A Benchmark and Evaluation of Automated Language of Study Extraction from Computational Linguistics Publications

Henry Gagnier[*]
Pittsford Sutherland High School
Pittsford, New York, USA
henrygagnier9@gmail.com

Ashwin Kirubakaran[*]
Edison Academy Magnet School
Edison, New Jersey, USA
ashwinkiru10@gmail.com

Abstract

Language of study is an aspect of computational linguistics papers that is useful for analyses of trends and diversity in computational linguistics. This study introduces the first benchmark and evaluation of automated language of study extraction from computational linguistics publications. The benchmark containing 431 publications from the ACL Anthology, with 62 languages analyzed, was annotated. SciBERT and four large language models (LLMs), GPT-4o mini, Gemini 2.5 Flash, Claude 3.5 Haiku, and DeepSeek 3.2, were evaluated on the benchmark using different parts of the ACL Anthology papers. GPT-4o mini achieved the best exact match and Jaccard agreement scores of 0.646 and 0.687, respectively, which is slightly less than the agreement in human annotation. Gemini 2.5 Flash achieved the best micro F1 of 0.633. Models using the abstract for extraction were competitive with models using the full text, showing that accuracy can be achieved in language of study extraction without high computational costs. These findings demonstrate that LLMs are able to accurately identify the languages of study in computational linguistics papers, potentially reducing the time and cost of analyses in computational linguistics.

1 Introduction

Large language models (LLMs) have shown immense potential for information extraction in scientific texts in recent years (Cheung et al., 2023; Dagdelen et al., 2024; Dunn et al., 2022; Xu et al., 2024; Jami et al., 2024), although their use for language of study extraction remains largely unexplored. Language of study extraction is the extraction of the languages that are analyzed within computational linguistics papers, which allows for diversity and trends in computational linguistics and natural language processing to be accurately analyzed, often done manually or using methods with limited accuracy (Held et al., 2023; Joshi et al., 2020).

In the past, language of study extraction has involved manual surveys of papers (Bender, 2009), which may be time-consuming and inherently costly. Bender (2011) introduces the "Bender Rule," calling for researchers to state the name of the languages they study, especially when the language is English. Ducel et al. (2022) found that only half of ACL papers respect the Bender Rule, introducing difficulty to language of study extraction and studies of trends and diversity in computational linguistics, especially for the English language. The language of study of computational linguistics papers is important and often not explicitly or clearly stated, making it difficult to easily extract.

Work has begun to emerge using and proposing new methods to gauge the inclusion of languages in computational linguistics and natural language processing. Joshi et al. (2020) measures language diversity and inclusion in computational linguistics conferences through frequency-based techniques, measuring the mentions of languages in scientific papers. Schwartz (2022) uses a similar method, which uses the number of times a language is mentioned in ACL paper abstracts. Held et al. (2023) presents a method where the plain text is searched for mentions of languages. Then, GPT-3.5-Turbo, an LLM, was used to filter these sentences through few-shot prompting to remove languages mentioned in passing or as homonyms to analyze coloniality in natural language processing.

Previous methods have flaws that may limit the accuracy of the language of study extraction and not effectively convey the frequencies at which languages are studied. The method presented in Joshi et al. (2020) may be impacted by homonyms and mentions of languages that do not represent contributions to the research and natural language processing of the mentioned languages. The method

[*]These authors contributed equally to this work.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 366–374
March 26, 2026 ©2026 Association for Computational Linguistics

in Held et al. (2023) may also experience significant error, as the usage of select sentences rather than the full text for filtration may limit the understanding and accuracy of information extraction by GPT-3.5-Turbo. Recognizing the importance of this work, which studies trends in computational linguistics, it is shown that the determination of the language of study in computational linguistics papers is vital.

No study or benchmark exists to evaluate the accuracy of machine learning models to extract the language of study from computational linguistics papers. The purpose of this study is to (1) construct a reliable benchmark for language of study extraction, (2) explore the usage of large language models for language of study extraction using varying input, and (3) identify problems and challenges for language of study extraction. This work aims to construct the first benchmark for language of study extraction and find models and inputs that are able to most accurately classify language of study to improve the efficiency and accuracy of studies analyzing trends in computational linguistics.

2 Methodology

2.1 Benchmark Construction

The title, abstract, and full text of computational linguistics publications were acquired from ACL OCL (Rohatgi et al., 2023). ACL OCL is a scholarly corpus derived from the ACL Anthology, which is the prime resource for research papers in computational linguistics and natural language processing, maintained by the Association for Computational Linguistics. As of 2023, the ACL Anthology hosted 88,000 papers, with 3,000 non-English papers (Bollmann et al., 2023). In this study, papers not in English were excluded using langdetect v1.0.9. Papers were randomly sampled from ACL OCL to ensure diverse coverage across venues and years.

The final benchmark included 431 papers studying 62 languages. In order to facilitate future research in LLMs for information extraction and continue research on language of study extraction, the final benchmark is available publicly at https://github.com/henrygagnier/language-of-study-extraction-benchmark.

2.2 Inter-Annotator Agreement Statistics

A dataset of 431 papers was independently labeled by two annotators who were both native speakers of American English. Each annotator independently labeled the full dataset. Overall, inter-annotator agreement was 73.2% for exact matches, where annotators identified identical sets of a paper. Jaccard similarity was 0.816, indicating overlap in identified classes in cases with disagreements. After independent annotation, disagreements were found and resolved through discussion, yielding the final benchmark. Annotator guidelines are provided in Appendix A.

2.2.1 Overall Agreement

We present the agreement between the two annotators after independently labeling the benchmark with the languages of study (Table 1). We report Krippendorff's , Cohen's , exact match agreement, and Jaccard similarity. Overall, the agreement values indicate substantial consistency between the annotators.

Metric	Value
Krippendorff's α	0.778
Cohen's κ	0.779
Exact match agreement	0.732
Jaccard similarity	0.816

Table 1: Overall inter-annotator agreement statistics

2.2.2 Per-Language Agreement

We show the per-language agreement statistics, for languages with at least 10 positive annotations, excluding languages with low sample amounts due to their unreliability (Table 2). Agreement varies among languages, potentially reflecting ambiguity in language identification. While varying, the agreement was consistently moderate to high, ranging from 0.565 to 0.909 in the selected languages, and suggesting that the annotation guidelines were consistently applied across all languages.

2.3 Benchmark Language Distribution

Including 62 languages of study, the benchmark introduced represents many high-resource languages such as English and Mandarin, and many low-resource languages such as Ewe and Scottish Gaelic (Wiafe et al., 2025; Klejch et al., 2025), as the first benchmark for language of study extraction. Figure 1 visualizes the distribution of language analyzed by the papers in the final benchmark, displaying a large amount of studies on the

Language	n	α	Language	n	α	Language	n	α
English	204	0.693	Chinese	58	0.832	German	42	0.705
Japanese	43	0.840	French	36	0.671	Arabic	35	0.861
Portuguese	32	0.909	Spanish	31	0.771	Russian	20	0.741
Hindi	16	0.893	Korean	15	0.627	Croatian	12	0.907
Turkish	12	0.796	Indonesian	11	0.773	Czech	10	0.746
Italian	10	0.565						

Table 2: Per-language Krippendorff's α for languages with ten or more positive annotations

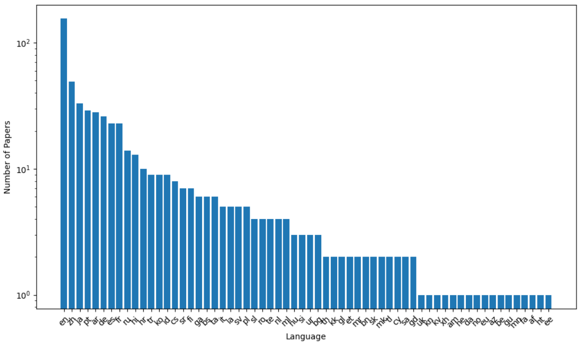

Figure 1: Distribution of the languages of study of the papers included in this benchmark on a logarithmic scale

English language, reflecting the current state of diversity in computational linguistics (Bender, 2009).

2.4 Language Models

2.4.1 SciBERT

SciBERT is a pretrained language model for scientific text, performing better than BERT on science-related tasks, trained on the full text of 1.14 million papers from Semantic Scholar (Beltagy et al., 2019). We use SciBERT as a baseline in this study and train SciBERT using a 70:15:15 train-test-validation split for multilabel classification with 3 epochs. SciBERT was evaluated on languages of study that had more than 20 studies in the benchmark to ensure sufficient training data, which were English (en), Portuguese (pt), Mandarin (zh), Arabic (ar), Japanese (ja), German (de), French (fr), and Spanish (es). To ensure comparability, the LLMs were evaluated on the same test set and performance metrics as SciBERT. We performed per-class threshold tuning to optimize the F1-score for each label independently, searching from 0.05 to 0.95 in increments of 0.01 to find the threshold that optimizes the F1 scores on the validation and training sets.

2.4.2 Large Language Models

We evaluated four large language models (LLMs) on an identical subset of the benchmark that SciBERT was evaluated on, evaluating GPT-4o mini (OpenAI et al., 2024), Gemini 2.5 Flash (Comanici et al., 2025), Claude 3.5 Haiku (Anthropic, 2024), and Deepseek 3.2 (DeepSeek-AI et al., 2025). We selected these models to represent a diverse set of recent large language models that are publicly accessible and optimized for practical use. In all large language model prompts, the annotation guidelines (Appendix A) were supplied. In all models, the temperature was set to 0 to ensure deterministic outputs and reproducibility across all experiments.

3 Results

We now look at the overall performance of the four LLMs and SciBERT on the test set using both the full text and the abstract, or the abstract only, to generate predictions (3). Overall, GPT-4o mini using the full text had the best exact match and Jaccard similarity of 0.646 and 0.687, respectively, and Gemini 2.5 Flash had the best micro and macro F1 scores of 0.633 and 0.622, respectively.GPT-4o mini achieved an exact match accuracy of 0.646, approaching the inter-annotator exact match rate of 0.732, showing that GPT-4o mini had a strong performance when compared to human annotations. GPT-4o mini, using only the abstract, had an exact match agreement of 0.600, or 0.046 less than when the model was using the full text. Model performance surprisingly increased or remained the same in Gemini 2.5 Flash, Claude 3.5 Haiku, and DeepSeek 3.2 when the full text was removed. Using only the abstract is much more computationally efficient than using the full text and provides comparable results.

Model performance was extremely varied, with GPT-4o mini having a precision of 0.735 and a recall of 0.463, while Claude 3.5 Haiku had a precision of 0.362 and 0.778, using the full text. Exact match agreement ranged from 0.262 to 0.646, and micro F1 ranged from 0.494 to 0.633. Different LLMs classify the language of study completely differently, with GPT-4o mini underclassifying and Claude 3.5 Haiku overclassifying. Before use for language of study extraction, models should be tested for their tendencies to misclassify and their accuracy, which was extremely variable among models. As expected, SciBERT has a much worse performance than all LLMs, with low precision and high recall. To better understand areas where the model struggles, we now look at language-specific results (Table 4) and qualitative error analysis. Model performance ranged widely throughout each language, with the best F1 scores being from 0.40 to 0.83. Many models had difficulty with English, likely due to its status as a common language of study in computational linguistics. We conduct a qualitative error analysis (Appendix D) and find that all models label papers that do not study English as otherwise. The confusion in the English language can likely be partially explained by model misinterpretation and hallucination. In other languages, models may be focused on the themes of papers without critically examining the actual experimentation of the paper, and some papers may require an amount of reasoning that the models cannot perform.

4 Discussion

This study evaluates LLMs and SciBERT for language of study extraction from computational linguistics publications. We construct the first benchmark for language of study extraction with high agreement and codify the practice of annotating which languages a paper uses, and using the annotations for analysis. Evaluating models, we find that LLM outputs align with the human-annotated labels and demonstrate significant potential for language of study extraction. Models reached exact match and micro F1 scores of 0.646 and 0.633, respectively. These results directly display that LLMs can effectively automate the extraction of the language of study in computational linguistics papers, allowing for a scalable and timely solution for analyzing trends in computational linguistics research.

This study's results align with previous natural language processing research, showing the effectiveness of LLMs for scientific document parsing and structured information extraction (de Haan et al., 2025; Nguyen et al., 2023; Dagdelen et al., 2024). While SciBERT was pretrained on scientific text (Beltagy et al., 2019), likely enhancing its performance on scientific papers and scientific literature-related tasks, such as language of study extraction, it performed poorly as a baseline. Transformer models may be less scalable for language of study extraction, as they require training data, which may not be available for understudied languages in computational linguistics.

Multiple unexpected findings were found in this study. When LLMs were prompted only using the abstract of the paper, performance dropped marginally, and in some cases, increased. Using the abstract of a paper significantly decreases the number of tokens used, decreasing the computational cost of model usage. Excluding the full text may have decreased the complexity of the prompt, limiting LLM hallucination. In cost-limited studies, using the abstract and LLMs is a solution to accurately extract the language of study without high computational costs. Model performance was also extremely variable across LLMs, with some models grossly underclassifying and some models grossly overclassifying. F1 and accuracy were also variable across models. If LLMs are to be used for language of study extraction, models must be tested on high-quality benchmarks, such as the one presented in this study, in order to evaluate model biases and performance. We also found that many LLM errors are not caused by misclassification of homonyms or mentions of languages in passing. LLM errors are largely caused by biases towards English, misinterpretation of the text, and a lack of ability for complex reasoning. To further improve model accuracy, future work should be performed.

Few-shot prompting, prompt engineering, context engineering, majority voting, agentic workflows, and other ablations should be tested, as many misclassifications may come from a misunderstanding of the provided paper's text or the annotation guidelines. More open source models should be included in evaluations, such as Qwen or Llama. Future work should also expand benchmarks to larger and more diverse samples of computational linguistics papers, enabling evaluation on low-resource languages. Additionally, studying temporal trends in computational linguistics using the automated extraction techniques presented in this paper may provide valuable and more accurate insights than previous studies into topics like priorities and rep-

Model	Exact Match	Jaccard	Micro F1	Macro F1	Micro Prec.	Micro Rec.
GPT-4o mini (full text + abstract)	**0.646**	**0.687**	0.568	0.555	**0.735**	0.463
Gemini 2.5 Flash (full text + abstract)	0.477	0.525	**0.633**	**0.622**	0.576	0.704
Claude 3.5 Haiku (full text + abstract)	0.262	0.409	0.494	0.485	0.362	**0.778**
DeepSeek 3.2 (full text + abstract)	0.492	0.551	0.625	0.619	0.603	0.648
SciBERT (full text + abstract)	0.077	0.251	0.365	0.388	0.242	0.741
GPT-4o mini (abstract only)	0.600	0.623	0.488	0.509	0.714	0.370
Gemini 2.5 Flash (abstract only)	0.477	0.533	0.627	0.620	0.578	0.685
Claude 3.5 Haiku (abstract only)	0.277	0.399	0.491	0.474	0.363	0.759
DeepSeek 3.2 (abstract only)	0.523	0.594	0.632	0.594	0.600	0.667
SciBERT (abstract only)	0.000	0.134	0.251	0.304	0.148	0.833

Table 3: Overall performance of various models on language of study extraction including exact match accuracy, Jaccard similarity, and micro and macro F1, precision, and recall.

resentation in computational linguistics research and coloniality in computational linguistics.

This research codifies and creates the first benchmark for language of study extraction and establishes automated language extraction using LLMs as an effective tool for analysis in computational linguistics. The performance of LLMs suggests that they can be used accurately for language of study tracking in analyses of the computational linguistics field, while maintaining the speed of previous methods. This methodology provides a foundation for more inclusive natural language processing research through the creation of a high-quality benchmark and the investigation of LLMs and SciBERT for language of study extraction.

5 Conclusions

This study creates a benchmark for language of study extraction using papers from the ACL Anthology, and presents results of language of study extraction using four LLMs and SciBERT. We use the abstract, and the abstract and the full text as inputs to the LLMs and SciBERT to evaluate the efficacy of cost-effective solutions.

GPT-4o mini using the full text achieved an exact match score of 0.646 and a Jaccard agreement score of 0.687. Gemini 2.5 Flash using the full text achieved micro and macro F1 scores of 0.633 and 0.622, respectively. Model performance without using the full text was comparable to model performance with the full text, suggesting that the use of the full text often isn't necessary for LLM-based language of study extraction. Qualitative error analysis reveals that errors are likely caused by model misunderstanding and bias, which should be mitigated in future work.

These findings suggest that the use of LLMs is a promising method for information extraction in scientific texts, especially to improve accuracy in language of study extraction in computational linguistics papers.

Limitations

This research has many limitations that must be considered. First, the annotation of publications is time-consuming due to the length of scientific papers. For that reason, this benchmark contains 431 annotated papers through a two-annotator setup, representing a relatively small sample of the ACL Anthology, potentially limiting the generalizability of these findings to the entire field of computational linguistics, and many low-resource and less studied languages in computational linguistics. This evaluation was limited to languages with more than twenty occurrences in the benchmark to ensure results were based on significant samples and that SciBERT would have training data. Finally, this approach excluded many low-resource languages in this analysis, which are extremely important in computational linguistics research.

References

Anthropic. 2024. Claude 3.5 model card. https://www.anthropic.com/news/claude-3-5.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv*.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the ACL Anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore. Association for Computational Linguistics.

Jerry Junyang Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2023. Polyie: A dataset of information extraction from polymer material scientific literature.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan. The COLING 2016 Organizing Committee.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Anna Currey and Kenneth Heafield. 2018. Unsupervised source hierarchies for low-resource neural machine translation. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 6–12, Melbourne, Australia. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1).

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Emily Herron, Vanessa Lama, Rui Pan, Azton Wells, and Nesar Ramachandra. 2025. Astromlab 4: Benchmark-topping performance in astronomy qamp;a with a 70b-parameter domain-specialized reasoning model.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? the #BenderRule in LREC and ACL articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, Marseille, France. European Language Resources Association.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models.

Sohaila Eltanbouly, May Bashendy, and Tamer Elsayed. 2019. Simple but not naïve: Fine-grained Arabic dialect identification using only n-grams. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 214–218, Florence, Italy. Association for Computational Linguistics.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp.

Harshitha Chandra Jami, Pushp Raj Singh, Avan Kumar, Bhavik R. Bakshi, Manojkumar Ramteke, and Hariprasad Kodamana. 2024. Ccu-llama: A knowledge extraction llm for carbon capture and utilization by mining scientific literature data. *Industrial amp; Engineering Chemistry Research*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A practitioner's guide to building asr models for low-resource languages: A case study on scottish gaelic.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciuca, Charles O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Jason Jingsh Li, Josh Peek, Kartheik Iyer, Tomasz Rozanski, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodriguez Mendez, Thang Bui, Alyssa Goodman, and 5 others. 2023. AstroLLaMA: Towards specialized foundation models in astronomy. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 49–55, Bali, Indonesia. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Prakhar Sharma and Sumegh Roychowdhury. 2019. IIT-KGP at MEDIQA 2019: Recognizing question entailment using sci-BERT stacked with a gradient boosting classifier. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 471–477, Florence, Italy. Association for Computational Linguistics.

Isaac Wiafe, Akon Obu Ekpezu, Raynard Dodzi Helegah, Fiifi Baffoe Payin Winful, Elikem Doe Atsakpo, Charles Nutrokpor, and Kafui Kwashie Solaga. 2025. Building an Ewe language dataset: Towards enhancing automatic speech recognition technologies for low resource languages. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 328–338, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6).

A Annotator Guidelines

The following guidelines were constructed around the current application of the language of study, particularly to gauge trends in research in computational linguistics. These guidelines were supplied to annotators for usage during independent annotation and disagreement resolution.

A.1 Inclusion Criteria

A language qualifies as a language of study in a paper if it meets any of the following conditions:

- Models are trained or evaluated on the language.
- Datasets or corpora in the language are used.
- Linguistic analysis is performed on the language.
- The language appears as source or target in translation experiments.

In multilingual or code-switching settings, all languages involved in the analysis or experimentation should be included.

A.2 Exclusion Criteria

A language is not considered a language of study if it is only:

- Mentioned in passing, as a comparison, or in related work.
- Appearing in citations, illustrative examples, or etymological discussions.
- Listed in background or introductory material without being part of the study.
- Included in a dataset or pretraining corpus that is not used in the reported experiments.
- Analyzed only in its historical form, without connection to the modern language.

B LLM Prompt

We used the following prompt for all large language model evaluations. The placeholder `{paper_text}` was replaced with the corresponding paper text or abstract for inference.

```
You are an expert in computational linguistics.

Task:
Given the paper text below, identify the natural
language(s) that are actively studied, modeled, or
used in experiments.

Include:

  • Languages used in datasets, training,
    evaluation, or experiments

  • Languages analyzed in multilingual,
    cross-lingual, or code-switching settings

  • Translation source and target languages

Exclude:

  • Languages mentioned only in background,
    citations, or related work

  • Languages used only for motivation or
    comparison

Rules:

  • Output only lowercase ISO 639-1 codes

  • Separate multiple languages with commas

  • If no language can be identified, output
    “none”

  • Do not output explanations or brackets

Paper:
{paper_text}
```

C Language-Specific Results

We present the results of each LLM and SciBERT for each of the languages on which all models were evaluated. Language-specific performance is widely varied across different languages.

D Error Analysis

To complement the quantitative analysis presented and exemplify model errors, we now present a short qualitative error analysis of examples of errors made by LLMs when predicting the languages that a paper studies.

- **Lack of focus on experiment:** In Çöltekin and Rama (2016), the language identification of 13 different languages is studied. Some models output that no languages are studied. This is likely because the large scope of the paper, studying many languages, does not count all the languages, despite the prompt annotation guidelines stating that in multilingual settings, all languages analyzed or experimented on should be included. In (Currey and Heafield, 2018), the translation between low-resource languages and English is studied. As the paper is about low-resource languages, some models output that English is not studied, even though it is used in MT pairs.

- **Complex reasoning:** In (Sharma and Roychowdhury, 2019), some models output that English is not studied. This case is complex to properly extract. In the introduction, it is stated over multiple sentences that given English sentences, the system developed makes conclusions. In this paper, determining if the study experiments on English, while explicitly mentioning English, is complex and requires reasoning to fully distinguish from a mention of English in passing.

- **Bias towards English:** In (Eltanbouly et al., 2019) studies Arabic dialect identification. Some models output that English is studied, while the word "English" is not mentioned in the paper. In (Barriga Martínez et al., 2021), the Otomi language is studied. Similarly to Eltanbouly et al. (2019), "English" is not mentioned, and some models output that English is studied. This error is potentially caused by model hallucination, and the bias of English being a common language of study in computational linguistics.

Model	**en**	**zh**	**pt**	**ar**	**ja**	**de**	**es**	**fr**
GPT-4o mini (full text + abstract)	0.46	0.77	0.40	**0.83**	0.75	0.50	0.40	0.33
Gemini 2.5 Flash (full text + abstract)	0.59	**0.83**	0.40	**0.83**	0.75	**0.67**	**0.50**	**0.40**
Claude 3.5 Haiku (full text + abstract)	0.53	0.67	**0.50**	**0.83**	0.67	0.20	0.19	0.29
Deepseek 3.2 (full text + abstract)	0.57	**0.83**	0.40	**0.83**	0.75	**0.67**	**0.50**	**0.40**
SciBERT (abstract + full paper)	0.54	0.71	0.40	0.21	**0.80**	0.20	0.15	0.12
GPT-4o mini (abstract only)	0.32	0.77	0.40	0.60	0.75	0.50	0.40	0.33
Gemini 2.5 Flash (abstract only)	0.58	**0.83**	0.40	**0.83**	0.75	**0.67**	**0.50**	**0.40**
Claude 3.5 Haiku (abstract only)	0.53	0.67	**0.50**	**0.83**	0.55	0.22	0.21	0.29
Deepseek 3.2 (abstract only)	**0.60**	**0.83**	0.40	**0.83**	0.75	0.50	**0.50**	0.33
SciBERT (abstract only)	0.52	0.50	0.40	0.28	0.38	0.06	0.10	0.12

Table 4: F1 scores of language of study extraction models on eight languages using full text + abstract or abstract only inputs.