolut beste Freunde". Shorter sentences might be more common and simpler in their structure facilitating translation.

Both models generally demonstrate a high proficiency in identifying personal names (e.g., Obama, Amy, Lynn Buford...); however, the models tend to lack semantic understanding of text. For instance, (English) names like “Miller” and “Tailor” are typically not translated to “Müller” and “Schneider”. But we observed that "Professor Lamb" was frequently mistranslated as "Professor Lamm" (the German equivalent of "lamb"). Another example demonstrating that models lack semantic understanding is the following: Models exhibit difficulties in handling sentences with verbs that possess multiple meanings, particularly in brief sentences where context is limited. A notable example is the translation of "Obama receives Netanyahu", which was inaccurately rendered as "Obama erhält Netanjahu" interpreting "receives" as "get" rather than the intended "meets with". This tendency to default to the most common meaning of ambiguous verbs highlights a limitation in the models' contextual understanding. But because this issue was shown on both models – Reference and Phoneme – the limitation must be caused by the used Eole model rather than in the phoneme approach.

Comparative analysis of the outputs from the Reference Model and the Phoneme Model reveals distinct phrasing tendencies. For instance, the Reference Model output "Studenten sagten, sie würden sich auf seine Klasse freuen" contrasts with the Phoneme Model's prediction of "Studenten sagten, sie freuen sich auf seine Klasse." This observation suggests that the Phoneme Model may have a propensity to generate sentences in the present tense rather than the conditional mood.

6 Conclusion

This research illustrates that the development of the Eole sequence-to-sequence (seq2seq) model from scratch, utilizing a phonemic dataset, produces results that are on par with those achieved through the use of a traditional text dataset. Nonetheless, the adoption of a phonemic methodology provides several unique benefits:

1. Generating a phonemic representation of spoken language is simpler than producing standardized, official text, making phonemes particularly advantageous for under-resourced languages that may lack such standardized forms.
2. Specifically for Text-to-Speech (TTS) tasks, working directly with phonemes eliminates the necessity of translating standard text into phonemic representations. This benefit is not only crucial for under-resourced languages, which may lack established rules for such conversion, but it also streamlines any pipeline that transforms text into speech. Minimizing processing steps leads to reduced computation time, quicker outcomes, and diminished latency.

In conclusion, the integration of phonemic representations across the entire translation pipeline, not solely at the input stage, has the potential to yield outcomes of comparable quality in terms of BLEU score to those obtained through conventional methodologies. In particular, the present study concentrates on the exploration of the Eole seq2seq model, acknowledging its inherent constraints such as limited semantic understanding impacting both the reference and phoneme-based translation. Thus, our work can be seen as providing first evidence towards the benefits of phoneme-based translation. To establish the applicability of the proposed approach more broadly, subsequent research involving alternative models and more datasets is warranted. Furthermore, the application of XAI techniques might be beneficial to understand model behavior and possible strategies for improvements in more depth (Schneider, 2024). Leveraging strategies common in deep learning and LLMs like reflection or generation and refinement might also further improve model outcomes (Schneider, 2025; Schneider & Vlachos, 2024). Nonetheless, the findings of this investigation provide encouraging indications of the efficacy of the approach. On a larger scale, our work can contribute towards sustainable AI by reducing data needs (Schneider, Seidel, Basalla, & vom Brocke, 2023).

Ethical Considerations and Limitations

To our knowledge, this research does not encompass any ethical concerns. We used the well-established WMT17 dataset and trained the Eole

model from scratch, thus avoiding any privacy concerns.

A primary limitation identified during this study is the requirement for sequence-to-sequence (seq2seq) models to accommodate Unicode characters, a necessity stemming from the extensive employment of these characters in the representation of phonemes.

References

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2022). Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning. In M. Melero, S. Sakti, & C. Soria (Eds.), *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages* (pp. 16–22). European Language Resources Association.

Gupta, M., & Kumar, P. (2021). Robust neural language translation model formulation using Seq2seq approach. *Fusion: Practice and Applications*, *5*(2), 61–67.

Jiang, J., Ahmed, Z., Carson-Berndsen, J., Cahill, P., & Way, A. (2011). Phonetic representation-based speech translation. In Machine Translation Summit (Ed.), *Proceedings of Machine Translation Summit XIII: Papers*.

Kim, J., Kong, J., & Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. arXiv. https://doi.org/10.48550/arXiv.2106.06103

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In M. Bansal & H. Ji (Eds.), *Proceedings of ACL 2017, System Demonstrations* (pp. 67–72). Association for Computational Linguistics.

Lavie, A. (2011). Evaluating the Output of Machine Translation Systems. In *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., & Zhou, M. (2018). *Neural Speech Synthesis with Transformer Network*. arXiv. https://doi.org/10.48550/arXiv.1809.08895

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., & Mesgarani, N. (2023). *StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models*. arXiv. https://doi.org/10.48550/arXiv.2306.07691

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. In P. Isabelle (Ed.), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (p. 311). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*. arXiv. https://doi.org/10.48550/arXiv.2006.04558

Schneider, J. (2024). Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, *57*(11), 289.

Schneider, J. (2025). Improving Next Tokens via Second-Last Predictions with Generate and Refine. *Proceedings of Symposium on Intelligent Data Analysis(IDA)*.

Schneider, J., Seidel, S., Basalla, M., & vom Brocke, J. (2023). Reuse, Reduce, Support: Design Principles for Green Data Mining. *Business & Information Systems Engineering*, *65*(1).

Schneider, J., & Vlachos, M. (2024). Reflective-net: Learning from explanations. *Data Mining and Knowledge Discovery*, *38*(5), 2975–2996.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., . . . Wu, Y. (2017). *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. arXiv. https://doi.org/10.48550/arXiv.1712.05884

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in Machine Translation. *Engineering*, *18*, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

`embed2discover`: the NLP Tool for Human-In-The-Loop, Dictionary-Based Content Analysis

Oleg Bakhteev[1,3], Luis Salamanca[2,3], Laurence Brandenberger [4], Sophia Schlosser[4],

[1]EPFL, Switzerland [2]ETHZ, Switzerland
[3]Swiss Data Science Center, Switzerland [4]University of Zurich, Switzerland
Correspondence: oleg.bakhteev@epfl.ch

Abstract

Guided dictionary-based content analysis has emerged as an effective way to process large-scale text corpora. However, the reproducibility of these analysis efforts is often not guaranteed. We propose a human-in-the-loop approach to dictionary-based content analysis, where users get control over the training pipeline by chunking the process into four distinct steps. Compared to end-to-end and/or purely LLM-based approaches, where the learning and inference process is difficult to understand and, hence, to steer, we advocate for a human-in-the-loop methodology. We demonstrate how, through minimal labeling and intervention, the user can guide the process and achieve competitive performance. [1]

1 Introduction

The use of machine learning (ML) and (large) language models for the content analysis of text-based data has grown in popularity (e.g., Ampel et al., 2025; Grimmer et al., 2021; Zhao et al., 2025; Xiao et al., 2023; Kroon et al., 2024), but researchers are often wary of employing such methods for fear of retrieving unreliable information (Jordan et al., 2023; Chatsiou and Mikhaylov, 2020; Wilkerson and Casas, 2017), in some cases without any information on the model's true performance to assess its validity. This could be quite hindering, demeaning the validity of these approaches, especially when analysing text sources from really specialized domains. In this paper, we introduce `embed2discover`, a tool for dictionary-based, supervised content analysis of (large-scale) text data. Our tool *assists* human coders (henceforth called 'users') in discovering topics and themes in (large) text corpora and classifying text excerpts by combining methodologies from natural language processing (NLP) with language models and human annotations.

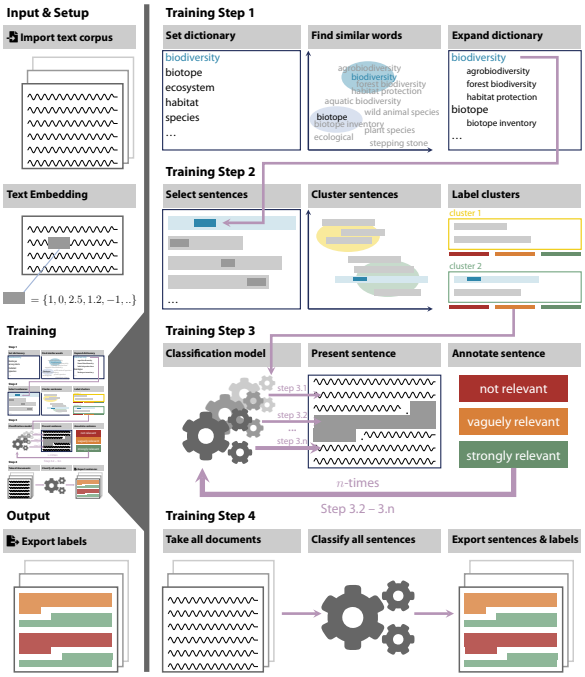


Figure 1: Overview over `embed2discover`.

With advances in text processing and increased archival retrieval efforts, text-based data sources have become increasingly popular. These data sources are rich in information and offer valuable insights into both the author's intent and the broader context in which the text was created. This surge in text availability has driven the development of new tools and methodologies for efficient processing, organizing, cleaning, and classifying large-scale text data. A common approach in this domain is content analysis, which enables researchers to identify patterns and commonalities within text-based data. For instance, scholars have used content analysis to examine populist elements in political speeches (e.g., Jagers and Walgrave, 2007) or studying polarization dynamics (e.g., Fisher et al., 2013). Traditionally, when such

[1]The code and demo can be found at `https://gitlab.datascience.ch/democrasci/embed2discover`.

classifications serve as the foundation for further research (e.g., Nussio and Clayton, 2024), extensive codebooks have guided human annotators or coders in categorizing text passages. However, as the volume of text data continues to grow, manual annotation reaches its limits, often necessitating the use of computer-assisted or fully automated content analysis. Despite these advancements, concerns remain regarding the reliability and replicability of automated content analysis, particularly when (large) language models are involved (e.g., Nelson et al., 2021; Kitto et al., 2023). The issue of replicability is especially challenging, as model-driven classifications can be difficult to reproduce without full transparency in their underlying processes. Besides, the performance can substantially vary between different domains.

We propose addressing concerns about reliability and replicability by integrating computer-assisted methods with a human-in-the-loop approach. Our goal is to provide scholars with a tool for assisted content analysis, ensuring that each step remains subject to human judgment. `embed2discover` is designed to combine the strengths of both worlds: leveraging advanced NLP methods while incorporating efficient human annotations. The tool enables users to hand-label meaningful sentences within text data through a user-friendly interface, progressively training a model to identify relevant sentences (embedded in paragraphs) with increasing accuracy. Fig. 1 provides an overview of `embed2discover`. First, the user defines a set of keywords related to the topic of interest to guide the content analysis. The tool then facilitates the training of a classification model using active learning. At each step, the user evaluates progress through relevant metrics, guiding the content analysis in the direction desired by the user. The output of `embed2discover` can then serve as the foundation for downstream analyses, uncovering insights into both the originator and the broader context of the text source under investigation.

2 Related Work

2.1 Dictionary-based Content Analysis

Traditionally, content analysis is performed as an expert-guided, human annotation process, where researchers devise (elaborate) coding schemes and then proceed to process text data manually and code the read text according to the schemes.[2] To speed up the hand-annotation process, computer-assisted content analysis was developed. Early approaches focus on automated content classifications (e.g., Andersen et al., 1992; Carley, 1994; Cowie and Lehnert, 1996), computer-assisted identification of grammatical patterns (e.g., Franzosi et al., 2012), or topic extraction (e.g., Lee and Kim, 2008). The promise of automated or semi-automated content analysis is increased efficacy, allowing researchers to either broaden or deepen their analysis through the use of expanding data sources (Laurer et al., 2024; Chatsiou and Mikhaylov, 2020). But fears potential users have to employ these techniques revolve around replicability, validity, and reliability of the coded results (Jordan et al., 2023; Baden et al., 2022; Muddiman et al., 2019).

The concern with using automated content analysis is that both supervised and unsupervised methods usually classify text into predefined categories, either using a dictionary (common in unsupervised approaches) or hand-annotated texts (i.e., sentences, paragraphs, or documents) (Wilkerson and Casas, 2017). Whereas dictionary approaches have considerably sped up the annotation process, they are also heavily biased (Carley, 1990; Vourvachis and Woodward, 2015; Van Atteveldt et al., 2021). The biggest issue resides in the fact that dictionaries are fixed words (or n-grams) that do not account for (i) linguistic flexibility, (ii) linguistic changes over time, and (iii) translation biases (Van Atteveldt et al., 2021). For supervised methods, the user has to set up a classification model and feed it with hand-annotated texts and allow a model to learn distinguishing characteristics from the text (i.e., existence of words, n-grams, linguistic structures). Then, the supervised machine learning (SML) models generally assign weights to these distinguishing characteristics and, given enough training data, can assign categories to new texts based on the content and the learned weights (for applications, see Hanna, 2013; King et al., 2013).

2.2 Known Shortcomings of Current Content Analysis Tools

Several drawbacks make researchers weary of applying these unsupervised and supervised models to classify text: (1) Coding schemes based on dic-

[2]Note that human annotation has been criticized in the literature, especially when it comes to defining human annotations as 'gold standards' and 'ground truth datasets' (e.g., Song et al., 2020; Mikhaylov et al., 2012).

tionaries limit the coded texts linguistically, do not account for word changes over time (if temporal data is used, see (Greene et al., 2019)), and restrict the found texts. This is particularly problematic for concepts that are fuzzy in nature or have ill-defined boundaries, such as populism, inequality, or biodiversity. (2) For supervised approaches, the researcher does not know apriori how many labeled texts it has to provide the SML in order to achieve a high enough classification score, especially in really unbalanced datasets. This makes the use of SML methods less desirable, as it strengthens the idea that these methods are unreliable and constitute a 'black box'. The latter argument holds especially true for generative LLMs (Huang et al., 2025). (3) For supervised approaches, the researcher has to define a clear coding scheme to provide the labeled texts. Drafting these coding schemes entails a lot of work.

3 Backend: The Mechanics Behind `embed2discover`

3.1 Designing goals

The primary goal of `embed2discover` is to offer a simple, reliable, and interpretable tool for content analysis. The core design principles of our system can be summarized as follows:

- **Interpretability**: The toolbox provides a pipeline that takes, as its initial input, only a small set of seed words and outputs a fully annotated corpus of sentences. To ensure transparency and traceability, we prioritize simple yet interpretable methods over more sophisticated ones. Each step includes visualizations and statistical summaries, helping users understand and evaluate the pipeline's performance.
- **Efficiency**: Our focus is on building a tool that runs smoothly on any modern computer without excessive resource consumption. Since our toolbox requires intensive user interaction, all steps must run efficiently, even for large corpora. To achieve this, we prioritize lightweight models whenever possible. We precompute all word and sentence embeddings for the given corpus, allowing users to run the toolbox without a GPU after embeddings are computed. Additionally, we use shallow classifiers as a baseline instead of deep learning models and avoid language model fine-tuning.
- **Configurability and extensibility**: The toolbox is designed to be highly configurable. Each step has its own configuration, represented by a YAML file. Users can modify configurations for specific projects and training sessions via the frontend application without altering global settings or restarting the application. The toolbox components—including nearest neighbor retrieval, classification, clustering methods, and active learning strategies—can be implemented externally and integrated via configuration. This flexibility allows users to extend the toolbox without needing to inspect or modify its source code.

3.2 Input: The Text Corpus and Embeddings

User experiments in the toolbox are organized into projects, where each project is associated with a corpus and corresponding word and sentence embeddings. The toolbox also allows multiple parallel experiments using the same corpus, with each experiment contained within a separate project, maintaining its own configuration and results. The text corpus is structured as non-formatted text files, with each document stored in an individual file and assigned a unique ID. For text preprocessing, i.e. word and sentence tokenization, we employ the `Spacy` library (Honnibal et al., 2020). To support multilingual corpora, we employ the `cld2` [3] language detector to identify document language, which guides language-specific preprocessing and is also used as input for language-aware embeddings. Since the initial training steps operate at the word level, obtaining high-quality word embeddings is crucial. In many cases, it is beneficial to consider not only distinct word embeddings but also phrase embeddings that account for short, frequent phrases. For phrase extraction, we use the `Gensim` library (Řehůřek and Sojka, 2011). The phrase extraction parameters, including the maximum n-gram length and threshold, can be configured during corpus upload.

The functionality of the toolbox is based on word and sentence embeddings (Sahlgren, 2008). The toolbox supports averaged *word embeddings* (Bommasani et al., 2020) derived from contextualized sentence embedding models such as `BERT` (Devlin et al., 2018). Regarding *sentence embeddings*, the toolbox relies on the `Sentence-BERT` (Reimers and

[3] `https://github.com/CLD2Owners/cld2` For the Python version of the library, we use the package from `https://github.com/GregBowyer/cld2-cffi`.

Gurevych, 2019) library. The tool allows caching embeddings for the target corpus, facilitating the handling of large-scale document corpora without wasting significant computational resources.

3.3 The Four Training Steps

The training process is divided into four steps (see Fig. 1):

1. **Dictionary Expansion**: Expands the initial dictionary with additional domain-related words and phrases.
2. **Coarse Classification**: Extracts sentences containing dictionary words and clusters them by semantic similarity. The user labels selected clusters to create an initial dataset annotation.
3. **Refined Classification**: Trains the model iteratively using active learning, where the user labels sentences tracking the increase of model performance.
4. **Full Classification**: Trains a final model based on previous labels and classifies all sentences in the corpus.

For all the steps, we calibrate the algorithm's hyperparameters (including k in KNN algorithm, classification model parameters, number of clusters in K-means algorithm) automatically using `Optuna` library (Akiba et al., 2019). Steps one, three, and four involve training a classification model. For the classification model, we mainly consider kernel logistic regression from (Pedregosa et al., 2011), but any other classification model can be used. We also perform model confidence calibration to make the model confidence aligned with class probabilities. This is important both from the perspective of confidence interpretability and for the active learning step 3, which utilizes the model confidence to filter new sentences to annotate.

For the **Dictionary expansion** step, the user provides a set of initial keywords (i.e., a dictionary). To expand the initial dictionary list with semantically similar words, we use a binary classification model. It uses the word embeddings, treating the dictionary and neighboring words as positive examples, and randomly sampled words as negative ones. These are likely to be irrelevant given the size of the full corpus vocabulary. The final selection threshold is determined via cross-validation. This step also supports both individual words and extracted phrases, as described in Section 3.2. The user can control the expansion of the dictionary via two key access points: (1) the model classification probability threshold, controlling the confidence of the model to add the words into the dictionary, and (2) the relative frequency threshold, allowing to discard too frequent words from the dictionary. In practice, the user can select the confidence threshold using cross-validation; the results of it are provided during the step run.

In the **Coarse Classification** step, we identify all sentences containing words from the expanded dictionary, compute their embeddings, and cluster them using K-means. Semantically similar sentences from user-selected clusters are then used as the initial input for building the classification model. The user then assigns labels to clusters as *strongly relevant*, *vaguely relevant*, or *not relevant*, see Fig. 6, Section A.1 in the Appendix. To improve interpretability, we visualize clusters by displaying centroid embeddings in a 3D space using Multi-Dimensional Scaling (MDS) (Borg and Groenen, 2007), together with cluster homogeneity and size (see Fig. 5, Section A.1 in the Appendix).

In **Refine classification**, the user is provided with a set of sentences to label. After completing one labeling round, the classification is updated. By choosing a rather simple model, we retrain the model from scratch after every round. We follow standard approaches to active learning in selecting the sentences for annotation. By default, we use the following active learning strategy: (1) We estimate a best classification model probability threshold by F1-score using a cross-validation procedure. (2) We take a pool of sentences with confidence higher than the obtained threshold. (3) From this pool, we select sentences with the highest confidence, sentences with the lowest confidence, and sentences randomly sampled from the pool. We believe that this strategy allows users to rapidly gather a representative annotated dataset. For later stage iterations, the user can move to an active learning strategy that promotes the annotation of less confident sentences (Li and Sethi, 2006), use the strategy described before, or switch between the two. Note that, similar to other components, the user can extend `embed2discover` with their own strategies.

In **Full classification**, the user can apply the classification to each sentence in the corpus. At this step, the model is retrained from scratch using the labels from the active learning step.

All steps in the pipeline support multilingual

embeddings, allowing users to work with corpora in multiple languages simultaneously. Here, we leverage the properties of multilingual embeddings: for the **Dictionary expansion step**, we search for neighbouring words with similar embeddings not only in the target language but across all languages selected by the user. The **Coarse classification step** uses the expanded dictionary to retrieve relevant sentences, thereby naturally supporting multilingual settings. For the **Refined classification**, the user can annotate sentences in selected languages and then generalize the model to all languages present in the corpus during the **Full classification step**, again relying on the cross-lingual capabilities of the embeddings.

3.4 Output: Fully-Labeled Corpus and Classification Evaluations

`embed2discover` generates a classification output at the sentence level. While the text corpus can be imported as separate files (e.g., representing distinct documents), the tool maintains sentence-level classification for two key reasons: (1) **Transparency**: Users can independently determine how to classify a document based on the number or proportion of positively identified sentences. (2) **Flexibility**: By providing model probability scores, the sentence-level output allows for further refinement in subsequent analyses, if necessary.

In addition to the main output, `embed2discover` generates output data at each intermediate step, also capturing: the dictionary, the coarse classification, and the annotated sentences. This facilitates replication and enables further extensions and future research. Moreover, each stage of the process is accompanied by performance visualizations, including: (1) hyperparameter optimization for all the steps, (2) precision-recall evaluation, F_1 and $F_{0.5}$ scores dependency from the threshold for the dictionary expansion, refine classification and full classification steps, (3) annotation progress tracking for the refine classification step. These features ensure both interpretability and transparency, making `embed2discover` a robust and trustable tool for supervised content analysis.

4 Frontend: The Design of the Web-Application

The frontend web interface is built using Flask [4] framework and implements three main views that support corpus handling, content analysis, and documentation and system settings. The communication with the backend is performed using REST API, which allows the application of different scenarios of the toolbox usage. Overall, the frontend has the following pages:

1. The *Corpora* page (Fig. 2) allows users to upload their own corpus or work with an existing, openly-published corpus.
2. The *Embedding* page allows users to precompute word and sentence embeddings for the given corpus, required in subsequent steps.
3. The *Project* page allows users to modify existing projects or create new ones. Project-setting includes the correspondence between the used corpus and used embeddings. The project page also allows to change the configuration YAML file for the specific training step.
4. The *Training* pages correspond to the four training steps: dictionary expansion, coarse classification, refined classification, and full processing. Each page allows the user to select a project, adjust the configuration of each step, run new experiments and look at previously obtained results.
5. The *System* page shows logs both for the backend and frontend and also allows users to kill the training step currently running.
6. The *Documentation* page holds the documentation of the tool, including instructions on how to use the WebApp, or download the toolbox to use it offline and allow a further customization.

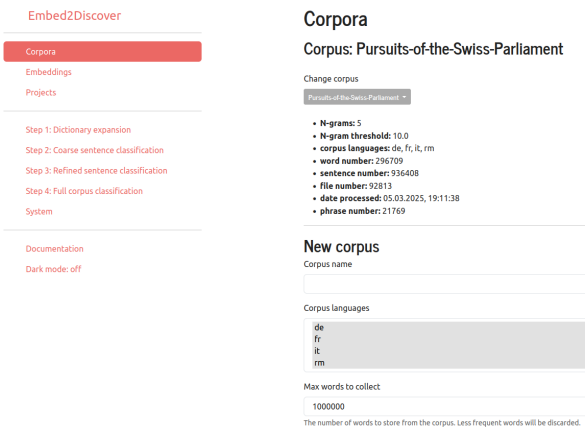


Figure 2: `embed2discover` frontend: a corpus page

[4] https://flask.palletsprojects.com/en/stable/

5 Usage and Evaluation

In this section, we conduct a human study to evaluate `embed2discover`. First, we assess its recall against 3,159 expertly annotated documents. Second, we leverage `embed2discover`'s ability to label large-scale corpora more efficiently than expert annotations and examine the quality of its coding outputs. For all the experiments we use SwissBERT language model (Vamvas et al., 2023) fine-tuned on the text data from the Swiss Parliament (Salamanca et al., 2024).

5.1 Testcase: Pursuits of the Swiss Parliament

Dataset description In order to evaluate the performance of `embed2discover`, we use text data from the Swiss Parliament (Salamanca et al., 2024). The text corresponds to submitted parliamentary pursuits, including interpellations, questions, motions, postulates, initiatives, and federal drafts. We use the full text of the submitted pursuits, totaling $N = 94,404$ documents. To evaluate the performance of `embed2discover`, we train a model to identify parliamentary pursuits on the topic of public service broadcasting. We detail results on parliamentary activity on public service broadcasting in Switzerland between 1891 and 2024 in Section A.3 in the Appendix.

Ground-truth annotation To assess the true recall capability of `embed2discover`, we annotated 3,159 documents related to the broader topic of 'technology and communication' through expert annotation. Each of the 3,159 texts was assigned a TRUE/FALSE label, indicating whether the document's main topic concerned the regulation of Swiss public service broadcasting. The expert annotations were conducted by two political scientists. We assessed their intercoder-reliability by comparing 500 documents. Their agreement rate was 98.5%, with a Cohen's Kappa of 0.968.

Texts for which experts were uncertain were excluded from the analysis. Since the experts annotated the bill texts independently of the annotation performed using `embed2discover`, some bill texts contained sentences that had already been annotated in the toolbox. These bills were also excluded from the evaluation. The final evaluation dataset consisted of $2,872$ texts, all written in German. The texts contain between 1 and 132 sentences, with a median of 6 sentences. The proportion of relevant bill texts in the corpus is 0.28.

5.2 Assessing Annotation Progress

We begin with the following German words as our base dictionary: 'Rundfunk' 'öffentlich rechtlicher Rundfunk' 'SRG' 'RTS' 'SRF' 'RSI' 'RTR' 'Radio und Fernsehen' 'Lokalradio' 'Regionalfernsehen' 'Regional-TV' 'Service-public-Auftrag' 'Schweizer Fernsehen SF' 'Schweizer Radio DRS' 'Fernsehprogramm' 'Fernsehkanal' 'Fernsehsendung' 'Sendekonzession' 'Rundfunkabgabe' 'Serafe' 'Billag'.

We expand the dictionary using the default settings, resulting in 204 additional chosen words, such as 'Staatssender', 'Fernsehempfangsgebühren', 'Spartenprogramm', 'Programmproduktion' or 'Beromünster'. In the second step, we evaluate 57 clusters, 12 of which are labeled as highly relevant. We then perform 120 active learning steps, 38 of which use the least-confidence sampling strategy.

During the refine classification step (step 3), the user can evaluate their progress using three distinct metrics and plots.

1. *The progress graph* (Fig. 3a) tracks the percentage of sentences labeled over the course of active learning steps. The fluctuating nature of the curves reflects the model's sampling strategy, where more uncertain or certain samples are prioritized for annotation.

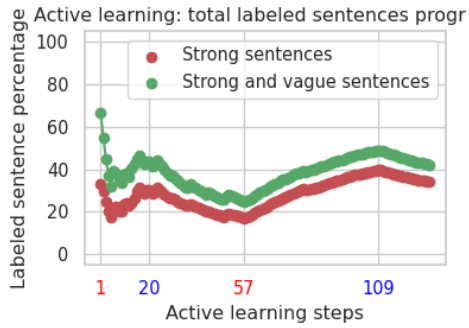
2. *The precision-recall curve* (Fig. 3b) illustrates the classifier's trade-off between precision and recall across different threshold levels, averaged over 20-fold cross-validation. The high precision-recall score suggests that the model is effectively capturing relevant annotations while minimizing false positives.

3. *The F1 curve*, (Fig. 4, Section A.1 in the Appendix) shows how the F1 score—balancing precision and recall—varies across classification probability thresholds, with the best F1 score highlighted.

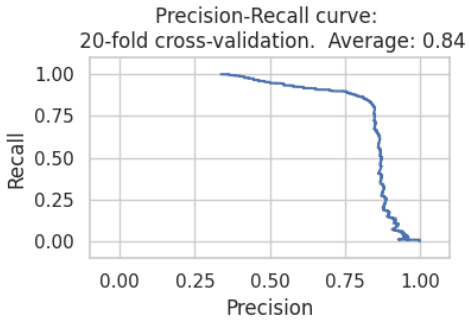
These evaluations help the user assess annotation progress and determine whether to proceed to step 4.

5.3 Performance evaluation

To assess the efficiency of `embed2discover`, we compare our model against multiple baselines:



(a) Progress graph: Steps using the confident active label strategy are shown in red, while steps using the least confident strategy are shown in blue.



(b) Precision-Recall graph for the last annotation step.

Figure 3: Evaluation of Annotation Progress

- **SML**: We performed a 20-fold cross-validation on the ground-truth dataset to evaluate how well the model would perform if a portion of the dataset were annotated without the assistance of `embed2discover`. In this setting, we used the same embedding model as in `embed2discover` but treated the classification as a binary problem, in contrast to `embed2discover`, which employs a three-class classification scheme. Apart from this difference, all other steps, including hyperparameter optimization, were conducted in the same manner as in `embed2discover`. Since we use the same number of annotated labels here, we can estimate the impact of active learning performed by `embed2discover`.

- **LLM-d**: LLM request using our initial dictionary.

- **LLM-z**: LLM in a zero-shot mode (Brown et al., 2020). In this mode, we described the classification task to the LLM without providing any examples.

- **LLM-o**: LLM in a one-shot mode (Brown et al., 2020). In this setting, we provided the LLM with one example of pursuit texts for each class and asked it to classify a new text.

For the LLM-based baselines, we employed Llama-3.1-8B-Instruct (Grattafiori et al., 2024). We intentionally used a relatively small LLM to ensure fair comparisons across models given the same hardware constraints. Specifically, we used a GPU with 16GB of memory, matching the hardware used for embedding computation in `embed2discover`. For some texts, the LLM failed to provide responses due to hallucinations or out-of-memory issues. In such cases, we used ground-truth labels, which may have slightly overestimated the performance of the LLM-based baselines. A complete list of LLM prompts can be found in Section A.2 in the Appendix.

In addition to the considered baselines, we evaluated multiple modes of `embed2discover` usage:

- **coarse**: In the coarse classification mode, we used only the data obtained during the initial coarse classification step without any active learning iterations.

- **final-b**: In the final binary classification, we utilized all annotations from `embed2discover` for the 'non-relevant' and 'strongly relevant' classes to train a binary classification model.

- **final**: Full classification model with all three classes. The model is trained in a three-class classification setting, the confidence scores of the non-relevant and vaguely relevant classes are combined into a single non-relevant category. We then apply thresholds to distinguish only between 'non-relevant' and 'relevant' sentences, which allows to match the predictions with the annotated gold standard.

For all models, we computed precision, recall, F1-score, and the approximate time required for manual annotation. The results are presented in Table 1.

Our model achieves a high recall of 0.89, successfully identifying the vast majority of relevant documents. A high recall is particularly important in our case, as the primary objective is to maximize the retrieval of relevant texts concerning the regulation of Swiss public service broadcasting. While the precision is lower (0.77), indicating that about

Model	N	Time, hours	Prec.	Rec.	F1
SML	20	0.2	0.32	0.82	0.45
SML	1191	11	0.65	0.74	0.69
LLM-d	0	0	0.29	0.79	0.42
LLM-z	0	0	0.28	**0.91**	0.43
LLM-o	-	-	0.32	0.73	0.44
`embed2discover`					
coarse	20	1.2	0.6	0.89	0.71
final-b	1191	3.6	**0.77**	0.89	**0.82**
final	1294	3.6	0.73	0.89	0.81

Table 1: Experiment results. The results for SML and `embed2discover` models are averaged over 20 runs. LMM-d refers to LLM with dictionary; LLM-z refers to LLM zero-shot; LLM-o refers to LLM one-shot; final-b refers to final binary. The "Time" column represents the approximate time required for manual annotation using each model. The "N" column indicates the number of manually annotated items. We do not report the time and number of manually annotated labels for LLM-o, as they depend on the selection of in-context examples and the example selection strategy.

212 out of 922 retrieved documents are false positives – this remains an acceptable trade-off within our research context. At step 70, precision rates were 0.65, while recall remained at around 0.8, indicating that with longer training, a higher precision can be achieved through further annotation. Since our approach prioritizes recall over precision, a certain level of false positives is tolerable, as they can be efficiently filtered during manual post-processing.

Compared to LLM-based methods, we observe that our approach achieves significantly better precision. While LLMs may yield better results with postprocessing, prompt tuning, or more sophisticated methods than one-shot learning, our method is more reliable because we monitor the behavior of the models used at each step. The results also show that the proposed approach is performing well in comparison to the SML, which used the same amount of data for training, and to the LLM-based method.

5.4 Time and Efficiency

The manual annotation of the $N = 3,159$ documents took 28 hours. This excludes the preparation of the codebook and the pre-discussions and only entails labeling work. Contrarily, the human annotation process using `embed2discover` for the $1,294$ texts took 3.6 hours, allowing us to automatically annotate the entire corpus afterward. The human annotation of the clustering step (57 clusters) took 70 minutes. The annotation of ten sentences per active learning step took 1-3 minutes per step, amounting to 147 minutes of human annotation. The computational processing time on `embed2discover` amounts to 45h. The dictionary expansion step took 4.5 minutes, the coarse classification step took 2.5 hours, and the active learning steps took between 5 and 40 minutes to update the model and locate new sentences for the user to label.

The toolbox was run on a server with 16 GB of RAM and an 8-core CPU. Corpus preprocessing and embedding computation took 6.5 hours, where a 16 GB GPU was used for the latter.

6 Conclusion

We present `embed2discover`, a human-in-the-loop, automated content analysis tool that enables users to classify text data based on a dictionary-driven approach. Our toolbox is designed to be fast, efficient, and **replicable**. Besides, it provides a larger level of interpretability by giving back the control to the domain expert.

One of the biggest challenges social scientists face when performing dictionary-based content analysis using computational approaches is the inability to fully replicate the procedure due to the black-box nature of most tools. While such tools offer rapid annotation and scalability for large-scale corpora, the labeled data often lacks reliability for downstream tasks.

`embed2discover` takes a different approach. By breaking down the annotation process into four distinct steps and employing lean computational methods, the user maintains full control over the process and can systematically evaluate the performance of each step. We demonstrate the effectiveness of our tool by classifying real-world documents from the Swiss Parliament.

Future work includes integrating additional active learning strategies, further optimizing computational efficiency, and expanding user support for argument mining and stance detection. We believe that `embed2discover` will serve as a valuable tool for researchers seeking a transparent and interpretable approach to automated text classification.

Limitations

While `embed2discover` offers a transparent and replicable approach to dictionary-based content analysis, certain limitations remain.

Replicability vs. Advanced Classification Models. Unlike black-box (large) language models (LLMs), our toolbox prioritizes simple, interpretable classification methods to ensure replicability and user control. More advanced classification approaches—such as transformer-based models or GPT-based classifiers—could potentially enhance classification accuracy. However, these methods often sacrifice transparency, making it difficult for users to trace and reproduce results. Besides, they can highly underperform when the text comes from really specific domains. To balance flexibility with control, `embed2discover` allows users to swap out classification models via configuration settings. We encourage users to explore these options when prioritizing speed or resource efficiency over strict replicability. The toolbox configurability will also enable it to keep up with emerging technologies by allowing the integration of more powerful embedding approaches.

Post-Annotation Analysis. `embed2discover` does not provide currently built-in functionality for post-annotation analysis. Users must export labeled data for further processing in external tools. Future versions of the toolbox may include integrated support for common post-annotation tasks, such as summary statistics, validation checks, and visual analytics.

Multi-User Annotations and Inter-Annotator Agreement. At present, `embed2discover` is designed for individual use and does not natively support multi-user annotation workflows. Additionally, we do not currently provide inter-annotator agreement (IAA) measures to assess the consistency of multi-user annotations. Extending the toolbox to allow collaborative annotation and incorporating IAA metrics would be valuable enhancements for future development.

Acknowledgments

The authors would like to thankfully acknowledge the funding through the SDSC Project C21-06 "EvolvingDemocrasci".

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Benjamin Ampel, Chi-Heng Yang, James Hu, and Hsinchun Chen. 2025. Large language models for conducting advanced text analytics information systems research. *ACM Transactions on Management Information Systems*, 16(1):1–27.

Peggy M Andersen, Philip J Hayes, Steven P Weinstein, Alison K Huettner, Linda M Schmandt, and Irene Nirenburg. 1992. Automatic extraction of facts from press releases to generate news stories. In *Third conference on applied natural language processing*, pages 170–177.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Ingwer Borg and Patrick JF Groenen. 2007. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

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 (NeurIPS 2020)*, 33:1877–1901.

Kathleen Carley. 1990. Content analysis. *The encyclopedia of language and linguistics*, 2:725–730.

Kathleen Carley. 1994. Extracting culture through textual analysis. *Poetics*, 22(4):291–312.

Kakia Chatsiou and Slava Jankin Mikhaylov. 2020. Deep learning for political science. *The SAGE handbook of research methods in political science and international relations*, pages 1053–1078.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dana R Fisher, Joseph Waggle, and Philip Leifeld. 2013. Where does political polarization come from? locating polarization within the us climate change debate. *American Behavioral Scientist*, 57(1):70–92.

Roberto Franzosi, Gianluca De Fazio, and Stefania Vicari. 2012. Ways of measuring agency: an application of quantitative narrative analysis to lynchings in georgia (1875–1930). *Sociological Methodology*, 42(1):1–42.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kevin T Greene, Baekkwan Park, and Michael Colaresi. 2019. Machine learning human rights and wrongs: How the successes and failures of supervised learning algorithms can inform the debate about information effects. *Political Analysis*, 27(2):223–230.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419.

Alexander Hanna. 2013. Computer-aided content analysis of digitally enabled movements. *Mobilization: An International Quarterly*, 18(4):367–388.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Jan Jagers and Stefaan Walgrave. 2007. Populism as political communication style: An empirical study of political parties' discourse in belgium. *European journal of political research*, 46(3):319–345.

Soren Jordan, Hannah L Paul, and Andrew Q Philips. 2023. How to cautiously uncover the "black box" of machine learning models for legislative scholars. *Legislative Studies Quarterly*, 48(1):165–202.

Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How censorship in china allows government criticism but silences collective expression. *American political science Review*, 107(2):326–343.

Kirsty Kitto, Catherine A Manly, Rebecca Ferguson, and Oleksandra Poquet. 2023. Towards more replicable content analysis for learning analytics. In *LAK23: 13th international learning analytics and knowledge conference*, pages 303–314.

Anne Kroon, Kasper Welbers, Damian Trilling, and Wouter van Atteveldt. 2024. Advancing automated content analysis for a new era of media effects research: The key role of transfer learning. *Communication Methods and Measures*, 18(2):142–162.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Sungjick Lee and Han-joon Kim. 2008. News keyword extraction for topic tracking. In *2008 fourth international conference on networked computing and advanced information management*, volume 2, pages 554–559. IEEE.

Mingkun Li and Ishwar K Sethi. 2006. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261.

Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

Ashley Muddiman, Shannon C McGregor, and Natalie Jomini Stroud. 2019. (re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, 36(2):214–226.

Laura K Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237.

Enzo Nussio and Govinda Clayton. 2024. Introducing the lynching in latin america (lyla) dataset. *Journal of Peace Research*, pages 1–18.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*, arXiv:1908.10084:1–11.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of linguistics*, 20:33–53.

Luis Salamanca, Laurence Brandenberger, Lilian Gasser, Sophia Schlosser, Marta Balode, Vincent Jung, Fernando Perez-Cruz, and Frank Schweitzer. 2024. Processing large-scale archival records: The case of the swiss parliamentary records. *Swiss Political Science Review*, 30(2):140–153.

Hyunjin Song, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich,

Fabienne Lind, Sebastian Galyga, and Hajo G Boomgaarden. 2020. In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4):550–572.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. Swissbert: The multilingual language model for switzerland. *ArXiv Preprint:2303.13310*, pages 1–15.

Wouter Van Atteveldt, Mariken ACG Van der Velden, and Mark Boukes. 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140.

Petros Vourvachis and Thérèse Woodward. 2015. Content analysis in social and environmental reporting research: trends and challenges. *Journal of Applied Accounting Research*, 16(2):166–195.

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(1):529–544.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *28th Inter- national Conference on Intelligent User Interfaces (IUI '23 Companion)*, pages 27–31, Sidney, NSW, Australia.

Chengshuai Zhao, Zhen Tan, Chau-Wai Wong, Xinyan Zhao, Tianlong Chen, and Huan Liu. 2025. Scale: Towards collaborative content analysis in social science with large language model agents and human intervention. *arXiv preprint arXiv:2502.10937*.

A Appendix

A.1 `embed2discover` interface

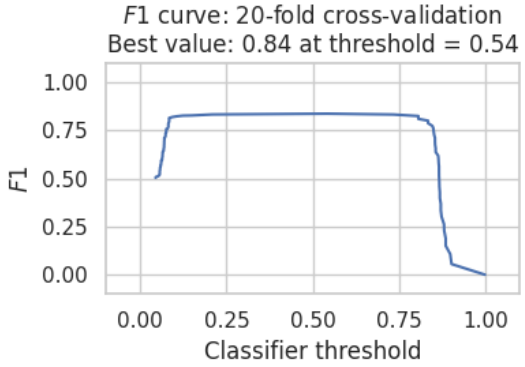


Figure 4: F1 graph

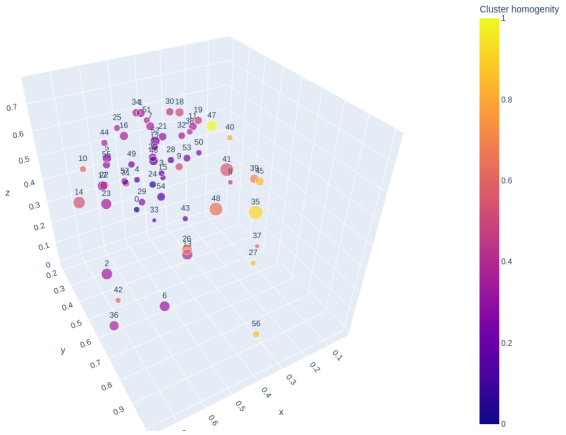


Figure 5: A visualization of the clustering results is provided. For each cluster, we project its centroid embedding into a 3D space. The size of the centroid point reflects the number of sentences in the cluster, while its color represents cluster heterogeneity, measured as the average distance between cluster points and their centroid.

Figure 6: An annotation for the refined classification step.

A.2 LLM Prompts Utilized in the Experiment

LLM request using an initial dictionary

```
    You are an AI model that
       evaluates whether a given
       text is relevant to a set
       of keywords.
Instructions:

    You will be given a set of
       keywords and a text.
    Respond with 1 if the text is
        relevant to the keywords.
    Respond with 0 if the text is
        not relevant.
    Output only 0 or 1, nothing
       else.
```

```
Input Format:

    Keywords: [{}]
    Text: "{}"

Write only 0 or 1, don't comment
   answer.
```

LLM in a zero-shot mode

```
   You are an AI model that
      determines whether a given
      text is about Public
      Service Broadcasting (PSB).

Instruction:
Classify the text as 1 if it is
   related to public service
   broadcasting, or 0 if it is
   not.


Classification Task:
- Text: "{}"

Write only 0 or 1, don't comment
   answer.
```

LLM in a one-shot mode

```
You are an AI model that
   classifies whether a given
   text belongs to category 1 or
   0 based on provided examples.

Examples:
    - Text: "{}"
    - Label: {}

    - Text: "{}"
    - Label: {}

Classification Task:
    - Text: "{}"

Write only 0 or 1, don't comment
   answer.
```

For each text to be classified, we randomly sample a pair of texts from different classes and insert them into the prompt. The order of the classes in the examples is selected randomly.

A.3 Public Service Broadcasting Debates in the Swiss Parliament

The Swiss Parliament has addressed 851 parliamentary pursuits related to public service broadcasting over the past 130 years (see Fig. 7). The first recorded pursuit on the topic dealt with the provided Sunday programs of the telephone broadcast in the year 1905. With the founding of Switzerland's first radio broadcasting company in 1931, parliamentary discussions primarily focused on regulating the distribution network. The topic gained traction in the 1970s, when debates emerged regarding the financing of government-led broadcasting studios. These discussions continued throughout the 1980s, during which test runs were conducted to allow commercial breaks during broadcasts.

In 1991, the Swiss Parliament approved the federal enactment draft on the structuring of the Swiss broadcasting system. This reform triggered further debates on financing through collection fees—first via Billag and later through Serafe—as well as discussions on content neutrality, appropriateness, and the governmental authority responsible for monitoring broadcasting standards.

In 2012, Swiss radio and television broadcasting companies were merged into a single entity, SRF, marking a significant structural change in the national broadcasting landscape.

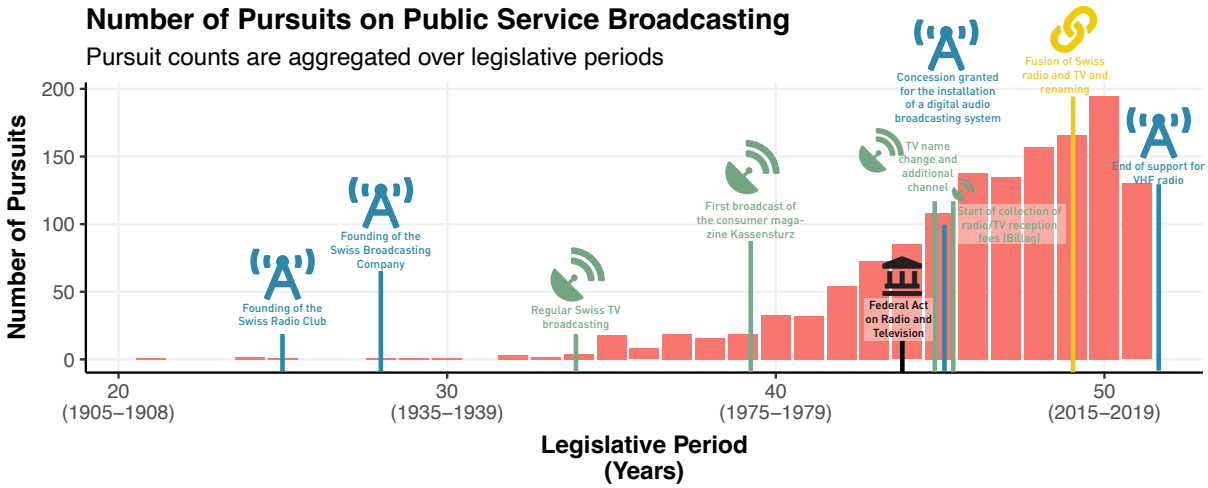


Figure 7: Distribution of public service broadcasting related pursuits in the Swiss parliament

Chapter 2

Corpus Track

SwissCoco2025 - The Swiss Corpora Collection 2025

Mark Cieliebak, Jonathan Gerber, Manuela Hürlimann

Zurich University of Applied Sciences (ZHAW) and
Swiss Association for Natural Language Processing (SwissNLP)
[ciel, gerj, hueu]@zhaw.ch

Abstract

The Swiss Corpora Collection (SwissCoco) gives an overview of language resources that are relevant for Switzerland. It focuses on text, speech, and documents. The Swiss Corpora Collection 2025 is the first iteration of the collection and contains 30 corpora. All corpora are also listed at https://swissnlp.org/swiss-corpora-collection.

1 Introduction

Language resources such as text and speech corpora are fundamental building blocks for research and applications in Natural Language Processing, Computational Linguistics, Large Language Models, Machine Learning and Artificial Intelligence in general. They provide the essential data for training and evaluating models and for conducting linguistic analyses. While a wealth of resources exists for major world languages, finding and accessing corpora for specific linguistic varieties, dialects, and smaller languages remains a significant challenge. This is particularly true for Switzerland, with its four national languages and diverse regional dialects, notably Swiss German, the predominant spoken variety in the German-speaking part of the country.

To address this challenge in the Swiss context, we introduce the **Swiss Corpora Collection (SwissCoco)**, an initiative to systematically identify, catalogue, and provide an overview of language resources relevant to Switzerland. SwissCoco is organized by the Swiss Association for Natural Language Processing (SwissNLP) at https://swissnlp.org/swiss-corpora-collection. This paper presents the inaugural collection, **SwissCoco2025**, which brings together publicly available corpora that are essential for advancing research for Swiss languages and linguistic varieties.

For a corpus to be included in SwissCoco, it must meet at least one of the following criteria that define it as "Switzerland-related":

1. is produced in Switzerland,
2. represents one or more Swiss national or regional languages (German, French, Italian, Romansh, Swiss German dialects),
3. reflects Swiss-specific linguistic varieties,
4. originates from a Swiss entity, or
5. contains content relevant to Swiss contexts.

Furthermore, all corpora in the collection must be publicly available to ensure accessibility for the wider research community. **SwissCoco2025** is the first iteration of this collection and contains 30 corpora. The full list and detailed information on each corpus are also available on the SwissNLP website at https://swissnlp.org/swiss-corpora-collection. New corpora are added throughout the year on the website and will be formally documented in the next annual survey paper in 2026.

2 Corpus Curation

The curation of the corpora followed a methodology centered on two core principles, as outlined in the introduction: corpora must be **Switzerland-related** and **publicly available**.

The collection process for the inaugural Swiss Corpora Collection 2025 leveraged existing knowledge within the SwissNLP community and was complemented by targeted web searches and manual information extraction. More precisely, the initial list of corpora was compiled from resources already known to the authors and other members of SwissNLP. This list was then expanded through

web searches to identify additional publicly available resources. The focus was on identifying a comprehensive set of foundational corpora that have been previously used in research.

The SwissText 2025 Conference hosted a special "Corpora Track", which called for Switzerland-related submissions. Two corpora were published within this track (SwissGPC and SPC_R), which are also included in the corpora list of SwissCoco2025.

The identified corpora were individually reviewed against the inclusion criteria. Each entry in the collection includes metadata such as the corpus name, a brief description, the languages covered, the data type (e.g., text, speech), and a link to the resource, which were extracted manually from the corpus' documentation and references. The resulting collection, as presented in the following section, provides the most comprehensive overview of publicly available Switzerland-related language resources to date.[1]

3 SwissCoco2025

The Swiss Corpora Collection 2025 contains 30 corpora and datasets that are relevant for the Swiss context. Table 1 gives an overview of the corpora. The full details including annotation types, license information, contact addresses and other relevant metadata, are provided in Appendix A.

In SwissCoco2025, 18 corpora contain texts, 12 contain speech recordings, and one contains web pages with text and images. There are 14 corpora with Swiss German data (text or speech) and one specific for Valais German.

4 Conclusion

This paper presents the inaugural edition of the Swiss Corpora Collection (SwissCoco2025), a comprehensive overview of language resources relevant to Switzerland. We have curated a collection of 30 publicly available corpora, spanning various modalities, including text, speech, and documents, and covering the country's national and regional languages. This collection, which is also publicly listed at https://swissnlp.org/swiss-corpora-collection, is a centralized resource to facilitate NLP research for Swiss-related linguistic varieties.

Building on the foundation of SwissCoco2025, we plan to systematically expand the collection by identifying new resources and regularly updating the metadata of existing ones. By providing this curated collection, we hope to lower the entry barrier for new researchers and contribute to the development of robust and effective NLP systems tailored to Switzerland's unique multilingual and multi-dialectal landscape.

[1] If you have suggestions for additional corpora, or if you notice any errors, please send an email to info@swissnlp.org

Table 1: All Corpora of SwissCoco2025. Column "Com. Use" indicates whether commercial use is possible. See Appendix for more details.

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
1	**ArchiMob**: Archives de la mobilisation	Speech, Text	Swiss German	Transcripts of interviews on the mobilisation in the Second World War in Switzerland	500k tokens	No	Link
2	**BCMS-MT**: Map Task Corpus of Heritage BCMS	Speech, Text	Swiss German	Spontaneous dialogues in Swiss German. Recordings are annotated with dialogue acts and speaker characteristics.	3 hours	Yes	Link
3	**CEASR**: Corpus for Evaluating Automatic Speech Recognition	Speech, Transcripts	German, English	Audio recordings from nine English and six German speech corpora and accompanying transcriptions generated by seven different ASR systems.	56 hours, 1360 speakers	Yes	Link
4	**CHEU-lex**: CHEU-lex Corpus	Text	German, French, Italian	Parallel and comparable corpus of Swiss and European Union (EU) legislation.	Not specified	Yes	Link
5	**DS21 Corpus**: Corpus of Historical Legal Texts	Text: Documents	German, French, Italian, Romansh, Latin	Historical Swiss legal texts from the early Middle Ages to 1798. Based on the Collection of Swiss Law Sources.	Varies by canton and volume	Yes	Link
6	**GSA-Data**: German Speaking Area Data	Text	German, Swiss German, Austrian	German, Austrian and Swiss German Jodels with geolocations	16.8M posts	Unk.	Link
7	**LEDGAR**: Multi-label Corpus for Text Classification of Legal Provisions in Contracts	Text: Legal Documents	English	Legal judgments from the Swiss Federal Supreme Court, intended for legal document analysis.	60k legal documents, 100k provisions, 12k labels	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
8	**LEX.CH.IT**: Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian	Text: Documents	Italian	Monolingual corpus of Swiss normative acts, including 366 federal acts, from 1974 to 2018.	366 acts	No	Link
9	**MediaParl**: MediaParl Bilingual Database	Speech	French, German	Bilingual speech database with recordings from the Valais Parliament.	16k sentences, 210 speakers	No	Link
10	**NOAH's Corpus**: NOAH's Corpus of Swiss German Dialects	Text: Various Sources	Swiss German	Swiss German texts from various genres, including Wikipedia articles, news, blogs, and novels. Manually annotated with Part-of-Speech tags.	73k tokens	No	Link
11	**SB-10k**: German Sentiment Corpus	Text: Tweets	German	German tweets from 2017, each annotated by 3 annotators with sentiment labels "positive", "negative", "neutral", "mixed", or "unknown".	9700 tweets	Yes	Link
12	**SB-CH**: Swiss German Sentiment Corpus	Text: Social Media	Swiss German	Swiss German sentences from Facebook comments and online chats. Includes manual sentiment labels for some sentences.	166k sentences, 2800 with sentiment labels	Yes	Link
13	**SDATS Corpus**: Swiss German Dialects Across Time and Space Corpus	Speech	Swiss German	Spoken Swiss German recordings from 1,000 speakers across 125 localities.	1k speakers, 125 localities, 300 variables	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
14	**SDS-200**: Schweizer Dialeksammlung-200	Speech	Swiss German	Swiss German audio recordings with transcripts in Standard German. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification.	200 hours, 4000 speakers	Yes	Link
15	**SMG-CH**: Social Media Variety Geolocation - Swiss German	Text	Swiss German	Swiss German Jodels with geolocations	29k Jodel conversations	Unk.	Link
16	**SPC_R**: Swiss Parliaments Corpus Re-Imagined	Speech, Text	Swiss German, Standard German	Enhanced long-form speech-text corpus of Swiss German parliamentary debates, with high-quality, corrected transcriptions.	751 hours	Yes	Link
17	**STT4SG-350**: Speech-to-Text for Swiss German-350	Speech	Swiss German	Swiss German audio recordings with transcripts in Standard German. Balanced distribution across dialects and demographics such as gender. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification. Dedicated test set with approx. 5 hours of audio of identical sentences spoken in 7 different dialects.	343 hours, 316 speakers	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
18	**Swiss Politics Corpus**: Swiss Politics Corpus	Text: Documents	German, French, Italian	A database of who said what and when in both chambers of the Swiss parliament over the past 127 years, based on digitized proceedings with oldest documents being from 1891	40k documents	No	Link
19	**Swiss SMS Corpus**: Swiss SMS Corpus	Text: SMS	Swiss German, German, French, Italian, Romansh	SMS messages crowdsourced from the Swiss public.	26k SMS messages, 650k tokens	No	Link
20	**SwissCrawl**: SwissCrawl Web Corpus	Text: Web	Swiss German	Large-scale, multilingual web crawl of the .ch domain.	560k sentences	No	Link
21	**SwissDial**: Parallel Multidialectal Corpus of Spoken Swiss German	Speech	Swiss German	Audio recordings of eight major Swiss German dialects and corresponding transcripts in Swiss German and Standard German	26 hours, 8 speakers	No	Link
22	**SwissGPC**: Swiss German Podcasts Corpus	Speech	Swiss German	Links to Swiss podcast from Swiss TV broadcasters	5000 hours	No	Link
23	**TEVOID**: Temporal Voice Idiosyncrasy	Speech, Text	Swiss German (Zurich)	Recordings of read and spontaneous speech of different speakers of Zurich German for the research on idiosyncratic differences	16 sentences	Yes	Link
24	**TRANSLIT**: Large-scale Name Transliteration Resource	Text	180 languages	Variations of person and geolocation names in various languages.	1.6 million entries	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
25	**VarDial_GDI17**: German Dialect Identification at VarDial 2017	Text	Swiss German	Text containing dialects from the cantons of Basel (BS), Bern (BE), Lucerne (LU), and Zurich (ZH)	18k sentences	Unk.	Link
26	**Walliserdeutsch ASR Corpus**: ASR and Translation Corpus for Walliserdeutsch	Speech	Walliserdeutsch German	Broadcast news from a local radio station in Walliserdeutsch dialect.	8.1 hours	Yes	Link
27	**WebClasSeg25**: WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset	Text, Image	25 languages	Webpages from public sector websites of Europe	2580 webpages	Yes	Link
28	**What's up, Switzerland?**: Swiss Chat Corpus	Text: Messages	Swiss German, German, French, Italian, Romansh, Spanish, Slavic languages	WhatsApp chat messages, gathered in Summer 2014	760k messages, 5.1mio tokens	No	Link
29	**ZHCORPUS**: Zurich Corpus of Vowel and Voice Quality	Speech, Text	Swiss German (Zurich)	Focused on sounds of the long Standard German vowels produced with varying basic production parameters	34k utterances, 70 speaker	Yes	Link
30	**ZTC_BAS**: Zurich Tangram Corpus	Speech	Swiss German (Zurich)	Recordings of Swiss German dialect from Zurich, including transcriptions.	2 hours	Unk.	Link

APPENDIX: Full Corpora Details

This appendix provides the full details for each of the corpora listed in SwissCoco2025. The data provides detailed information about each resource, including its full name, authors/creators, languages, modalities, size, publication year, licensing details, and a direct link to the data or its corresponding publication.

1: ArchiMob - Archives de la mobilisation

Reference: Samardžić et al. (2016)
Modalities: Speech, Text
Languages: Swiss German
Description: Transcripts of interviews on the mobilisation in the Second World War in Switzerland
Size: 500k tokens
Annotations: Transcripts, Audio, metadata
Publication Year: 2016
License(s): CC NC 4.0
Commercial Use: No
Website: Link
Reference: Tanja Samardzic, Yves Scherrer, Elvira Glaser: *ArchiMob - A corpus of Spoken Swiss German*. LREC 2016.
Paper Link: Link
Contact: tanja.samardzic@uzh.ch

2: BCMS-MT - Map Task Corpus of Heritage BCMS

Reference: Lemmenmeier-Batinić et al. (2023)
Modalities: Speech, Text
Languages: Swiss German
Description: Spontaneous dialogues in Swiss German. Recordings are annotated with dialogue acts and speaker characteristics.
Size: 3 hours
Annotations: Dialogue acts, speaker characteristics
Publication Year: 2023
License(s): Not specified
Commercial Use: Yes
Website: Link
Reference: Dolores Lemmenmeier-Batinić, Josip Batinić, Anastasia Escher: *Map Task Corpus of Heritage BCMS spoken by second-generation speakers in Switzerland*. Language Resources and Evaluation 57.4, 2023.
Paper Link: Link
Contact: dolores.lemmenmeier@uzh.ch

3: CEASR - Corpus for Evaluating Automatic Speech Recognition

Reference: Ulasik et al. (2020)
Modalities: Speech, Transcripts
Languages: German, English
Description: Audio recordings from nine English and six German speech corpora and accompanying transcriptions generated by seven different ASR systems.
Size: 56 hours, 1360 speakers
Annotations: Transcripts from different ASR engines, meta-data such as gender, accent etc.
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, Mark Cieliebak: *CEASR: A Corpus for Evaluating Automatic Speech Recognition*. LREC 2020.
Paper Link: Link
Contact: ciel@zhaw.ch

4: CHEU-lex - CHEU-lex Corpus

Reference: Felici (2025)
Modalities: Text
Languages: German, French, Italian
Description: Parallel and comparable corpus of Swiss and European Union (EU) legislation.
Size: Not specified
Annotations: Structural, morphosyntactic, and content-related information
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Annarita Felici: *CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation*. Applied Corpus Linguistics 2025.
Paper Link: Link
Contact: Annarita.Felici@unige.ch

5: DS21 Corpus - Corpus of Historical Legal Texts

Reference: Höfler and Piotrowski (2011)
Modalities: Text: Documents
Languages: German, French, Italian, Romansh, Latin
Description: Historical Swiss legal texts from the early Middle Ages to 1798. Based on the Collection of Swiss Law Sources.
Size: Varies by canton and volume
Annotations: Transcribed, annotated, commented
Remarks: Dataset link does not work
Publication Year: 2011
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Alternative Names: Collection of Swiss Law Sources
Reference: Stefan Höfler, Michael Piotrowsk: *Building Corpora for the Philological Study of Swiss Legal Texts*. 2011.
Paper Link: Link
Contact: Not available

6: GSA-Data - German Speaking Area Data

Reference: Hovy and Purschke (2018)
Modalities: Text
Languages: German, Swiss German, Austrian
Description: German, Austrian and Swiss German Jodels with geolocations
Size: 16.8M posts
Annotations: Coordinates
Publication Year: 2018
License(s): not specified
Commercial Use: Unk.
Website: Link
Reference: Dirk Hovy, Christoph Purschke: *Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting*. EMNLP 2018.
Paper Link: Link
Contact: dirk.hovy@unibocconi.it

7: LEDGAR - Multilabel Corpus for Text Classification of Legal Provisions in Contracts

Reference: Tuggener et al. (2020)
Modalities: Text: Legal Documents
Languages: English
Description: Legal judgments from the Swiss Federal Supreme Court, intended for legal document analysis.
Size: 60k legal documents, 100k provisions, 12k labels
Annotations: Labeled provisions in contracts and legal texts
Remarks: Created by ZHAW and a Swiss Startup
Publication Year: 2020
License(s): MIT License
Commercial Use: Yes
Website: Link
Reference: Don Tuggener, Pius von Däniken, Thomas Peetz, Mark Cieliebak: *LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts*. LREC 2020.
Paper Link: Link
Contact: tuge@zhaw.ch

8: LEX.CH.IT - Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian

Reference: Canavese (2019)
Modalities: Text: Documents
Languages: Italian
Description: Monolingual corpus of Swiss normative acts, including 366 federal acts, from 1974 to 2018.
Size: 366 acts
Annotations: -
Remarks: No download link found
Publication Year: 2019
License(s): CC BY 4.0
Commercial Use: No
Website: Link
Reference: Paolo Canavese: *LEX.CH.IT: A Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian*. Comparative Legilinguistics 40.1, 2019.
Paper Link: Link
Contact: Paolo.Canavese@unige.ch

9: MediaParl - MediaParl Bilingual Database

Reference: Imseng et al. (2012)
Modalities: Speech
Languages: French, German
Description: Bilingual speech database with

recordings from the Valais Parliament.
Size: 16k sentences, 210 speakers
Annotations: Transcripts, speaker metadata, language tags
Remarks: Split German-French approx. 50:50
Publication Year: 2012
License(s): Non-commercial research only
Commercial Use: No
Website: Link
Reference: David Imseng, Hervé Bourlard, Holger Caesar, Philip N. Garner, Gwénolé Lecorvé, Alexandre Nanchen: *MediaParl: Bilingual mixed language accented speech database*. Spoken Language Technology 2012.
Paper Link: Link
Contact: Not available

10: NOAH's Corpus - NOAH's Corpus of Swiss German Dialects

Reference: Hollenstein and Aepli (2014)
Modalities: Text: Various Sources
Languages: Swiss German
Description: Swiss German texts from various genres, including Wikipedia articles, news, blogs, and novels. Manually annotated with Part-of-Speech tags.
Size: 73k tokens
Annotations: Part-of-Speech tags
Versions: V1.0 from 2014, NOAH 3.0 contains 114k tokens
Publication Year: 2014
License(s): CC Attribution 4.0
Commercial Use: No
Website: Link
Reference: Nora Hollenstein, Noemi Aepli: *Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging*. VarDial 2014.
Paper Link: Link
Contact: Not available

11: SB-10k - German Sentiment Corpus

Reference: Cieliebak et al. (2017)
Modalities: Text: Tweets
Languages: German
Description: German tweets from 2017, each annotated by 3 annotators with sentiment labels "positive", "negative", "neutral", "mixed", or "unknown".
Size: 9700 tweets
Annotations: Sentiment labels
Publication Year: 2017
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Mark Cieliebak, Jan Deriu, Fatih Uzdilli, Dominic Egger: *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*. SocialNLP 2017.
Paper Link: Link
Contact: info@spinningbytes.com

12: SB-CH - Swiss German Sentiment Corpus

Reference: Grubenmann et al. (2018)
Modalities: Text: Social Media
Languages: Swiss German
Description: Swiss German sentences from Facebook comments and online chats. Includes manual sentiment labels for some sentences.
Size: 166k sentences, 2800 with sentiment labels
Annotations: Sentiment labels
Publication Year: 2018
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Ralf Grubenmann, Don Tuggener, Pius von Daniken, Jan Deriu, Mark Cieliebak : *Towards a Corpus of Swiss German Annotated with Sentiment*. LREC 2018.
Paper Link: Link
Contact: info@spinningbytes.com

13: SDATS Corpus - Swiss German Dialects Across Time and Space Corpus

Reference: Leemann et al. (2020)
Modalities: Speech
Languages: Swiss German
Description: Spoken Swiss German recordings from 1,000 speakers across 125 localities.
Size: 1k speakers, 125 localities, 300 variables
Annotations: Sociolinguistic and psycholinguistic metadata, phonetic variables
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Adrian Leemann, Péter Jeszenszky Carina Steiner, Jan Messerli, Melanie Studerus:

SDATS Corpus – Swiss German dialects across time and space. 2020.
Paper Link: Link
Contact: carina.steiner@phbern.ch

14: SDS-200 - Schweizer Dialeksammlung-200

Reference: Plüss et al. (2022)
Modalities: Speech
Languages: Swiss German
Description: Swiss German audio recordings with transcripts in Standard German. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification.
Size: 200 hours, 4000 speakers
Annotations: Transcripts, dialect information, age group, gender.
Remarks: Same data format as STT4SG-350
Publication Year: 2022
License(s): META-SHARE NonCommercial NoRedistribution
Commercial Use: Yes
Website: Link
Reference: Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, Manfred Vogel: *SDS-200: A Swiss German Speech to Standard German Text Corpus*. LREC 2022.
Paper Link: Link
Contact: info@swissnlp.org

15: SMG-CH - Social Media Variety Geolocation - Swiss German

Reference: Gaman et al. (2020)
Modalities: Text
Languages: Swiss German
Description: Swiss German Jodels with geolocations
Size: 29k Jodel conversations
Annotations: Sentences, Coordinates
Remarks: Dataset used in VarDial 2020
Publication Year: 2020
License(s): Not specified
Commercial Use: Unk.
Website: Link
Reference: Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, Marcos Zampieri: *A Report on the VarDial Evaluation Campaign 2020*. VarDial 2020.
Paper Link: Link
Contact: yves.scherrer@helsinki.fi

16: SPC_R - Swiss Parliaments Corpus Re-Imagined

Reference: Timmel et al. (2025)
Modalities: Speech, Text
Languages: Swiss German, Standard German
Description: Enhanced long-form speech-text corpus of Swiss German parliamentary debates, with high-quality, corrected transcriptions.
Size: 751 hours
Annotations: Transcriptions, LLM-based correction, predicted BLEU scores
Remarks: An extension of the original Swiss Parliaments Corpus
Publication Year: 2025
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Vincenzo Timmel, Manfred Vogel, Daniel Perruchoud, Reza Kakooee: *Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced Transcription with RAG-based Correction and Predicted BLEU*. arXiv 2025.
Paper Link: Link
Contact: info@swissnlp.org

17: STT4SG-350 - Speech-to-Text for Swiss German-350

Reference: Plüss et al. (2023)
Modalities: Speech
Languages: Swiss German
Description: Swiss German audio recordings with transcripts in Standard German. Balanced distribution across dialects and demographics such as gender. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification. Dedicated test set with approx. 5 hours of audio of identical sentences spoken in 7 different dialects.
Size: 343 hours, 316 speakers
Annotations: Transcripts, dialect information, age group, gender.
Versions: Extension of Swiss Parliament Corpus SPC

Remarks: Same data format as SDS-200
Publication Year: 2023
License(s): META-SHARE NonCommercial NoRedistribution
Commercial Use: Yes
Website: Link
Reference: Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, Mark Cieliebak: *STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions*. ACL 2023.
Paper Link: Link
Contact: info@swissnlp.org

18: Swiss Politics Corpus - Swiss Politics Corpus

Reference: Salamanca (2018)
Modalities: Text: Documents
Languages: German, French, Italian
Description: A database of who said what and when in both chambers of the Swiss parliament over the past 127 years, based on digitized proceedings with oldest documents being from 1891
Size: 40k documents
Annotations: labeled text lines and paragraphs
Publication Year: 2018
License(s): MIT License
Commercial Use: No
Website: Link
Reference: Luis Salamanca, Lilian Gasser, Laurence Brandenberger, Frank Schweitzer: *A trip through Swiss politics and history*. Blog Post 2018.
Paper Link: Link
Contact: Not available

19: Swiss SMS Corpus - Swiss SMS Corpus

Reference: Stark et al. (2009-2015)
Modalities: Text: SMS
Languages: Swiss German, German, French, Italian, Romansh
Description: SMS messages crowdsourced from the Swiss public.
Size: 26k SMS messages, 650k tokens
Annotations: Language, PoS tags
Remarks: 41% Swiss German, 28% German, 18% French, 6% Italian, 4% Romansh
Publication Year: 2009
License(s): CC-NY-NC
Commercial Use: No
Website: Link
Reference: Elisabeth Stark, Simone Ueberwasser, Beni Ruef: *Swiss SMS Corpus*. 2009-2015.
Contact: sms@cl.uzh.ch

20: SwissCrawl - SwissCrawl Web Corpus

Reference: Linder et al. (2020)
Modalities: Text: Web
Languages: Swiss German
Description: Large-scale, multilingual web crawl of the .ch domain.
Size: 560k sentences
Annotations: Crawling data
Remarks: 89% of sentences in Swiss German
Publication Year: 2020
License(s): CC BY-NC 4.0
Commercial Use: No
Website: Link
Reference: Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, Andreas FischerLucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, Andreas Fischer: *Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German*. LREC 2020.
Paper Link: Link
Contact: andreas.fischer@hefr.ch

21: SwissDial - Parallel Multidialectal Corpus of Spoken Swiss German

Reference: Dogan-Schönberger et al. (2021)
Modalities: Speech
Languages: Swiss German
Description: Audio recordings of eight major Swiss German dialects and corresponding transcripts in Swiss German and Standard German
Size: 26 hours, 8 speakers
Annotations: Transcripts and dialect information
Versions: V1.0 from 2021, V1.1 contains additional 7726 recorded GR sentences.
Publication Year: 2021
License(s): Research use only, commercial use restricted
Commercial Use: No
Website: Link
Reference: Pelin Dogan-Schonberger, Julian

Mäder, Thomas Hofmann: *SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German*. arXiv 2021.
Paper Link: Link
Contact: Not available

22: SwissGPC - Swiss German Podcasts Corpus

Reference: Stucki et al. (2025)
Modalities: Speech
Languages: Swiss German
Description: Links to Swiss podcast from Swiss TV broadcasters
Size: 5000 hours
Annotations: Transcripts, audio, language tags
Publication Year: 2025
License(s): CC BY 4.0 for Link Collection
Commercial Use: No
Website: Link
Reference: Samuel Stucki, Jan Deriu, Mark Cieliebak: *SwissGPC v1.0 - The Swiss German Podcasts Corpus*. SwissText 2025.
Paper Link: Link
Contact: deri@zhaw.ch

23: TEVOID - Temporal Voice Idiosyncrasy

Reference: Dellwo et al. (2012)
Modalities: Speech, Text
Languages: Swiss German (Zurich)
Description: Recordings of read and spontaneous speech of different speakers of Zurich German for the research on idiosyncratic differences
Size: 16 sentences
Annotations: Transcripts, audio
Publication Year: 2012
License(s): Not specified
Commercial Use: Yes
Website: Link
Reference: Volker Dellwo, Adrian Leemann, Marie-José Kolly: *Speaker idiosyncratic rhythmic features in the speech signal*. Interspeech 2012.
Paper Link: Link
Contact: volker.dellwo@uzh.ch

24: TRANSLIT - Large-scale Name Transliteration Resource

Reference: Benites et al. (2020)
Modalities: Text
Languages: 180 languages
Description: Variations of person and geolocation names in various languages.
Size: 1.6 million entries
Annotations: Sentiment labels
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, Mark Cieliebak: *TRANSLIT: A Large-scale Name Transliteration Resource*. LREC 2020.
Paper Link: Link
Contact: Not available

25: VarDial_GDI17 - German Dialect Identification at VarDial 2017

Reference: Zampieri et al. (2017)
Modalities: Text
Languages: Swiss German
Description: Text containing dialects from the cantons of Basel (BS), Bern (BE), Lucerne (LU), and Zurich (ZH)
Size: 18k sentences
Annotations: Dialects
Publication Year: 2017
License(s): MIT License
Commercial Use: Unk.
Website: Link
Reference: Marcos Zampieri, Shervin Malmasi, Nikola Ljubesi, Preslav Nakov, Ahmed Ali, Jorg Tiedemann, Yves Scherrer, Noemi Aepli: *Findings of the VarDial Evaluation Campaign 2017*. VarDial 2017.
Paper Link: Link
Contact: simon.clematide@cl.uzh.ch

26: Walliserdeutsch ASR Corpus - ASR and Translation Corpus for Walliserdeutsch

Reference: Garner et al. (2014)
Modalities: Speech
Languages: Walliserdeutsch, German
Description: Broadcast news from a local radio

station in Walliserdeutsch dialect.
Size: 8.1 hours
Annotations: Transcribed speech and translated text
Publication Year: 2014
License(s): Not specified
Commercial Use: Yes
Website: Link
Alternative Names: Walliserdeutsch ASR; Walliserdeutsch speech corpus
Reference: Philip N. Garner, David Imseng, Thomas Meyer: *Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch*. Interspeech 2014.
Paper Link: Link
Contact: dimseng@idiap.ch

27: WebClasSeg25 - WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset

Reference: Gerber et al. (2025)
Modalities: Text, Image
Languages: 25 languages
Description: Webpages from public sector websites of Europe
Size: 2580 webpages
Annotations: Sentiment labels, Crawling Data, Screenshots
Publication Year: 2025
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Jonathan Gerber, Jasmin Saxer, Kimia Rabishokr, Bruno Kreiner, Andreas Weiler: *WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset - Integrating Functional and Maturity-Based Analysis*. SIGIR 2025.
Paper Link: Link
Contact: jonathan.gerber@zhaw.ch

28: What's up, Switzerland? - Swiss Chat Corpus

Reference: Ueberwasser and Stark (2017)
Modalities: Text: Messages
Languages: Swiss German, German, French, Italian, Romansh, Spanish, Slavic languages
Description: WhatsApp chat messages, gathered in Summer 2014
Size: 760k messages, 5.1mio tokens
Annotations: Language tags
Publication Year: 2014
License(s): Non-commercial research only
Commercial Use: No
Website: Link
Reference: Simone Ueberwasser, Elisabeth Stark: *What's up, Switzerland? A corpus-based research project in a multilingual country*. Linguistik online 84/5, 2017.
Paper Link: Link
Contact: estark@rom.uzh.ch

29: ZHCORPUS - Zurich Corpus of Vowel and Voice Quality

Reference: Maurer et al. (2018)
Modalities: Speech, Text
Languages: Swiss German (Zurich)
Description: Focused on sounds of the long Standard German vowels produced with varying basic production parameters
Size: 34k utterances, 70 speaker
Annotations: Transcripts, audio
Publication Year: 2018
License(s): Research use only, commercial use restricted
Commercial Use: Yes
Website: Link
Reference: Dieter Maurer, Christian d'Heureuse, Heidy Suter, Volker Dellwo, Daniel Friedrichs, Thayabaran Kathiresan: *The Zurich Corpus of Vowel and Voice Quality, Version 1.0*. Interspeech 2018.
Paper Link: Link
Contact: dieter.maurer@zhdk.ch

30: ZTC_BAS - Zurich Tangram Corpus

Reference: Kalmanovitch (2016)
Modalities: Speech
Languages: Swiss German (Zurich)
Description: Recordings of Swiss German dialect from Zurich, including transcriptions.
Size: 2 hours
Annotations: Transcripts, audio
Publication Year: 2019
License(s): Not specified
Commercial Use: Unk.
Website: Link
Reference: Yshai Kalmanovitch, Wolfgang Kesselheim: *The Zurich Tangram Corpus - BAS*

Edition. 2019.
Paper Link: Link
Contact: bas@bas.uni-muenchen.de

References

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Paolo Canavese. 2019. LEX.CH.IT: A corpus for microdiachronic linguistic investigations of swiss normative acts in italian. *Comparative Legilinguistics*, 40(1):43–65.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Volker Dellwo, Adrian Leemann, and Marie-José Kolly. 2012. Speaker idiosyncratic rhythmic features in the speech signal. Interspeech Conference Proceedings.

Pelin Dogan-Schönberger, Julina Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German.

Annarita Felici. 2025. CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation. *Applied Corpus Linguistics*, 5(3):100151.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Philip N Garner, David Imseng, and Thomas Meyer. 2014. Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. In *INTERSPEECH*, pages 2118–2122.

Jonathan Gerber, Jasmin Saxer, Kimia Rabishokr, Bruno Kreiner, and Andreas Weiler. 2025. Webclasseg-25: A dual-classified webpage segmentation dataset - integrating functional and maturity-based analysis. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3792–3801, New York, NY, USA. Association for Computing Machinery.

Ralf Grubenmann, Don Tuggener, Pius von Däniken, Jan Deriu, and Mark Cieliebak. 2018. SB-CH: A Swiss German corpus with sentiment annotations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Höfler and Michael Piotrowski. 2011. Building corpora for the philological study of Swiss legal texts. *Journal for Language Technology and Computational Linguistics*, 26(2):77–89.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

David Imseng, Hervé Bourlard, Holger Caesar, Philip N Garner, Gwénolé Lecorvé, and Alexandre Nanchen. 2012. MediaParl: Bilingual mixed language accented speech database. In *2012 IEEE spoken language technology workshop (SLT)*, pages 263–268. IEEE.

Yshai Kalmanovitch. 2016. Speech in interaction–the Zurich Tangram corpus. *Tagungsband der Tagung Phonetik und Phonolotie im Deutschsprachigen Raum*, 12:79–81.

Adrian Leemann, Péter Jeszenszky, Carina Steiner, Jan Messerli, and Melanie Studerus. 2020.

Dolores Lemmenmeier-Batinić, Josip Batinić, and Anastasia Escher. 2023. Map Task Corpus of Heritage BCMS spoken by second-generation speakers in Switzerland. *Language Resources and Evaluation*, 57(4):1607–1644.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the Internet: The case of Swiss German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Dieter Maurer, Christian d'Heureuse, Heidy Suter, Volker Dellwo, Daniel Friedrichs, and Thayabaran Kathiresan. 2018. The Zurich Corpus of vowel and voice quality, version 1.0.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 3250–3256, Marseille, France. European Language Resources Association.

Luis Salamanca. 2018. A trip through swiss politics and history. `https://www.datascience.ch/articles/trip-swiss-politics-history`. Accessed: 2025-10-17.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Elisabeth Stark, Simone Ueberwasser, and Beni Ruef. 2009-2015. www.sms4science.ch.

Samuel Stucki, Mark Cieliebak, and Jan Deriu. 2025. SwissGPC v1.0 – The Swiss German Podcasts Corpus. *arXiv preprint arXiv:2509.19866*.

Vincenzo Timmel, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025. Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced transcription with RAG-based correction and predicted BLEU. *arXiv preprint arXiv:2506.07726*.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online*, 84(5).

Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, and Mark Cieliebak. 2020. CEASR: A corpus for evaluating automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6477–6485, Marseille, France. European Language Resources Association.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced Transcription with RAG-based Correction and Predicted BLEU

Vincenzo Timmel[1], Manfred Vogel[1], Daniel Perruchoud[1], Reza Kakooee[1]

[1]University of Applied Sciences and Arts Northwestern Switzerland

{vincenzo.timmel, manfred.vogel, daniel.perruchoud, reza.kakooee}@fhnw.ch

Abstract

This paper presents a new long-form release of the Swiss Parliaments Corpus, converting entire multi-hour Swiss German debate sessions (each aligned with the official session protocols) into high-quality speech–text pairs. Our pipeline starts by transcribing all session audio into Standard German using Whisper Large-v3 under high-compute settings. We then apply a two-step GPT-4o correction process: first, GPT-4o ingests the raw Whisper output alongside the official protocols to refine misrecognitions, mainly named entities. Second, a separate GPT-4o pass evaluates each refined segment for semantic completeness. We filter out any segments whose Predicted BLEU score (derived from Whisper's average token log-probability) and GPT-4o evaluation score fall below a certain threshold. The final corpus contains 801 hours of audio, of which 751 hours pass our quality control. Compared to the original sentence level SPC release, our long-form dataset achieves a 6-point BLEU improvement, demonstrating the power of combining robust ASR, LLM-based correction, and data-driven filtering for low-resource, domain-specific speech corpora.

1 Introduction

Data scarcity in low-resource domains still hinders the development of Automatic Speech Recognition (ASR) systems. For Swiss German, (Plüss et al., 2021) contributed the Swiss Parliaments Corpus (SPC), including a meticulously prepared training dataset with high alignment quality of 176 hours of Swiss German speech paired with Standard German transcripts of Bernese parliamentary debates with a corresponding curated test dataset of 6 hours. The corpus was built using a forced sentence alignment procedure and alignment quality estimator that overcomes challenges such as sentence reordering and language mismatches between Swiss German audio and Standard German text. They used a global alignment algorithm based on Needleman-Wunsch and an Intersection over Union (IoU) estimator to filter out poor-quality alignments. Additional filters, such as character-per-second limits and language detection, ensured that only accurately aligned sentences were included.

The SPC_R corpus presented in this paper is an extension of the original SPC corpus focusing on the creation, curation, and release of datasets tailored to Swiss German NLP applications. Originally, crawled data from the parliament debates of the Grosser Rat Kanton Bern encompass 801 hours of session recordings in long-form with a length spanning from 28 to 242 minutes paired with official session protocols.

In contrast to (Plüss et al., 2021), which extracts sentences from parliamentary sessions by finding near-perfect matches between automatically generated transcriptions and the official session protocols, we incorporate an advanced transcription pipeline in SPC_R. This includes the Whisper Large-v3 model (Radford et al., 2023) for transcription, and a post-correction step using GPT-4o (Hurst et al., 2024), aligned with the official protocol to further enhance transcription quality and overall data accuracy.

In addition, the SPC_R corpus provides the data in long-form, whereas the original SPC is segmented at sentence level.

The primary contributions include:

- High-quality transcription by Whisper Large-v3 of approximately 801 hours of audio with high-compute settings, see Section 3.
- BLEU score (Papineni et al., 2002) prediction based on Whisper transcription outputs via linear regression.
- A two-step large language model (LLM) approach in which a first model corrects the tran-

scription and a second, independent model evaluates that correction.

This paper provides detailed insights into the methodology, experimental results, and implications for future NLP dataset releases in Swiss German.

2 Related Work

In the past years, several initiatives (Plüss et al., 2021, 2022, 2023; Dogan-Schönberger et al., 2021) made valuable contributions for the development of Swiss German ASR solutions; an overview of the released datasets is shown in Figure 1. However, these datasets are all at sentence level which typically does not improve ASR solutions for real-world situations (Timmel et al., 2024). Additionally, not all existing datasets can be used for commercial purposes.

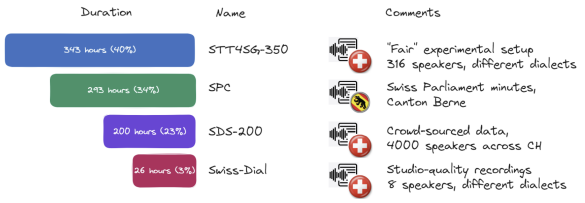


Figure 1: Overview of Swiss German speech to German text datasets. Usage of SPC is possible under MIT license, SDS-200 and STT4SG-350 under SwissNLP license. SwissDial can be used exclusively for research purposes.

3 Transcription with Whisper Large-v3

The starting point for the construction of the SPC_R Corpus is 801 hours of long-form audio from parliament debates of Grosser Rat Kanton Bern which we transcribe with Whisper Large-v3.

Our transcription pipeline uses Whisper Large-v3 via WhisperX (Bain et al., 2023) under high-compute settings, namely *beam_size* set to 10, *best_of* set to 10, and *log_prob_threshold* set to –2. All transcriptions are performed on an NVIDIA A4500 GPU with 20 GB of VRAM, using *float16* precision and a *batch_size* of 8. These high-compute settings further improve results, as shown in Figure 5. For all transcribed parliament sessions, we store Whisper's *avg_log_prob* output, which reflects the model's prediction confidence and exhibits strong predictive power for transcription quality, as described in Subsection 3.1.

3.1 BLEU Prediction

We observed a linear relationship between the confidence metric calculated by Whisper (Kim, 2023), as presented in Equation 1, and the BLEU score (sacreBLEU[1], more precisely) of datasets transcribed with Whisper.

$$\text{confidence} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \quad (1)$$

The confidence is derived from Whisper's segment-specific average log-probabilities *avg_log_prob*, which are averaged over the whole audio file. In Equation (1), p_i denotes the average log-probability for the ith segment, and N is the total number of segments in the entire audio file, where a segment is the text between two timestamps predicted by Whisper. Thus, the confidence is the exponential of the average *avg_log_prob* over a whole audio file.

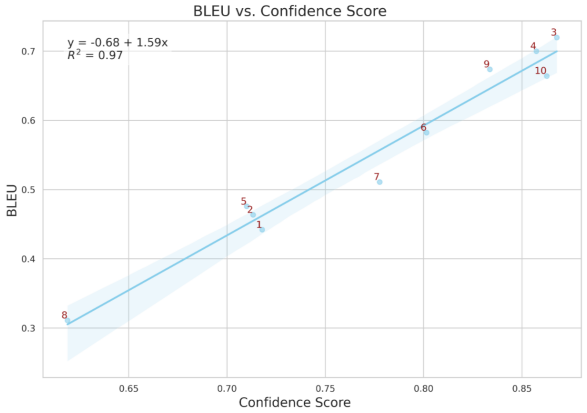


Figure 2: Linear relationship between BLEU score vs. Whisper confidence score for ten long-form conversations, represented by numbers 1-10. The blue shaded area represents the 95% confidence interval.

Figure 2 shows this linear relationship between the BLEU score (calculated between the transcription and a manually created ground truth) and the confidence on ten distinct, independent Swiss German datasets. Each dataset of approximately one hour (ca. 8'000 tokens) consists of manually transcribed Swiss German conversations (the ground truth) between two or more speakers (these datasets cannot be disclosed due to data privacy and NDA restrictions). Our analysis shows that higher confidence values are associated with higher BLEU

[1] https://github.com/mjpost/sacreBLEU (default settings: 4-gram, standard tokenization and smoothing)

scores in a near-linear fashion, indicating that the confidence metric is a strong predictor of transcription quality, suggesting its potential for assessing transcription performance.

A linear regression fitted to these data produced an intercept of -0.68 and a slope coefficient of 1.59 and allows to predict a BLEU score based solely on the confidence, called the Predicted BLEU, without first creating a ground truth.

Figure 3 shows the distribution of Predicted BLEU scores for all 131'291 segments of SPC_R, corresponding to a total of 801 hours of audio.

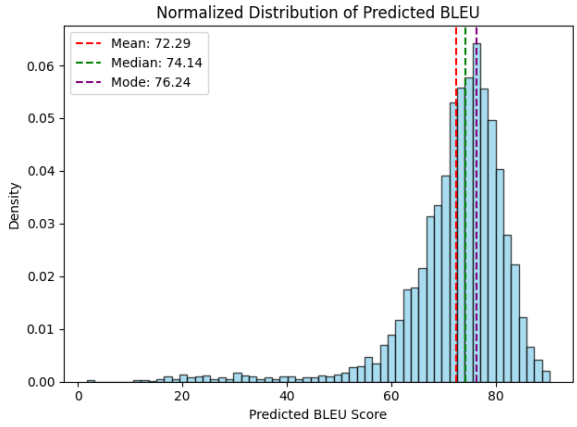


Figure 3: Distribution of Predicted BLEU scores across SPC_R (N = 131'291 data segments).

Figure 4 shows the cumulative proportion of data samples for a given Predicted BLEU score threshold. As the threshold rises, fewer samples qualify, underscoring the balance between transcription quality and the amount of available data.

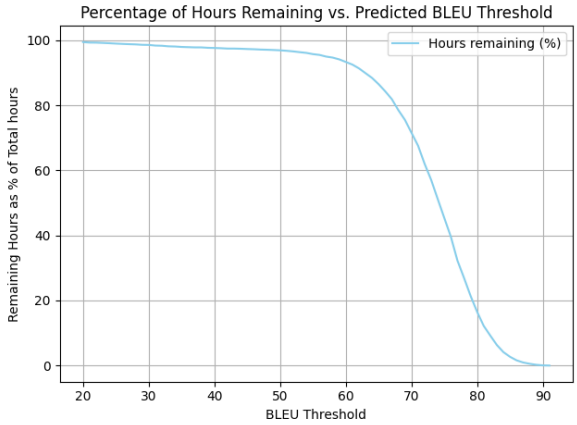


Figure 4: Percentage of data samples that have a BLEU score above the threshold.

Hence, the Predicted BLEU score derived from Whisper's *avg_log_prob* can be used to identify and select high-quality transcription segments (see Section 5).

4 Transcript correction using GPT-4o

Automated transcription with Whisper Large-v3 shows promising results but leads to errors in named entities (e.g., "Alba Rutschi" instead of "Alberucci") and other similar errors. To mitigate this, we introduce a two-step correction process using text-embedding-3-large GPT-4o and GPT-4o-mini (OpenAI, 2023):

1. **Correction Stage:** GPT-4o is used to refine the initial transcription by prompting it to correct errors, segment by segment. Corrections are based on information injected from the official manual summaries of the parliament session corresponding to the audio segment using Retrieval-Augmented Generation (RAG, see Subsection 4.1).

2. **Evaluation Stage:** Evaluation assessments of GPT-4o corrections use manual inspection on small data samples and GPT-4o-mini-as-a-Judge.

GPT-4.1 (OpenAI, 2025) was also evaluated but we found that it would repeatedly change conjugation of words, thus sometimes introducing new errors in the transcription. While still overall reducing the WER, it fixed less errors than GPT-4o.

4.1 Context provision via RAG

RAG (Lewis et al., 2020) is used to provide GPT-4o with factual context to correct the transcription.

We follow best practices (Wang et al., 2024), using Faiss (Douze et al., 2024) for efficient vector storage and retrieval, a sliding window approach and text-embedding-3-large as embedding model. Official manual summaries are ingested with *pyPDF* (Fenniak et al., 2024) using chunks of 600 characters with an overlap of 450. These values are chosen to consistently ensure a complete overlap between the transcription and the context from the chunk based on the maximum segment length of 423 characters. We pass the most relevant chunk to GPT-4o as context without re-ranking retrieved chunks.

Manual evaluation on 122 audio segments corresponding to 50 minutes of transcribed data shows that the correct chunk from the official manual summary is retrieved for 94.1% of the segments. This

high rate may be due to the ease of aligning session protocols with session transcriptions.

4.2 Correction Stage

In the correction stage, GPT-4o is given the context from subsection 4.1 and the transcription to be corrected, with an extensive, iteratively expanded system prompt specifying usage of the retrieved chunk and additional rules related to peculiarities of the Bernese dialect [2].

The pipeline run with high-compute settings improves the word error rate (WER) from 15.7% to 11.1% when evaluated on 50 minutes of manually transcribed data with temperature set to 0.1 to reduce variability and lower WER (see Figure 5).

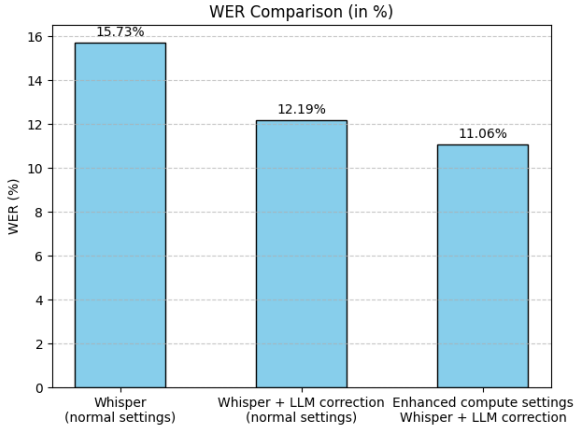


Figure 5: Word Error Rates (WER) for Whisper Large-v3 under three configurations: standard settings, after applying GPT-4o correction, and using high-compute settings (enhanced settings) with GPT-4o correction.

Additionally, when manually inspecting named entities such as places, names, legal references, and political parties, the correctness of named entity transcriptions increases from initial 72.2% with Whisper Large-v3 (52 out of 72) to 100% (72 out of 72) after applying GPT-4o correction.

Table 1 shows an example of the audio, the initial Whisper Large-v3 transcription, the context retrieved, and the output corrected with GPT-4o.

4.3 Evaluation Stage

At this stage, the quality of the transcription is evaluated in the following categories (referred to as judgment tokens hereafter):

[2]Rules include cases such as "vo dr" (audio) to be corrected from "vor der" to "von der" and "mier" (audio) to be corrected from "mir" to "wir".

Table 1: Example audio input, initial transcription with Whisper Large-v3, retrieved context (shortened) given to GPT-4o, and its output. GPT-4o is encouraged to keep the correction as close to the input as possible, so that the data can still be used to train an ASR system that relies on aligned audio and text.

Audio Input (transcribed)
dass ehr au verdaut händ, wenn ehr näbem outo send.
Whisper Large-v3 output (initial transcription)
dass er auch verdauert hat, wenn er neben dem Auto sitzt.
Context retrieved via RAG (given to GPT-4o as help for the correction.)
sodass Sie wieder leicht ernüchtert sind und verdaut haben, wenn Sie beim Auto ankommen werden.
GPT-4o output (final, corrected transcription)
dass Sie auch verdaut haben, wenn Sie neben dem Auto sind.

- **3) Fully correct:** All names, nouns, numbers, and abbreviations are accurately transcribed without any mistakes.
- **2) Minor error** (not affecting key terms)**:** All names, nouns, numbers, and abbreviations are correct. Small grammatical error present (e.g., incorrect conjugation or article).
- **1) Key term error:** At least one name, noun, number, or abbreviation is incorrect in the transcription.
- **0) No relevant excerpts:** The provided excerpt does not contain any relevant content, making evaluation and correction impossible.

Figure 6 presents output of the evaluation stage: 78.0% of transcripts are semantically identical, which means that the context is perfectly reflected in the transcription, after being corrected by GPT-4o.

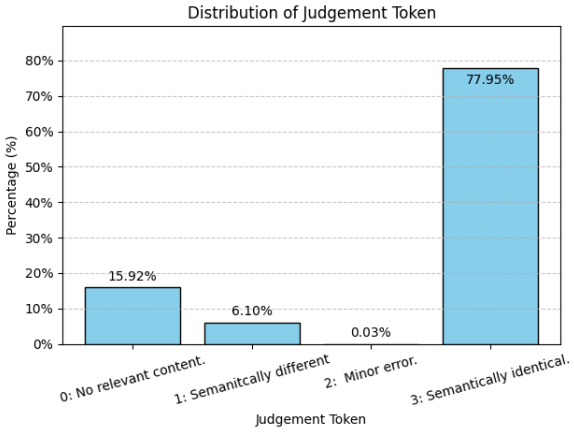


Figure 6: Distribution of the categorization of the final transcription quality using GPT-4o-mini-as-a-judge.

After analyzing 50 minutes of data, we discovered that the judgment category is reliable only when we collapse the label "token 0" into "token 1" and likewise merge "token 2" with "token 3." Grouping the classes this way raises categorization accuracy to 92.2%. Because GPT-4o-mini struggles to decide whether an error is due to missing context or to a genuine semantic change in the transcription, we fuse those tokens for the final data selection.

5 Selecting Data and Train/Test Split

For the construction of the SPC_R high-quality corpus, we combine findings from Section 3.1 (Predicted BLEU) and Section 4.2 (Judgement token) as presented in Figure 7.

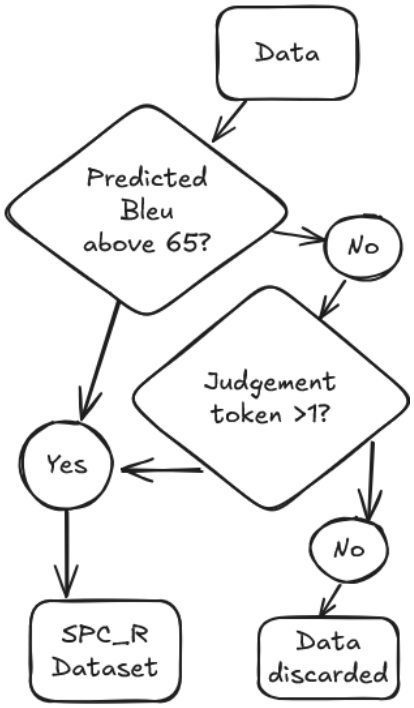


Figure 7: Logic used to build high-quality SPC_R corpus dataset. Size of initial dataset "Data" is 801 hours of audio, size of high-quality dataset "SPC_R" is 751 hours.

We select a Predicted BLEU score threshold of 65 for filtering based on prior research (Cloud) suggesting BLEU score above 60 to be indicative of transcription quality superior to general human levels. By choosing a slightly higher threshold, we reduce the variability indicated by the 95% confidence interval in Figure 2. While this does not guarantee perfect data, (Timmel et al., 2024) shows that imperfect, pseudo-labelled data can improve the quality of ASR models when used in combination with high-quality training data.

This leads to a high quality corpus of 751 hours of Swiss German audio with paired Standard German transcriptions. For the test set, 50 hours are selected with at least a BLEU score of 70 and segments being evaluated as category 3 (as described in Section 4.3). The train/test split is therefore 701/50 hours.

6 Availability and License

The dataset is publicly available on Hugging Face at i4ds/spc_r, the complete codebase (including the prompts) is publicly available on GitHub at i4ds/spc_r.

This dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) License, which allows sharing and adaptation provided that appropriate credit is given and any derivatives are licensed under the same terms.[3]

7 Conclusion

We present SPC_R, transcribed with Whisper Large-v3 on high-compute settings, corrected with context by GPT-4o, and evaluated for quality by GPT-4o-mini. This process results in a corpus of 751 hours of high-quality spoken Swiss German paired with Standard German text.

8 Future Work

There are several promising avenues for further enhancing the Swiss Parliaments Corpus. For instance, incorporating additional data sources beyond the Bernese parliamentary debates could broaden the dialectical and contextual diversity of the dataset, potentially leading to performance and robustness improvements of Swiss German ASR models. Exploring alternative transcription models, especially open source solutions, may offer cost or performance advantages over current approaches based on OpenAI models. Finally, there is also room to work with more nuanced evaluation metrics such as Para_{both} (Paonessa et al., 2023), which better capture semantic fidelity and the accurate transcription of named entities.

9 Limitations

Evaluation Metrics: Our evaluation relies primarily on standard metrics such as BLEU and WER.

[3]For more details, see https://creativecommons.org/licenses/by/4.0/.

These metrics, while useful, do not capture all aspects of transcription quality, as they can be misleading if a sentence conveys the correct semantics using different words, and especially in terms of correctly transcribing named entities, as they don't weight the greater impact of named entity errors on the comprehension of the transcription. In our experience, most of Whisper's errors, which reduce comprehension of the transcription, are now in the named entities, at least in Swiss German.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Google Cloud. Evaluate models | cloud translation. https://cloud.google.com/translate/docs/advanced/automl-evaluate. Accessed: 2025-03-12.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *Preprint*, arXiv:2103.11401.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and pypdf Contributors. 2024. The pypdf library.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jongwook Kim. 2023. Extract confidence. https://github.com/openai/whisper/discussions/1183#discussioncomment-1234567. GitHub Discussion Comment, Accessed: 2025-03-12.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

OpenAI. 2023. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2025-02-28.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Claudio Paonessa, Dominik Frefel, and Manfred Vogel. 2023. Improving metrics for speech translation. *arXiv preprint arXiv:2305.12918*.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Christian Scheller, Yanick Schraner, Claudio Paonessa, Larissa Schmidt, Julia Hartmann, Tanja Samardzic, Manfred Vogel, and Mark Cieliebak. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. In preparation.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. In *Proceedings of the Swiss Text Analytics Conference*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Fine-tuning whisper on low-resource languages for real-world applications. *arXiv preprint arXiv:2412.15726*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

SwissGPC v1.0 - The Swiss German Podcasts Corpus

Samuel Stucki, Mark Cieliebak, Jan Deriu
Centre for Artificial Intelligence
Zurich University of Applied Sciences (ZHAW), Winterthur
stku@zhaw.ch, ciel@zhaw.ch, deri@zhaw.ch

Abstract

We present SwissGPC v1.0, the first mid-to-large-scale corpus of spontaneous Swiss German speech, developed to support research in ASR, TTS, dialect identification, and related fields. The dataset consists of links to talk shows and podcasts hosted on Schweizer Radio und Fernsehen and YouTube, which contain approximately 5400 hours of raw audio. After segmentation and weak annotation, nearly 5000 hours of speech were retained, covering the seven major Swiss German dialect regions alongside Standard German.

We describe the corpus construction methodology, including an automated annotation pipeline, and provide statistics on dialect distribution, token counts, and segmentation characteristics. Unlike existing Swiss German speech corpora, which primarily feature controlled speech, this corpus captures natural, spontaneous conversations, making it a valuable resource for real-world speech applications.

Keywords: low-resource ASR dataset, Swiss German dialects, conversational speech corpus

1 Introduction

Swiss German is a family of dialects spoken in Switzerland and belongs to the Alemannic group of German dialects. It differs from Standard German in phonetics, grammar, vocabulary, and syntax. The dialects vary significantly across regions and are collectively spoken by approximately five million people. Unlike many other dialect groups, Swiss German is widely used in both professional and private settings and additionally serves as an expression and representation of a distinct Swiss nationality in the German-speaking part of the country. While it is primarily a spoken language, the rise of informal digital communication has led to an increase in written Swiss German. However, the absence of standardized orthography and its classification as a low-resource language make data collection for Automatic Speech Recognition (ASR) and other speech processing tasks particularly challenging.

There has been growing interest in the past years in researching ASR tasks on Swiss German dialects, which lead to the creation of several corpora such as the Swiss Parliament Corpus (Plüss et al., 2020), SwissDial (Dogan-Schönberger et al., 2021), SDS-200 (Plüss et al., 2022), and STT4SG-350 (Plüss et al., 2023). These corpora contain between 28 and 343 hours of audio and have since enabled various research endeavours (Sicard et al., 2023; Paonessa et al., 2023; Bollinger et al., 2023; Dolev et al., 2024).

However, these corpora are insufficient for data-intensive tasks such as Text-to-Speech (TTS). This paper thus presents the first version of the "Swiss German Podcasts Corpus (SwissGPC v1.0)", the first mid-to-large-scale[1] corpus for Swiss German: It contains links[2] to talk shows and podcasts collected from Schweizer Radio und Fernsehen (SRF) and YouTube (YT). These collected data contain approximately 5400 hours of raw audio, including speech from all dialect regions and Standard German. We utilized the 7 dialect regions outlined in (Plüss et al., 2023) to simplify the dialect classification. Only the source links of the utilized shows are released, as we do not possess the legal rights to distribute the audio or the annotated data of both SRF and YT.

2 Corpus Requirements

Our primary motivation for creating SwissGPC was to train a Zero-Shot Voice Adaptation Text-to-Speech (TTS) system for Swiss German dialects, for which large amounts of high-quality data are

[1] The dataset can be considered large-scale in the context of Swiss German corpora. However, compared to other languages such as English, German, or Mandarin it is still a small- to medium-sized corpus

[2] Due to copyright reasons, we can not provide the audio files, but only the links to the websites.

required. The dataset was thus created with the following goals in mind:

1. The corpus should be sufficiently large with a goal of 4000-5000 hours of primarily Swiss German speech.

2. The corpus must be sufficiently diverse in speakers to provide useful training data for TTS[3].

3. The speech must be recorded with a high-quality recording setup.

4. The corpus should cover a diverse set of topics.

Based on these goals, we decided to collect a large number of dialogues from podcasts that are primarily in Swiss German and to preprocess them to make them applicable for TTS and other speech processing tasks.

3 Data Annotation Pipeline

As outlined in the introduction, we do not have the rights to distribute the audio. We will only publish the links to the podcast sources that comprise the corpus. For SRF podcasts there exists an official API[4], while for YouTube, a third-party tool can be used such as pytube to download the files (specifically the pytubefix fork (JuanBindez, 2025), as the original library is not maintained anymore). Table 1 and 2 list the podcasts and their online source for SRF and YouTube, respectively. All sources combined link, at the time of publication, to 5404 hours of audio.

The data was weakly annotated using an automated pipeline, visualized in Figure 1. First, the raw audio was diarized and segmented on a speaker basis using pyannote (Bredin, 2023). The diarization step only tags actual speech, leading to silent and music segments being implicitly removed. The samples, containing only a single speaker based on the diarization, were cut to be between 2 and 15 seconds long. The time range was chosen to allow diverse sampling of shorter and longer segments, and additionally due to models downstream that required different lengths of audio for transcription or training. This resulted in a reduction of 7.84% from 5404 hours of raw audio to 4979 hours of actual speech with 1.76M unique samples. The segmented audio was then transcribed to Standard German since it has a standardized orthography. The transcription was performed using whisper-v3 (Radford et al., 2022) for its high performance in translating Swiss German speech to Standard German text (Paonessa et al., 2024). Using the approach of (Bolliger and Waldburger, 2024), we applied a wav2vec2 phoneme transcriber (Baevski et al., 2020; Xu et al., 2022). We classified the generated phoneme sequences with a Naïve Bayes n-gram classifier trained on the phonemicized STT4SG-350 corpus (Plüss et al., 2023) and an additional Standard German CommonVoice (Ardila et al., 2020) subset for Dialect Identification (DID). In total 8 different regions were thus used in classification: Basel, Bern, Central CH, Eastern CH, Grisons, Valais, Zurich, and Standard German. Further enrichment processes included the generation of Swiss German text using (Bollinger et al., 2023) and the creation of Mel-Spectrogram of the audio samples using (McFee et al., 2024).

SRF Podcast Name	Length (h)
#SRFglobal	36.97
100 Sekunden Wissen	186.75
Debriefing 404	245.14
Digital Podcast	428.05
Dini Mundart	39.39
Gast am Mittag	33.14
Geek-Sofa	317.28
SRF-Wissen	45.05
Kultur-Talk	55.84
Literaturclub - Zwei mit Buch	31.79
Medientalk	66.46
Pipifax	9.08
Podcast am Pistenrand	18.29
Samstagsrundschau	404.14
Sternstunde Philosophie	159.39
Sternstunde Religion	60.82
Sykora Gisler	152.22
Tagesgespräch	1661.33
Ufwärmrundi	60.98
Vetters Töne	25.42
Wetterfrage	67.68
Wirtschaftswoche	122.30
Wissenschaftsmagazin	393.61
Zivadiliring	50.03
Zytlupe	44.74
Total	4715.87

Table 1: List of SRF podcasts, links to the source, and hours of raw audio.

[3]Note that this will also be very helpful for downstream ASR tasks.

[4]https://developer.srgssr.ch/

YouTube Podcast Name	Length (h)
Auf Bewährung - Leben mit Gefängnis	3.00
Berner Jugendtreff	127.80
Ein Buch Ein Tee	3.73
expectations - geplant und ungeplant kinderfrei	16.84
Fadegrad	49.95
Feel Good Podcast	319.60
Finanz Fabio	58.44
Scho ghört	23.45
Sexologie - Wissen macht Lust	15.41
Über den Bücherrand	14.53
Ungerwegs Daheim	38.67
Wir müssen reden - Public Eye spricht Klartext	17.52
Total	688.93

Table 2: List of YouTube podcasts, links to the source, and hours of raw audio.

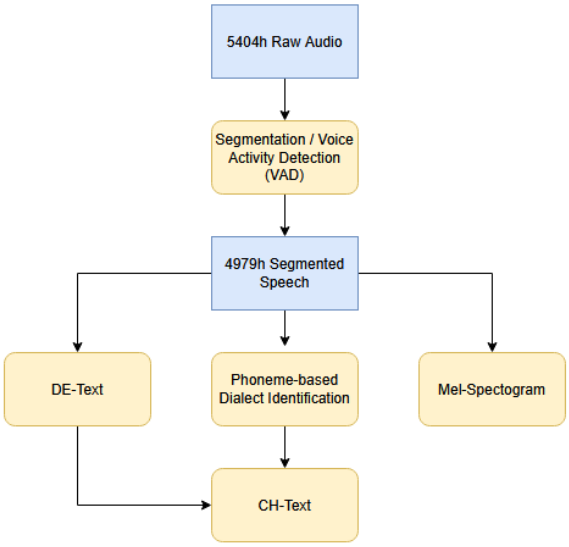


Figure 1: Automated Data Annotation Pipeline

3.1 Pipeline Evaluation

In order to ensure high quality of the automated annotations, we performed an evaluation of the performance of the individual steps of the annotation pipeline. This section presents the evaluation results. The *Zivadiliring* podcast[5] was selected for all evaluations due to its moderate size (approximately 50 hours of raw audio), exclusive use of Swiss German, minimal guest appearances, and the known dialects of its hosts—one from Eastern Switzerland and two from Zurich. These characteristics make it a representative sample of other podcasts.

The diarization was evaluated on a single episode lasting 42 minutes and 38 seconds, using (ela, 2024; Brugman and Russel, 2004) for manual annotation. The diarization pipeline achieved a Diarization Error Rate (DER) of 14.1%, which is comparable to its performance on the AISHELL-4 corpus (Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, 2017), where it reached 12.2%. The Standard German transcription was evaluated by manually transcribing 100 randomly selected audio samples, achieving a Word Error Rate (WER) of 0.30 ± 0.264. Example transcriptions are provided in Table 3. The observed deletions and substitutions can be attributed to the numerous linguistic differences between Swiss German and Standard German. These include the omission of past tense forms, instead preferring the perfect tense, as well as variations in auxiliary verbs, grammatical structures, and the use of Helvetisms or loanwords that either do not exist in Standard German or carry different meanings. Lastly, there is the inherent loss of information when transcribing the audio automatically from Swiss German to Standard German using whisper.

Dialect	Hypothesis	Reference
Eastern CH	Und dann ist quasi die Idee, wenn du als Burning Man gehst, dass du etwas wie einen Provider machst.	Dann ist quasi die Idee auch, dass du, wenn du an das Burning Man gehst, etwas providest.
Zurich	Aber es ist nicht gescheitert. Nein, ich bin ja so hyperemotional. Dann verplatzt es mich und dann bin ich aber wieder ruhig nach vier Sekunden. Aber ich habe dann schon wahrscheinlich ein bisschen umgewettert.	Aber es ist nicht gescheitert an der Wäsche. Nein, ich bin ja schon, ich bin ja so Hyperemotional, dann verplatzt es mich, dann bin ich aber auch wieder ruhig nach vier Sekunden. Aber habe dann wahrscheinlich schon herumgeflucht.
Zurich	Oder was ist er? Weisst du, mit dem Rettchen wüsstest du, über was wir reden. Was ist er gestern gewesen? Was ist er heute?	Oder was ist er? Weisst du damit wir wissen über was wir reden. Was war er gestern? Was ist er heute?

Table 3: Comparison of generated and manual annotated Standard German sentences

The Naïve Bayes classifier from (Bolliger and Waldburger, 2024) was retrained with an additional class for Standard German using the Common-

[5]https://www.srf.ch/audio/zivadiliring

Voice corpus (Ardila et al., 2020). A total of 30 hours of audio was sampled from CommonVoice, ensuring an age and gender distribution similar to that of the phonemicized STT4SG-350 corpus (Plüss et al., 2023), where each dialect region consists of 30 hours of speech. The classifier achieved a macro F1-score of 0.88 across the eight regions. When applied to the *Zivadiliring* episodes, nearly two-thirds of all samples were classified as Zurich and one-third as Eastern Switzerland, aligning with the hosts' origins. Additionally, in an episode where one of the hosts was replaced by a guest from Basel, the classifier correctly identified the samples as Basel.

Lastly, the Swiss German transcription of the same 100 samples used in the Standard German evaluation resulted in a Word Error Rate (WER) of 0.639±0.253. This high error rate is primarily attributed to the lack of a standardized writing system for Swiss German. Example transcriptions are provided in Table 4.

Dialect	Hypothesis	Reference
Zurich	Si känt scho mal din Name, fast. Er isch Content Creator, er isch berüehmt im Internet und er isch super.	Sie kennt scho mal din Name, fast. Er isch Content-Creator, er isch berüehmt im Internet und er isch
Eastern CH	I mein, wa de Onur alles seit. Nur will me zemme wohned isch jetzt nöd de Informationsfluss.	Ich meine, was dä Onur alles seit. Nur will mir zemme wohnet isch ezt do nöd de Informationsfluss.
Zurich	Drum händs so gfunde, ja du bisch irgendwie d'Muetter und denn au irgendwie nöd. Ich glaub, es git nöd die definiert Rolle. Aber ich han so gfunde d'Klaschtante isch no härzig.	Drum hät si d Mueter gfunde, dass si das au nöd gseh hät, dass Klatschstunde no härzig isch.

Table 4: Comparison of generated and manual annotated Swiss German sentences

4 Corpus Statistics

Raw Data. The raw audio is sourced from 25 SRF and 12 YouTube podcasts, comprising 15171 individual episodes with an average length of 1277.28 seconds (21.28 minutes). Episode durations are unevenly distributed, visualized in Figure 2, forming two distinct peaks: one between 100 and 200 seconds and another between 1,600 and 1,800 seconds. The first peak is primarily due to the podcast *100 Sekunden Wissen*, in which hosts provide information about various topics in around 100 seconds. In general, most podcasts produce episodes ranging from 20 to 30 minutes in length, as seen in Figure 4.

The largest podcast is *Tagesgespräch* with 1661.33 hours of raw audio, comprising nearly 31% of the total dataset, clearly visible in Figure 5. On average there are 410 episodes in a podcast, while the median is significantly lower at 104. Outlier episodes ($> 7200s, n = 32$) were typically special episodes, such as yearly recaps, video game playthroughs, or guest interviews. The longest episode in the dataset lasted 13846 seconds (3 hours and 50 minutes), while the shortest was just 19 seconds.

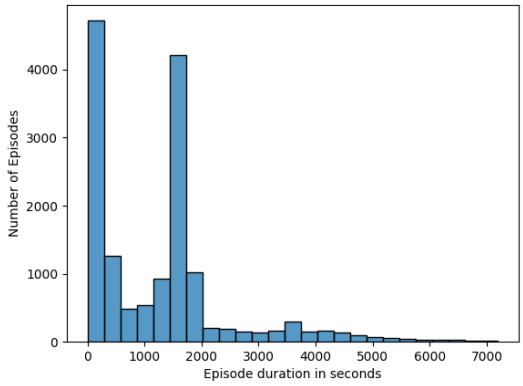


Figure 2: Distribution of episode durations in seconds of all podcasts in the corpus. Outliers ($n = 32$) with length > 7200 seconds are not shown.

Filtering. The filtering step, where we remove, for instance, samples with music only, reduced the data from 5404 hours of audio to 4979 hours of speech, which was segmented into 1.76M samples

Token Counts. After filtering, the data contains 55.85M tokens, calculated using spaCy (Honnibal et al., 2020). The token distribution is shown in 3. Since we did not segment the data on the sentence level, this led to a bimodal distribution of the tokens, visualized in Figure 3, with a large concentration of samples at 15 seconds. Training of a TTS model downstream then led to longer segments being generated better than shorter ones. Future work may improve this. The first peak with token counts of between 7 and 14 could be explained by the characteristics of spontaneous speech in a podcast

setting, in which hosts often interrupt each other in turns or simultaneously, leading to short segments of speech from individual hosts. The second peak with token counts between 40 and 53 can be explained by the generally more information-dense segments in podcasts or shows, where hosts have a monologue telling a story, reading a book or letter, or similar. Additionally, it was found that very short (< 7 tokens) and very large (≥ 65 tokens) samples were often erroneous or incoherent translations by whisper, either due to complex audio or simple mistranslations.

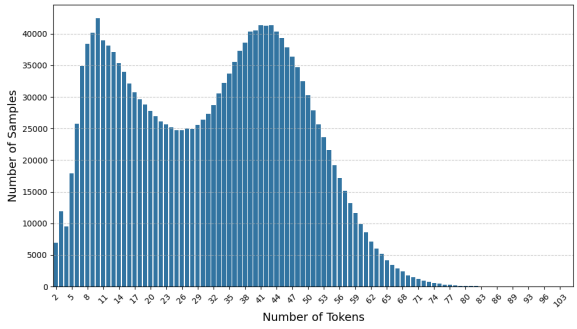


Figure 3: Standard German token distribution of segmented audio samples.

Dialects. At the dialect level, the corpus is highly unbalanced: the two largest regions, Standard German and Zurich, account for 57.53% of audio in the dataset, while the smallest region, Valais, represents only 0.79%. Table 5 provides further insight into the dialect distributions. Additionally, it was observed that Standard German tended to have segments with larger token counts than other dialect regions. This can be attributed to SRF broadcasting more formal and information-dense segments such as science, philosophy, and news programs in Standard German rather than Swiss German. This ensures that all Swiss residents can understand the content regardless of their familiarity with Swiss German dialects. The moderators of these programs are not required to be from Switzerland and could thus originate from Germany, Austria, or another German-speaking region. We hypothesized that the pronunciation of Swiss German speakers using Standard German (i.e., Swiss Standard German) may have a beneficial effect on model training. However, as we are currently unable to distinguish between them, both are grouped under the Standard German label and kept in the dataset.

Region	Samples (K)	Length (h)	% of Dataset	Tokens (M)
Basel	179	460.81	9.25%	5.35
Bern	293	771.38	15.49%	8.98
German	538	1685.72	33.86%	17.23
Grisons	57	151.33	3.04%	1.74
Central CH	121	341.22	6.85%	3.95
Eastern CH	121	350.60	7.04%	4.00
Valais	15	39.46	0.79%	0.43
Zurich	440	1178.58	23.67%	14.13
Total	1767	4979.09	100.00%	55.81

Table 5: Corpus statistics by dialect concerning number of samples, duration, percentage of total duration, and number of tokens.

5 Potential Use Cases

The Swiss German Podcasts Corpus can be a valuable resource for various NLP tasks, particularly for Swiss German. Unlike many existing datasets that focus on scripted or carefully controlled speech, our corpus contains spontaneous, natural, and uncontrolled speech. This makes it particularly useful for real-world applications where speech is often erratic, featuring hesitations, interjections, interruptions, and overlapping speakers. The large size of the corpus and the weak annotation make it particularly useful for weakly supervised learning approaches. An example task where this approach yielded very good results is in *Voice Adaptation for Swiss German dialects* using the XTTSv2 architecture (Casanova et al., 2024). Since the corpus contains a mix of Swiss German and Standard German, it can also serve as an excellent resource for training *Swiss German–to–Standard German machine translation models*. Such models can bridge the gap between spoken dialects and formal written language, enabling better transcription and translation.

6 Corpus Access

SwissGPC v1.0 will be accessible through the Swiss Association for Natural Language Processing (SwissNLP) website. The corpus will include:

1. A comprehensive list of links to all podcasts sourced from SRF and YouTube

2. The code for both downloading any podcast from SRF and YT and the automated annotation pipeline

7 Conclusion

We have presented the Swiss German Podcast Corpus (SwissGPC v1.0), the first mid-to-large-scale Swiss German speech corpus comprising approximately 5400 hours of raw audio (4979 hours of

speech after data cleaning). While the audio can not be released due to licensing concerns, we have provided references to individual podcasts, including an approach for downloading the audio. Additionally, we defined an automated annotation pipeline to weakly label the data for downstream use.

We are convinced that SwissGPC will enable interesting research in the Swiss German speech processing space, and we are excited to see applications utilizing it.

Limitations

The corpus represents a snapshot in time of the selected podcasts. Shows may release new episodes, remove existing ones, change name or location, be discontinued, or be taken offline as a whole. As a result, reproducing the results given here may prove challenging.

SwissGPC v1.0 is highly imbalanced on a dialectal basis, and future work may seek more audio from under-represented regions and add it to the corpus.

The list of podcasts from SRF is not exhaustive, as during the writing of this paper additional podcasts were found that could be utilized. Additionally, it should also be possible to crawl TV shows from SRF, such as *SRF bide lüt*, *Arena*, and more via their website[6] or YouTube channel[7], increasing the size of the corpus further.

References

2024. ELAN (Version 6.8) [Computer software].

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Laura Bolliger and Safiyya Waldburger. 2024. Automatische erkennung schweizerdeutscher dialekte anhand von audiodaten via phonemtranskriptionen. Technical report, ZHAW Zürcher Hochschule für Angewandte Wissenschaften.

Tobias Bollinger, Jan Deriu, and Manfred Vogel. 2023. Text-to-Speech Pipeline for Swiss German – A comparison. *Preprint*, arXiv:2305.19750.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German. *CoRR*, abs/2103.11401.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. In *Oriental COCOSDA 2017*, page Submitted.

JuanBindez. 2025. pytubefix: A fork of PyTube with fixes for compatibility issues. Accessed: 2025-03-11.

Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter, Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad,

[6] https://www.srf.ch/play/tv/sendungen
[7] https://www.youtube.com/@srfdoku

Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef van der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, and Waldir Pimenta. 2024. librosa/librosa: 0.10.2.post1.

Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect Transfer for Swiss German Speech Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15240–15254, Singapore. Association for Computational Linguistics.

Claudio Paonessa, Vincenzo Timmel, Manfred Vogel, and Daniel Perruchoud. 2024. Whisper fine-tuning for Swiss German: A data perspective. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 192–192, Chur, Switzerland. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. *CoRR*, abs/2010.02810.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Clément Sicard, Victor Gillioz, and Kajetan Pyszkowski. 2023. Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 76–83, Neuchatel, Switzerland. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Interspeech 2022*, pages 2113–2117.

A Corpus Statistics

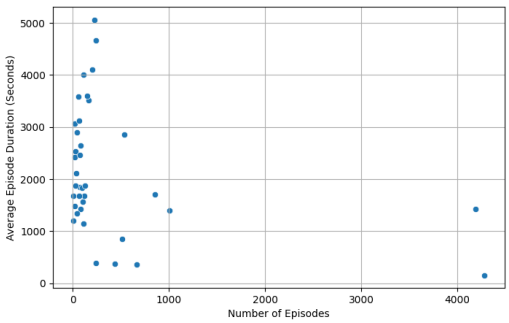


Figure 4: Comparison of total number of episodes in a podcast to the average duration per episode in the podcast.

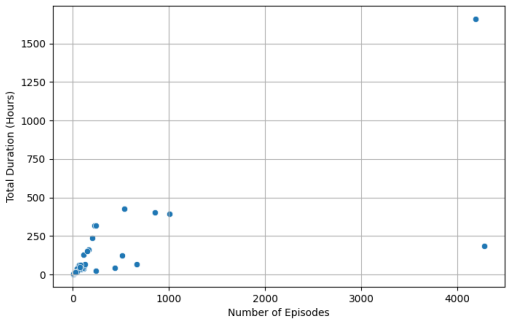


Figure 5: Comparison of the total number of episodes in a podcast to the total duration of all episodes combined in the podcast.

DE-Text	CH-Text	Dialect	Podcast
Die Kosten steigen natürlich dieses Jahr zwischen 6 und 7 Prozent, je nach Leistungserbringer-bereich. Aber letztes Jahr hatte man schon Defizite. Also schon letztes Jahr haben die Prämieneinnahmen die Ausgaben nicht gedeckt. Dieses Jahr wird es noch schlimmer sein. Das nächste	D Kösta stigen natürlich das Johr zwüscha sechs und sieba Prozent, je noch Leistigser-bringerberich. Aber letschts Johr hend d Prämieneinahma d Usgaba nit bedeckt kah.	Grisons	Samstags-rundschau
grossen Deutschschweizer Massenmedien, die noch eine regelmäs-sige Gamekritik gemacht haben. Alle anderen haben das schon viel länger aufgegeben als wir. Und auch bei den Spezialisten, die sich jetzt spezi-fisch für Gamer	Grosse Dütschschwizer Massemedia, wo no e regelmäs-sigi Gameskritik gmacht händ. Alli andere händ das scho vill länger ufgeh wie mir. Und au bi de Spezialiste, wo sich jetzt spezi-fisch für Games.	Zurich	Geek-Sofa
Es war eine Erleichterung, nachdem die UBS angekündigt hat, dass sie auf die staatlichen Garantien verzichtet, die wir im März sprechen mussten. Ohne viel Enthusi-asmus.	Es isch e Erliech-terig gsi, nochdem dUBS aakündigt het, dass sie uf di staatliche Garantie vozichtet hend, wome im März sproche müend.	Eastern CH	Tages-gespräch
nur noch mit Katalysatoren zulassen, so würde man längerfristig den Schadstof-fausstoss massiv beschränken kön-nen.	Nor no met Katalysatore zueloh, so wörd mer längerfrestigi de Schadstof-fusstoss massiv chönne beschränke.	Central CH	100 Sekun-den Wissen

Table 6: Examples of segmented samples in the corpus.

Chapter 3

Impact Track

Abstracts of Impact Track Presentations

Jonathan Gerber, Mark Cieliebak, Don Tuggener, Manuela Hürlimann

Zurich University of Applied Sciences (ZHAW) and
Swiss Association for Natural Language Processing (SwissNLP)
[gerj, ciel, tuge, hueu]@zhaw.ch

1 Assessing the Trustworthiness of Large Language Models on Domain-specific Questions

Sandra Mitrovic, Roberto Larcher and Jérôme Guzzi

Pre-trained Large Language Models (LLMs) can be leveraged to answer domain-specific questions using prompt engineering and retrieval-augmented generation. However, ensuring the trustworthiness of such systems remains a critical challenge. In this work, we propose a general methodology to evaluate the reliability of LLM-based modules by constructing large, representative, and unbiased datasets of questions and answers through automated variation generation. We define key metrics to assess correctness, robustness, and explainability. We apply our approach to a real-world use case in which a smart wheelchair provides answers about its functioning, exploiting RAG with ChatGPT as the underlying LLM. Our experimental results, based on a dataset of over 1,000 questions, reveal that while correctness and robustness are generally strong, the model struggles with open-ended questions, negations, and idiomatic expressions, with explainability being the most challenging aspect. Beyond the specific results (which heavily depend also on dataset at hand), we emphasize the generalizability of our methodology, which can be adapted to various domains. We are currently working on automating the evaluation pipeline to reduce reliance on human assessment and extending the methodology for real-time monitoring of LLM responses.

2 Building commercial GenAI-based solutions: Emerging use cases and best practices

Olmo Barberis and Keibel

Over the past three years, LLMs have impressed the world with their powerful capabilities to understand and generate human language. As with most technological innovations, it takes time and significant efforts to build successful productive solutions (and not just prototypes) around LLMs which generate real-world business revenues and ultimately render the upfront investment profitable. Companies and organizations around the world are still at a fairly early stage in exploring how to best leverage LLMs productively, but some trends and best practices are emerging as to which types of use case are worthwhile to pursue and how LLM-based solutions should be built. In this talk, we will pick up on some of these trends and best practices and sketch out what we believe most commercial GenAI projects might look like in a few years from today. In doing so, we take the perspective of projects which do not have unlimited budget. We will look more closely at some common challenges and ways to mitigate them. We will give some examples from real-life projects that focus on automating business processes.

3 Enhancing Qualitative Content Analysis via LLM Multi-Agent Systems

Norman Süsstrunk, Caroline Dalmus and Albert Weichselbraun

Despite the growing popularity of using large language models (LLMs) such as ChatGPT for qualitative content analysis, current approaches often rely on overly simplistic prompting strategies.

As Mayring (2025) highlights in his field report, the primary issue lies in the inadequacy of many prompt designs. General requests such as "Do a qualitative content analysis according to Mayring" result in superficial outputs that lack adherence to the step-by-step methodology central to rigorous qualitative analysis. Even with more structured prompts, ChatGPT frequently fails to follow essential procedural elements such as inductive category formation, abstraction level calibration, and coder agreement testing. The outcomes typically resemble rough summaries rather than methodologically grounded categorizations, leading to what Mayring refers to as "rough approximations and gross errors". These limitations are further exacerbated when applied to larger datasets or when theoretical grounding is required. Prompt-based approaches, even when refined, struggle to maintain the iterative and transparent logic required by Mayring's qualitative content analysis. As a result, the reliability and reproducibility of the outcomes remain questionable. To overcome these limitations, we propose the integration of multi-agent systems that mirror the structured, procedural logic of Mayring's methodology. Rather than relying on single, monolithic prompts, a system of specialized LLM agents can be deployed, with each agent responsible for a specific task aligned with Mayring's distinct techniques (e.g., inductive category formation, summarization, explication). For instance, individual agents can be designated to handle:

- Category definition
- Calibration of abstraction levels
- Identification and validation of coding units
- Verification of coder agreement

Crucially, these agents would operate under human oversight, ensuring interpretive validity and adherence to ethical and methodological standards. This agent-based architecture is inspired by recent advances in the design of LLM agents, where specialized agents collaborate under human guidance to plan, execute, and optimize complex experiments. By adapting this collaborative structure to qualitative content analysis, we can reflect Mayring's method not only in output, but also in process – step-by-step, transparent, and verifiable. This hybrid system presents a promising way to elevate current practices from surface-level approximations toward structured, scientifically grounded qualitative content analysis.

4 Entity Extraction, Linking, and Disambiguation Pipeline for News Documents

Tsvetan Rangelov, Yannick Suter and Guillaume Comte

At RepRisk we maintain a large database of news incidents coupled to com- panies accused of ESG issues violations. This data is used by asset managers, investors or institutions to make informed decisions about entities they are inter- ested in. This data is multilingual, with a long history, enriched by new compa- nies created every day and combines both human analysis and machine learning. Our pipeline addresses the complex challenge of associating news documents with corporate entities, a critical need for clients who rely on accurate, timely data. Faced with the absence of a single source of truth, duplicate records, and disparate naming conventions—where legal names, journalistic aliases, and outdated entries coexist—we developed a robust, multi-faceted solution. We leverage our unique dataset where texts are associated with IDs to identify entities corresponding to both legal names and the commonly-used variants as well as custom transliteration routines to address our varied multilingual data. Our approach integrates advanced entity extraction with candidate generation, recall-based linking for candidate selec- tion, and precision-based verification for optimal results. To enhance multilingual performance, we incorporate all this contextual information into cutting-edge trans- former models combined with large language models through tailored prompting. This comprehensive system not only resolves data inconsistencies across heteroge- neous sources but also sets a new benchmark for technical rigor and operational efficiency in real-time news content processing.

5 ErrorCatcher: LLM-Powered Editorial Quality Assurance for Reuters News

Luca Malagutti, Guilherme Thomaz and Claudia Schulz

In the fast-paced environment of news production, ensuring editorial quality while maintaining tight publication schedules remains a significant challenge. We present ErrorCatcher, an LLM-powered editorial quality assurance system devel-

oped at Reuters News to help journalists identify and correct both syntactic errors and style guide violations before publication. ErrorCatcher leverages a suite of specialized prompts designed in collaboration with experienced journalists to analyze news articles across multiple dimensions: grammatical correctness, adherence to in-house style guidelines, consistency in terminology, and in-story factual coherence. The system offers targeted feedback, identifying errors and suggesting corrections which reference relevant style guidelines. Our system addresses a significant challenge in integrating sizable organizational style guidelines by developing a hierarchical approach that categorizes style elements by priority and relevance, enabling the system to focus on the most pertinent rules, lowering costs and improving response coherence. We evaluate several leading LLMs as the backbone of our system, revealing that LLMs optimized for complex reasoning demonstrate superior capabilities in identifying subtle style inconsistencies and nuanced grammatical issues across journalistic content. Our preliminary deployment of ErrorCatcher as an internal tool has shown promising results, with journalists reporting improved workflow efficiency and heightened awareness of recurring style issues. We outline our approach to developing ErrorCatcher, discuss the technical and practical challenges of implementing AI editorial assistance in a global news environment, and share our progress in extending ErrorCatcher with additional capabilities while evaluating its performance.

6 Exploring NLP-Driven Personalized Support for Type 1 Diabetes Management: A Preliminary Study

Sandra Mitrovic, Federico Fontana, Andrea Zignoli, Christian Berchtold, Sam Scott and Laura Azzimonti

The widespread availability of wearable devices and sports monitoring applications has enabled individuals, including those with Type 1 diabetes (T1D), to easier track their physical activity. Given the importance of exercise in managing T1D, personalized feedback can play a critical role in optimizing workout routines while mitigating the risks of hypo- and hyper-glycemia. This study explores the feasibility of leveraging Natural Language Processing (NLP) models to generate tailored messages based on an individual's activity data and expert inputs. In particular, we consider two types of workouts: with negative-outcome (i.e., where the individual's glucose level went out of range, further subdivided into hypo- and hyper-glycemia) and with positive- outcome (i.e., where the individual's glucose level remained within the range). Negative-outcome workouts require a behavior change, and messages should advise the individual on how to adjust. Conversely, if the outcome is positive, the individual should be encouraged to maintain their current behavior. Driven by the potential future goal to integrate our approach into an app that prioritizes user privacy and transparency, we focus on evaluating several open-source NLP models to determine their effectiveness in producing high-quality, personalized messages. Furthermore, we consider two types of prompts. First, the simpler one, referred to as the observable prompt type, is based on the combination of a behavioral pattern (i.e., a more precise description of the out-of-range behavior selected from a pre-defined set of possibilities) and its accompanying expert-provided information. Second, the more complex one, referred to as the actionable prompt type, adds to the observable prompt type personalized actionable variables (derived by the underlying ML model1). Additionally, we implemented prompt refinement strategies to enhance message quality and safety, though further research is needed to optimize these approaches. We perform quantitative and qualitative evaluation of prompts. For example, within the qualitative evaluation we focused on prompt adherence, correctness, level of detail, emotional tone, and medical content comprehension. Contrary to expectations, our results reveal that models fine-tuned on medical data or those excelling in medical benchmarks do not necessarily generate superior messages for this application. Among the tested models, Mistral-7B-Instruct-v0.3 demonstrated the most promising performance, while others, including Starling-LM-7B-beta, gemma-2-2b-it, Llama-3.2-3B-Instruct, and JSL-MedPhi2- 2.7B, yielded suboptimal outcomes. This work serves as a proof of concept for the feasibility of using personalized NLP-driven messages in diabetes management, with the ultimate goal of driving behavior change. However, we acknowledge the limitations of our study, particularly regarding dataset size and the narrow scope of actionable variables considered. Future research should focus on expanding the dataset and refining both model selection and prompt engineering techniques to improve the relia-

bility and effectiveness of NLP-generated guidance in diabetes care.

7 GZIP-KNN for ChatGPT Text Detection: A Low-Resource Alternative to Supervised Methods

Matthias Berchtold, Sandra Mitrovic, Davide Andreoletti, Daniele Puccinelli and Omran Ayoub

With the increasing capability of Large Language Models (LLMs) to generate highly plausible and human-like text, the need for reliable AI-generated text detection has become critical. This need is additionally underpinned by recent findings of several studies showing that even adults often struggle to distinguish between human- and machine-authored content. Furthermore, misattributing authorship can lead to the spread of misinformation and the unethical appropriation of text. On the other hand, Transformer-based architectures, which power these models, are highly resource-intensive, adding another layer of complexity to their widespread use. In this study, we investigate the potential of GZIP-KNN, a recently proposed lightweight method, for detecting AI-generated text, specifically content generated by ChatGPT. We evaluate GZIP-KNN's predictive performance, training time, inference time, and memory footprint in comparison to logistic regression, eXtreme Gradient Boosting (XGB), and Gated Recurrent Unit (GRU). As our focus is on low-resource approaches, we do not consider pretrained models. Using five open datasets from different domains, we conduct two experiments. The first examines the trade-off between predictive performance and computational complexity in an in-domain setting. The second assesses performance under data and inference time constraints in an out-of-domain scenario. Experimental results indicate that GZIP-KNN achieves strong predictive accuracy, outperforming alternative methods even with limited data. However, its higher inference time limits its applicability in scenarios requiring rapid decision-making. Nonetheless, findings suggest that GZIP-KNN can match the performance of other methods when trained on only a small subset of available data in an out-of-domain context.

8 Presenting LLMs' collective intelligence approach for Multilingual Hallucination Detection

Sandra Mitrovic, Joseph Cornelius, David Kletz, Ljiljana Dolamic and Fabio Rinaldi

Hallucinations pose a crucial problem in the utilization of large language models (LLMs). The problem is even more pronounced as literature lacks the standardized definition of hallucinations. Furthermore, different LLMs may identify different parts of the same text as hallucinations and in general, different LLMs have different hallucination rates. The problem of identifying hallucinations is even more complex in the multilingual setup. In this study we present our approach to multilingual hallucination detection, as part of MuSHROOM ("Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes"), a SemEval-2025 Task-3. This task is complex as it consists in both detecting exact hallucination spans and determining the hallucination probability. Moreover, the task covers 14 different languages and provides no labeled data, apart from several validation instances for 3 different languages. Task used two evaluation metrics: intersection-over-union (IoU) and correlation (Corr). We tackle this problem simulating the original annotation process that uses multiple artificial annotators. Each artificial annotator is instantiated through a different LLM service combined with varying prompts. Subsequently, the outputs of individual annotators are aggregated into a single annotation using as final hallucination probability the ratio of annotators that denoted the span as a hallucination. We use six different LLM APIs and three different prompts, and we experimented also with different merging variants. Our approach shows great potential as it, in terms of IoU, scored 4th for French (out of 30 teams), 5 for Italian (out of 28 teams), 12 for English (out of 41 teams), and 15th for German (out of 28 teams). In terms of Corr, the results were even better as we ranked 1st , 3rd , 4th and 7th for English, German, French and Italian, respectively. Beside the quantitative results, where we established which models and prompts perform the best, we also performed extensive qualitative analysis, looking deeper in different aspects of differences between published ground truth and our system annotations.

9 Public Unveiling ESG Insights in Real-Time: A Live Demo of RepRisk's ML Pipeline

Guillaume Comte, Tsvetan Rangelov and Yannick Suter

At RepRisk, a leading ESG data provider, we harness the power of Natural Language Processing (NLP) and Machine Learning (ML) to extract critical ESG insights from news articles worldwide. Our advanced ML pipeline processes vast amounts of unstructured text to identify key ESG-related events, assess company involvement, and generate structured, actionable insights.

In this live demo, attendees will have the opportunity to select news articles of their choice, which will then be processed in real time through RepRisk's multi-stage ML pipeline. The system will extract ESG-relevant information, classify incidents, map companies to their identifiers, and generate predictive insights, all displayed dynamically in our interactive UI. Each prediction and extracted entity will be clickable, allowing users to explore related incidents and navigate company profiles directly on RepRisk's platform.

This session will not only showcase the sophistication of RepRisk's NLP-driven ESG analytics but also allow participants to experience firsthand the accuracy and depth of our AI models in transforming raw news into meaningful ESG intelligence.

10 RAG vs Long-Context LLMs: Choosing the Right Approach for NLP Applications

Elena Nazarenko

The landscape of Natural Language Processing (NLP) has been dramatically reshaped by the rise of Large Language Models (LLMs). Two key architectural approaches have emerged to address the challenges of integrating external knowledge and processing large volumes of text: Retrieval-Augmented Generation (RAG) and Long-Context LLMs. RAG systems excel at incorporating external knowledge sources into the generation process. By retrieving relevant documents or passages based on user queries, RAG enables LLMs to provide contextually accurate and up-to-date responses, mitigating the limitations of pre-trained models. This approach significantly expands an LLM's access to vast amounts of information at minimal cost. It is particularly valuable in applications requiring access to dynamic or proprietary information, such as question answering over internal knowledge bases, document summarization, and personalized recommendations. Conversely, recent models like Gemini 2.0, Claude 3, and GPT-4.5, with extended context windows (120K–2M tokens), have demonstrated exceptional capabilities in processing extensive text directly. This eliminates the need for external retrieval, potentially simplifying architecture and reducing latency. These models excel in scenarios where the entire relevant context is available, such as analyzing legal documents, processing scientific papers, or handling complex narrative tasks. However, the choice between RAG and Long-Context LLMs is not always straightforward. RAG systems introduce complexities related to retrieval quality, indexing, and latency, while Long-Context LLMs face challenges with computational cost, potential information dilution, and the "needle in a haystack" problem. This presentation aims to provide practical insights and guidance for NLP practitioners, enabling them to make informed decisions when selecting the most appropriate approach for their specific applications.

11 SYMBOL - Neurosymbolic AI for explainable and reliable AI in high-stake environments

Albert Weichselbraun and Norman Süsstrunk

Lack of explainability and reliability (e.g., due to hallucinations, misleading information, or biased outputs) are serious obstacles towards the adaptation of LLMs in high-stake environments. SYMBOL tackles this shortcoming by developing neurosymbolic AI models that combine embedding-based language models (sub-symbolic processing) with machine-readable domain knowledge (symbolic reasoning), organized in knowledge graphs. The project aims at bridging the semantic gap between user queries, the company's information systems (e.g., databases, customer relationship management systems, and software APIs) and its knowledge management infrastructure (e.g., domain ontologies, structured knowledge in databases and knowledge graphs, and corporate knowledge repositories). LLMs interpret user queries and translate them to the corresponding concepts in the knowledge graph. This enables processing of queries using symbolic AI which ensures very high reliability, since reasoning within symbolic AI components is deterministic. Symbolic reasoning upon domain-specific knowledge graphs also explains

query results and decisions based on (human understandable) concepts within these graphs, ensuring that system decisions are traceable and explainable to non-computer scientists. Once completed, SYMBOL will support clients in the wealth management industry by

- navigating and aiding users through complicated regulatory requirements;
- generating regulatory reports and analyses on demand, helping wealth management firms respond to audits, risk assessments, and evolving compliance mandates; and
- extracting deep business insights from their data, enabling proactive decision-making based on structured, regulatory-compliant intelligence.

By allowing non-technical users to interact with SYMBOL the project will eliminate barriers to data-driven decision-making, ensuring that compliance officers, portfolio managers, and auditors can extract the necessary information when it matters most – such as to support high-stake decision-making processes and on-site regulatory reviews. Although the wealth management use case is central to the SYMBOL project, we aim at adapting the developed neurosymbolic AI components to other high-stake environments in domains such as finance, medicine, and law.

12 Scaling RAG from Pilot to Production: Evaluation, Software practices and Safety

Louis Douge and Robert Simmen

We developed and deployed Life Guide Scout, a GenAI-powered underwriting assistant to more than 3,000 Life & Health underwriters worldwide. The system uses a Retrieval-Augmented Generation (RAG) setup to integrate Swiss Re's proprietary underwriting guidance and medical knowledge, thereby speeding up information retrieval. Fully integrated in the underwriter's workflow, it enables intuitive, efficient and trustworthy interactions with highly specific knowledge. The real challenge of productively deploying an LLM-based system lies in assessing its performance over time and across versions. We present a comprehensive evaluation methodology based on synthetically generated data. For instance, on the specific task of mentioning the right underwriting rating in Life Guide Scout's answer, we achieve an end-to-end 80% hierarchical recall, a metric particularly suited to our problem. We also examine the various failure modes and suggest mitigations. In addition to this programmatic approach, human feedback played a crucial role in refining Life Guide Scout through three key approaches: expert evaluations for structured assessments, user feedback within the application for real- time insights, and surveys and interviews to gauge adoption trends. This multi-layered approach ensured continuous iteration, improving accuracy, usability, and overall user satisfaction. Developing GenAI applications also requires a blend of new and traditional engineering practices. We share insights on prompt management techniques, structured outputs, and strategies for handling frequent LLM updates, including new models and versions. While LLMs introduce novel challenges, traditional software engineering practices remain critical. We detail unit, integration, and regression testing methods, which are essential for iterating on an LLM- centric application in a production environment. Given the risks of incorrect AI-generated outputs in an insurance context, we implemented pre- and post-processing techniques to reduce inaccuracies by leveraging the specificities of our problem. Enhancing transparency, we introduced source anchoring using IDs, which not only links references but also highlights the exact section or phrase within the source that the LLM used to generate its response. This improves user trust and allows for quick verification of information. GenAI introduces new risks related to safety and security. We conducted extensive adversarial attacks, or Red Teaming, on Life Guide Scout to uncover vulnerabilities and proactively mitigate risks, ensuring alignment with responsible AI principles. By stress-testing the system against adversarial scenarios, we strengthened safeguards, improving both security and reliability. Finally, we share our approach to developing conversational memory within a RAG setup while managing token usage effectively. Maintaining context across interactions enhances the user experience but presents engineering trade-offs that we addressed through targeted optimizations.

13 SetFit for Automated Essay Scoring: Extending Longformer to a Sentence Transformer

Leon Krug, Jannik Bundeli, Jannine Meier and Elena Nazarenko

Automated Essay Scoring (AES) demands models that can evaluate student essays with human-like consistency while maintaining computational efficiency. Although standard transformer models like DeBERTa can achieve strong performance, they are often resource-intensive and constrained by a 512-token input limit, which can lead to truncated context in longer essays. This limitation hinders the model's ability to capture argument flow, coherence, and global structure, which are crucial for accurate scoring. Additionally, many existing approaches also rely on prompt engineering, further restricting practical application. To address these challenges, we present a novel prompt-free approach using SetFit for AES that achieves competitive accuracy while significantly reducing computational overhead. Unlike traditional transformer-based models, SetFit enables sentence transformer fine-tuning with contrastive learning, making it suitable for essay scoring even in low-data regimes. We extend Longformer into a sentence transformer, allowing SetFit to process full-length essays within a 4096-token window. This overcomes the 512-token restriction of traditional transformers, ensuring that the model can evaluate entire essays rather than isolated sections. Our approach integrates SetFit's lightweight contrastive learning to optimize sentence embeddings, enabling efficient, prompt-free fine-tuning with significantly lower GPU requirements compared to full transformer fine-tuning. By using contrastive learning, our model learns rich representations of essay quality without needing large-scale labeled datasets. We train our model on AES-specific datasets, so it captures the complexity of essay evaluation metrics such as coherence, grammar, and argumentation strength. Our fine-tuned model has been publicly released on Hugging Face, where it has already gained over 6,000 downloads, reflecting strong community interest in efficient, long-text NLP solutions. Our results show that SetFit with an extended Longformer sentence transformer achieves competitive accuracy and offers a cost-effective, scalable alternative to resource-heavy methods. Beyond essay scoring, our approach is applicable to other long-form NLP tasks, including legal document analysis, research paper assessment, and educational content evaluation, providing a cost-effective alternative to computationally expensive transformer-based models.

14 Transforming Healthcare Documentation: Efficient AI-Powered Automation of Clinical Discharge Summaries for Inpatients

Chantal Zwick, Joseph Weibel, Daniel Olivier Perruchoud, Tristan Struja and Felice Burn

Large language models (LLMs) are widely used to speed up administrative processes across industries. In the medical sector, physicians spend up to 2/3 of their work time with administrative tasks. LLMs could substantially alleviate this burden, allowing for more time with patients. Given the complexity of summarizing information from multiple sources and the sensitivity of content contained in medical documents, LLMs need to be deployed with the utmost scrutiny on local hardware. We therefore assessed the quality and thoroughness of discharge notes generated by locally hosted state-of-the-art LLMs compared to human-written notes. Methods: History of present illness (HPI) as well as diagnoses and procedures (DXL) were extracted from patient records for three clinical scenarios: planned or elective chemotherapy (PEC), acute coronary syndrome (ACS), i.e. myocardial infarction, and acute lower back pain (ALBP). Three medium-sized LLMs, i.e. Mixtral 8x7B, Mixtral 8x22B and Llama 3.1 70B were prompted to generate discharge sum- maries based on HPI and DXL inputs. Three approaches of generating discharge notes were compared: prompting without examples (zero-shot approach), In-Context Learning (ICL) which utilized four examples of triplets consisting of HPI, DXL, and humanwritten discharge summaries (4-shot approach), and supervised fine-tuning (SFT) on Mixtral 8x7B with specific training sets (NP EC-train = 1028, NACS-train = 1920, NALBP-train = 1494). For evaluation, five simple and five complex samples were extracted for each of the three scenarios, resulting in 30 triplets of HPI, DXL and human-written dis- charge summaries. Using the different LLMs and different prompting approaches, this results in a total of 150 generated discharge summaries, which were assessed via BLEU, ROUGE-L, and BERTScore metrics. In addition, a blind panel of 6 specialists in internal medicine assessed the 150 summaries with a modified Physician Documentation Quality Instru-

ment (mPDQI-9) consisting of nine items rated on a 5-point Likert scale, with higher scores indicating better performance. Results: Our findings indicate that both ICL and SFT enhance the quality of the generated discharge summaries compared to the zero-shot approach. The improvements were most notable for SFT in the PEC scenario (median 32 vs 28 out of 45). In general, generated reports for simpler cases received higher human ratings compared to more complex cases, particularly for the PEC scenario, but hallucination was a problem. When benchmarked against their respective ground truth discharge summaries, we achieved a BERTScore of 0.75, a BLEU score of 0.18, and a ROUGE-L score of 0.35 for the simple cases with SFT, which was the best approach. Overall, zero-shot Mixtral 8x7B, 8x22B, and Llama 3.1 70B demonstrated similar performance based on the expert panel's assessment. Conclusion: Our findings demonstrate that LLMs create medical discharge summaries for simple clinical scenarios with acceptable quality, but struggle with more complex cases. This highlights the need for accurate prompting, technical solutions to hallucination, and high quality input data in training models. Addressing these challenges would alleviate much of the administrative burden for physicians, especially those in training, which currently spend only 30 % of their workdays directly with patients. This approach has the potential of enhancing workflow efficiency, reducing clinician burnout, and improving

15 Unlocking Model Potential: A Comprehensive Framework for Feature and Data Enhancement

Xavier Ferrer, Alessandro Caruso and Claudia Schulz

In the dynamic landscape of machine learning, optimizing model performance relies on a thorough analysis of feature spaces. This study introduces an innovative framework designed to refine and improve machine learning models through meticulous feature analysis. We explore the correlations between the features and the model predictions to identify areas of improvement and potential feature gaps. By targeting misclassified samples, we uncover patterns that may elude conventional models, enabling us to propose targeted adjustments in model architecture and feature engineering. We leverage SHAP (SHapley Additive exPlanations) analysis together with unsupervised learning techniques, such as PCA or t-SNE, to reveal nonlinear relationships and natural data groupings based on feature vectors. Furthermore, we employ K-Nearest Neighbors (KNN) and cluster analysis to detect annotation errors by identifying homogeneous feature vector clusters and to enhance data integrity by flagging potential misannotations for review. We applied the proposed framework to an entity matching project, where text-based features are compared between different documents to identify matching pairs. This approach allowed us to identify the limitations of our models and guide the creation of new features specifically designed to distinguish between samples with very similar feature vectors but different annotations. Clustering analysis also helped identify and correct erroneous annotations in the dataset, resulting in a significant improvement in model performance. Our framework not only identifies and corrects model weaknesses, but also proposes strategies to build more robust, accurate, and interpretable models, ultimately advancing their applicability in real-world scenarios. Although tailored for NLP challenges, the framework is also applicable beyond NLP for any feature-based ML model. This study serves as a guide for data scientists and machine learning practitioners seeking to optimize model performance through comprehensive feature analysis and enhancement techniques.

16 "Radikale Diskurse lichten" Automated Telegram monitoring for analysis & research

Lars Schmid

The RaDisli ("Radikale Diskurse Lichten") project introduces an automated, dynamic monitoring tool that systematically collects and analyzes extremist content from Telegram channels. The prototype leverages advanced NLP techniques to provide real-time analytical insights into radical discourse, speci cally supporting monitoring and analytical e orts within social work. Key features include individual ltering by channels, time range, and search terms. Each message undergoes automated classi cation into categories: "hate speech," "toxicity," "threat," and "extremism." The Streamlit-based web application visualizes activity patterns through heat maps, highlighting peak communication times. Word clouds summarize frequently used terms per channel or group, and topic modeling via Latent Dirichlet Allocation (LDA)

provides insights into prevalent themes within the discourse. Additionally, a network graph visualizes interconnections between channels based on forwarded messages, highlighting in uential hubs and dissemination pathways. Evaluations of the prototype indicate that the application signi cantly enhances analytical capabilities. Users report that the streamlined, image-free interface reduces emotional stress and allows for a more objective, neutral assessment of extremist content compared to direct interaction within Telegram. To date, the system has processed and analyzed over 3.1 million messages from more than 180 channels, demonstrating robust scalability and performance.