

Conference Programme and Book of Abstracts 2nd International Workshop on Language and Language Models WoLaLa2026

Dubrovnik, Croatia, 12 – 13 October 2026



Edited by Marko Tadić, Noémi Ligeti-Nagy

Organisers

ELTE Research Centre for Linguistics, Budapest, Hungary

Croatian Language Technologies Society

University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia

HR-CLARIN

2nd International Workshop on Language
and Language Models

WoLaLa2026

Conference Programme Book of Abstracts

Edited by
Marko Tadić, Noémi Ligeti-Nagy

12 – 13 October 2026
University of Zagreb
Centre for Advanced Academic Studies
Dubrovnik, Croatia



Croatian Language
Technologies Society
Zagreb, 2026

ISBN: 978-953-55375-6-4

WoLaLa2026 Committees

Program Committee

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia (chair)
Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, Romanian Academy, Romania
Eva Hajičová, Charles University Prague, Czech Republic
Enikő Héja, ELTE Research Centre for Linguistics, Hungary
Elena Irimia, Research Institute for Artificial Intelligence, Romanian Academy, Romania
Ágnes Kalivoda, ELTE Research Centre for Linguistics, Hungary
Svetla Koeva, Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria
András Kornai, HUN-REN Institute for Computer Science and Control, Hungary
Gábor Kovács, Alfréd Rényi Institute of Mathematics, Hungary
Alessandro Lenci, University of Pisa, Italy
Noémi Ligeti-Nagy, ELTE Research Centre for Linguistics, Hungary
Veronika Lipp, ELTE Research Centre for Linguistics, Hungary
Nikola Ljubešić, Jožef Stefan Institute, Slovenia
Petya Osenova, Sofia University "St. Kl. Ohridski", Bulgaria
Csaba Pléh, Central European University, Austria
Paul Rayson, Lancaster University, United Kingdom
Réka Sólyom, Károli Gáspár University of the Reformed Church in Hungary, Hungary
Frieda Steurs, Dutch Language Institute, Belgium
Dan Tufiş, Romanian Academy, Romania
Noémi Vadász, National Archives of Hungary, Hungary
Vincent Vandeghinste, Dutch Language Institute // Centre for Computational Linguistics, University of Leuven, Belgium
Tamás Váradi, ELTE Research Centre for Linguistics, Hungary
Martin Wynne, University of Oxford, United Kingdom
Zijian Győző Yang, ELTE Research Centre for Linguistics, Hungary

Advisory Board

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia (chair)
António Branco, University of Lisbon, Portugal
Eva Hajičová, Charles University Prague, Czech Republic
Erhard Hinrichs, University of Tübingen, Germany
András Kornai, HUN-REN Institute for Computer Science and Control, Hungary
Alessandro Lenci, University of Pisa, Italy
Csaba Pléh, Central European University, Austria
Gábor Prószéky, ELTE Research Centre for Linguistics // Pázmány Péter Catholic University
Paul Rayson, Lancaster University, United Kingdom
Frédérique Segond, National Institute for Research in Digital Science and Technology, France
Frieda Steurs, Dutch Language Institute, Belgium
Dan Tufiş, Romanian Academy, Romania
Hans Uszkoreit, German Research Center for Artificial Intelligence, Germany
Tamás Váradi, ELTE Research Centre for Linguistics, Hungary
Martin Wynne, University of Oxford, United Kingdom

Local Organizing Committee

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia (chair)
Daša Farkaš, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia
Vanja Štefanec, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia
Gaurish Thakkar, University of Zagreb, Faculty of Humanities and Social Sciences, Croatia

WoLaLa2026 Programme at Glance

Monday, 2026-10-12		Tuesday, 2026-10-13
08:30-09:00	Registration	
09:00-09:30	Welcome Remarks	Keynote 9:00-10:00
09:30-10:00	Keynote 9:15-10:15	Session 5 10:00-11:00
10:00-10:30	Session 1 10:15-11:15	Coffee break
10:30-11:00	Coffee break	Session 6 11:30-12:30
11:00-11:30	Session 2 11:45-12:45	Closing Remarks
11:30-12:00	Lunch break 12:45-13:45	
12:00-12:30	Session 3 13:45-15:05	
12:30-13:00	Coffee break	
13:00-13:30	Session 4 15:35-16:35	
13:30-14:00		
14:00-14:30	Guided city tour (in front of CAAS) 17:00-19:00	
14:30-15:00		
15:00-15:30		
15:30-16:00		
16:00-16:30		
16:30-17:00		
17:00-17:30		
17:30-18:00		
18:00-18:30		
18:30-19:00		
19:00-19:30		
19:30-20:00		
20:00-20:30	Conference Dinner (Restaurant Orhan) 19.30-22:30	
20:30-21:00		
21:00-21:30		
21:30-22:00		
22:00-22:30		

WoLaLa2026 Detailed Programme

Monday, 2026-10-12

8:30–9:00	Registration
9:00–9:15	Opening and Welcome Remarks
9:15–10:15	Marko Grobelnik (keynote): <i>Mechanistic Interpretability as an experimental method for Computational Linguistics</i> Chair: Marko Tadić
Session 1	LLM Knowledge, Reasoning & Human Comparison Chair: Marko Tadić
10:15–10:35	Anna Mikó, Anna Babarczy: <i>Epistemic vigilance and the effect of LLM-based assistance on stress during cognitively demanding tasks: Dissociating state anxiety from perceived workload</i>
10:35–10:55	Noémi Ligeti-Nagy, Máttyás Osváth: <i>Does Answer Format Matter? Multiple-Choice vs. Open-Ended Evaluation of Cultural Knowledge in Hungarian LLMs</i>
10:55–11:15	Marina Sokolova, Natalia Moskvina, Nayara Mirio E. Silva: <i>Code-Switching as a Parsing Cue: Humans vs. LLMs</i>
11:15–11:45	Coffee Break
Session 2	Lexical Semantics, Terminology & Lexicography Chair: Enikő Héja
11:45–12:05	Claus Zinn, Erhard Hinrichs: <i>Evaluating the Usage of LLMs for an Example-Based Word Sense Disambiguation Task</i>
12:05–12:25	Barbara Heinisch: <i>Specialised Language and Large Language Models: Terminology as Pain Point in LLM-Produced Translations of University Texts in German?</i>
12:25–12:45	Alexey Matyushin, Daria Ryzhova, Anastasiya Klenova, Adelina Makarova, Lada Starostina: <i>LLM-Assisted Serbian–Russian Lexicography: The Case of Medical Terminology</i>
12:45–13:45	Lunch Break
Session 3	Metadata, Classification & Annotation Chair: Claus Zinn
13:45–14:05	Réka Dodé: <i>The Applicability of Keywords and Additional Metadata for Fields of Science Classification</i>
14:05–14:25	Gaurish Thakkar, Marko Tadić, Nives Mikelić Preradović: <i>Co-text-Aware Croatian Sentiment Annotation Using Open-Source Large Language Models</i>
14:25–14:45	Réka Dodé, Zijian Győző Yang, Flóra Földesi, Enikő Héja: <i>Prompt-Based Metadata Extraction with Llama, Qwen and Gemma on Hungarian Scientific Corpora</i>
14:45–15:05	Alexandr Parahonco, Victoria Bobicev, Maria Carp, Verginica Barbu-Mititelu: <i>From Metadata to Synthetic Legal Text: A Two-Stratum Romanian Generation Pilot for ADAMO</i>

15:05–15:35	Coffee Break
Session 4	Applications: RAG, Language Learning & Community/Society Chair: Gábor Prószéky
15:35–15:55	Boglárka Vermeki: <i>Corpus-Grounded Retrieval-Augmented Generation for Beginner Hungarian: Comparing Pedagogical and Authentic Child-Language Grounding in Communicative Task Generation</i>
15:55–16:15	Rodrigo Santos, Luís Gomes, Afonso Vaz, António Branco: <i>RAG vs. Fine-Tuning for Chatbot Enhancement: Empirical Study with a Real-World Public Services Use Case</i>
16:15–16:35	Sahil Kale: <i>Analysing Temporal Self-Knowledge Signals in Large Language Models</i>
17:00–19:00	Guided city tour (starting in front of CAAS)
19:30–22:30	Conference Dinner (Restaurant Orhan)

Tuesday, 2026-10-13

9:00–10:00	Will Lamb (keynote): <i>Benchmarks aren't enough: Modelling the sociolinguistic impact of LLMs</i> Chair: Martin Wynne
Session 5	Language Resources, Tools & Community Infrastructure Chair: Erhard Hinrichs
10:00–10:20	Flóra Földesi, Zsófia Ludányi, Noémi Ligeti-Nagy: <i>Large Language Models as Intelligent Gateways for Rule-Based Spelling Consultancy: A Case Study on Hungarian Orthography</i> (helyesiras.mta.hu)
10:20–10:40	Bence Sárosy: <i>Augmenting Fake News Corpus with Information from LLMs</i>
10:40–11:00	Tamás Váradi: <i>WoLaLa Wiki: A Research Wiki for Language and Language Model</i> (demo)
11:00–11:30	Coffee Break
Session 6	LLM Applications for Texts, Libraries & Digital Humanities Chair: Tamás Váradi
11:30–11:50	Nicolae Rusu, Victoria Bobicev, Daniela Istrati: <i>Evaluation of Semantic and Visual Coherence in AI-generated Journalistic Content</i>
11:50–12:10	Adam Majczyk, Maciej Ogrodniczuk: <i>Retrieval-Augmented Generation System for the Polish PressMint Corpus</i>
12:10–12:30	Berenike Herrmann, Andreas Witt: <i>Textification. On the role of text as an interface in the age of LLMs</i>
12:30–12:45	Closing remarks
14:00–22:00	Free time / self-funded excursion to Trsteno and Ston with dinner in Mali Ston

Extended Abstracts

(ordered by the surname of the first author)

Réka Dodé

ELTE Research Centre for Linguistics, Budapest, Hungary, dode.reka@nytud.elte.hu

The applicability of keywords and additional metadata for fields of science classification

Keywords: Hungarian Scientific Text, Field of Science, Prompt Engineering, Metadata-based Classification, ChatGPT 5.2, PULI-LlumiX-Llama-3.1

I Introduction

One of the fundamental difficulties of fields of science classification lies in the construction of appropriate training and evaluation corpora (Dodé et al., 2026). In several smaller scientific subfields, both the number of domain experts and the quantity of available Hungarian-language publications are limited, which significantly constrains the development and training of reliable classification models. As a consequence, the merging of related scientific categories or the reduction of the classification hierarchy may become necessary in order to ensure sufficient quantity and quality of training data. This challenge is especially relevant in multilingual and low-resource environments, where disciplinary corpora are unevenly distributed across scientific fields and metadata quality may vary considerably.

Recent advances in generative artificial intelligence have demonstrated that large language models are capable of performing complex classification tasks using zero-shot prompting without task-specific fine-tuning (Kojima et al., 2023). The present study investigates the extent to which keywords, titles, and librarian-assigned subject headings contribute to fields of science classification performed in Hungarian by generative language models using a zero-shot approach, that is, without the use of task-specific training data or fine-tuning.

The relationship between terminology, metadata, and fields of science classification has become increasingly important in the context of digital scholarly communication and artificial intelligence-based information retrieval (i.e. Pottier et al., 2024), (Hersh, 2024). Keywords, subject headings, and classification labels are therefore not independent metadata elements, but interconnected representations of disciplinary knowledge operating at different levels of abstraction (Schwing et al., 2012, Pottier et al., 2024). Their primary function is to support indexing, retrieval, and the organization of scientific information (Corrin et al., 2022, Schilhan et al., 2021).

Keywords occupy a particularly important position within this system. Unlike broader classification labels, keywords are selected directly by authors and are intended to represent the most relevant concepts of a publication in a concise form. From a terminological perspective, keywords may be interpreted as privileged terminological units that condense the thematic and conceptual focus of a scientific text (Hulth, 2003, Dodé, 2024). Because they emerge from the terminology of a given scientific domain, they also reflect the disciplinary embedding of the publication itself (Kwon, 2018). In this sense, keywords not only facilitate retrieval but also provide partial access to the conceptual structure of the underlying scientific subdomain.

However, the effectiveness of keywords in fields of science classification largely depends on their specificity and disciplinary distinctiveness. Certain keywords function as highly discriminative terminological markers, while others are too general, polysemous, or interdisciplinary to support reliable categorization (Dodé, 2024, Kwon, 2018, Uddin and Khan, 2016). This issue becomes particularly relevant in automated classification systems relying on machine learning or large language models (LLMs), where the interpretation of metadata strongly influences classification outcomes (Reynolds and McDonnell, 2021, Sebastiani, 2002).

The research is based on the assumption that keywords, subject headings, and the titles represent different layers of the same disciplinary knowledge structure. Ideally, these metadata elements complement one another in information retrieval systems. The study therefore examines how the gradual addition of metadata elements — including keywords, titles, and

librarian-assigned subject headings — influences the accuracy of fields of science classification performed by large language models, and whether the inclusion of additional metadata improves classification quality and reduces ambiguity in model outputs.

The study further assumes that the automation of fields of science classification remains a highly complex and multidimensional task that depends not only on the semantic quality and specificity of metadata, but also on the architecture, resource capacity, and interpretive behaviour of the applied language models. Consequently, scientific classification should not be interpreted merely as a technical process of label assignment, but rather as an interaction between disciplinary terminology, metadata organization, and artificial intelligence–based semantic interpretation.

2 Methods and Data

The empirical analysis was conducted on a corpus of Hungarian-language scientific publications obtained from the Contenta repository (SZTE Klebelsberg Könyvtár és Levéltár, n.d.). The dataset consisted of 34 publications selected according to the availability of complete metadata records: each publication contained author-provided keywords, librarian-assigned subject headings, and Frascati-based fields of science labels classified to the second hierarchical level (OECD, 2007). Publications containing only first-level scientific labels were excluded from the analysis in order to ensure sufficient disciplinary specificity. The Frascati second-level categories provide relatively fine-grained disciplinary distinctions, making the classification task substantially more complex than broader top-level categorization.

The study employed a zero-shot prompting methodology (Syed and Gadesha, 2025) using generative language models. Two models were compared: ChatGPT 5.2 (OpenAI, n.d.) and PULI-LlumiX-Llama-3.1 Chat (aka ChatPULI) (Yang et al., 2025). Both systems received prompts in isolated sessions, meaning that every experiment was conducted in a newly initiated conversation without previous contextual interaction. This methodological decision ensured that earlier prompts or responses could not influence subsequent classifications through conversational memory effects.

The experiments focused exclusively on metadata-based classification rather than full-text analysis. The models were not provided with the body text or abstracts of the publications; instead, they received different combinations of metadata elements. This approach allowed the investigation of how individual metadata types – keywords, title, subject heading – influence classification quality and disciplinary disambiguation.

The evaluation process relied on librarian-assigned scientific labels as the gold standard. In cases where multiple valid scientific labels were assigned to a publication, all labels were retained, and a model prediction was considered correct if it matched any of the accepted categories. This approach reflects the inherently multidimensional nature of scientific classification, where publications may legitimately belong to more than one disciplinary area.

3 Experiments

Three prompt-based experiments were conducted in order to examine how different metadata configurations influence fields of science classification performed by large language models:

- In the first experiment, the models (ChatGPT 5.2 and ChatPULI) received only author-provided keywords together with the complete list of 45 Frascati scientific categories. The prompt instructed the models to select the category most closely related to the provided keywords. This experimental condition tested the independent classificatory value of keywords without additional contextual information.
- The second experiment introduced publication titles in addition to the keywords. The expectation was that titles would provide additional contextual cues and improve disciplinary precision.
- The third experiment incorporated librarian-assigned subject headings alongside the keywords and titles.

4 Results

The experiments revealed that the quality of fields of science classification performed by large language models is dependent on the specificity, consistency, and disciplinary relevance of metadata elements. Keywords alone often proved insufficient for reliable classification, particularly when they consisted of abstract or interdisciplinary terminology. Polysemous and highly general terms may reduce classification accuracy because they lack strong disciplinary discriminative power.

The inclusion of titles produced mixed effects. Although titles occasionally improved contextual interpretation, they sometimes introduced competing terminological signals that destabilized classification decisions. This finding suggests that titles do not necessarily function as neutral contextual enhancements; instead, they may amplify certain concepts in ways that alter the interpretive focus of the models.

The most significant improvement emerged after the inclusion of librarian-assigned subject headings. Because subject headings are developed precisely for thematic indexing and semantic organization, they naturally function as strong disciplinary indicators. Their controlled and standardized nature reduced ambiguity and supported more accurate scientific categorization. The findings therefore highlight the continuing relevance of library science methodologies in the era of artificial intelligence–based information retrieval.

Clear differences also emerged between the two examined models. ChatPULI displayed greater sensitivity to metadata variation and produced less stable classifications overall. Its decisions frequently shifted across prompt versions, and several publications received entirely different scientific labels depending on the metadata combination used. ChatGPT demonstrated greater consistency and thematic stability, even when metadata were incomplete or weakly informative.

The experiments also highlighted certain limitations of manual fields of science classification systems. In some cases, the classifications generated by the language models differed from the librarian-assigned scientific categories, suggesting that disciplinary categorization may itself contain interpretive ambiguities or inconsistencies. These findings indicate that AI-assisted classification systems could potentially contribute not only to automated indexing, but also to metadata validation and quality control within library and information science workflows.

Overall, the study confirms that metadata elements represent interconnected layers of disciplinary knowledge organization. Keywords, titles, subject headings, and fields of science classification labels should therefore not be interpreted as isolated descriptors but as complementary components of scientific information architecture. The results further demonstrate that controlled metadata systems remain highly valuable in AI-assisted retrieval and classification environments, particularly when combined with the contextual interpretation capabilities of large language models.

References

- Corrin, L., Thompson, K., Hwang, G.-J., & Lodge, J. M. (2022). The importance of choosing the right keywords for educational technology publications. *Australasian Journal of Educational Technology*, 38(2), 1–8. <https://doi.org/10.14742/ajet.8087>
- Dodé, R. (2024). Kulcsszavak és terminusok vizsgálata a real repozitóriumnak anyagán – pilotkutatás. In K. Fogarasi, D. Ittész, & D. Máty (Eds.), *Tudásmegosztás, információkezelés, alkalmazhatóság: I. nyelvhasználat*. Akadémiai Kiadó. <https://doi.org/10.1556/9789636640910>
- Dodé, R., Varga, K., Yang, Z. G., Madarász, G., Osváth, M., & Héja, E. (2026). Opportunities and challenges in classifying hungarian scientific texts by field of science [Special Issue, manuscript under review]. *Acta Linguistica Academica*.
- Hersh, W. (2024). Search still matters: Information retrieval in the era of generative ai. *Journal of the American Medical Informatics Association*, 31(9), 2159–2161. <https://doi.org/10.1093/jamia/ocae014>
- Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 216–223. <https://doi.org/10.3115/1119355.1119383>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. <https://arxiv.org/abs/2205.11916>

- Kwon, S. (2018). Characteristics of interdisciplinary research in author keywords appearing in Korean journals. *Malaysian Journal of Library & Information Science*, 23(2), 77–93. <https://doi.org/10.22452/mjlis.vol23no2.5>
- OECD. (2007, February). Revised field of science and technology classification in the Frascati manual. <https://www.aris-rs.si/en/gradivo/sifranti/inc/Frascati.pdf>
- OpenAI. (n.d.). ChatGPT. <https://openai.com/hu-HU/index/chatgpt/>
- Pottier, P., Lagisz, M., Burke, S., Drobnik, S. M., Downing, P. A., Macartney, E. L., Martinig, A. R., Mizuno, A., Morrison, K., Pollo, P., Ricolfi, L., Tam, J., Williams, C., Yang, Y., & Nakagawa, S. (2024). Title, abstract and keywords: A practical guide to maximize the visibility and impact of academic papers. *Proceedings of the Royal Society B*, 291(2027), 20241222. <https://doi.org/10.1098/rspb.2024.1222>
- Reynolds, L., & McDonnell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv*. <https://doi.org/10.48550/arXiv.2102.07350>
- Schilhan, L., Kaier, C., & Lackner, K. (2021). Increasing visibility and discoverability of scholarly publications with academic search engine optimization. *Insights: The UKSG Journal*, 34. <https://doi.org/10.1629/uksg.534>
- Schwing, T., McCutcheon, S., & Maurer, M. (2012). Uniqueness matters: Author-supplied keywords and LCSH in the library catalog. *Cataloging & Classification Quarterly*, 50(8), 903–928. <https://doi.org/10.1080/01639374.2012.703164>
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47. <https://doi.org/10.1145/505282.505283>
- Syed, M., & Gadesha, V. (2025, January). What is zero-shot prompting? <https://www.ibm.com/think/topics/zero-shot-prompting>
- SZTE Klebelsberg Könyvtár és Levéltár. (n.d.). Contenta repozitóriumbok. <https://www.ek.szte.hu/kezdoadal/mit-keres/contenta-repozitoriumok/>
- Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, 10(4), 1166–1177. <https://doi.org/10.1016/j.joi.2016.10.004>
- Yang, Z. G., Bánfi, Á., Dodé, R., Ferenczi, G., Földesi, F., Hatvani, P., Héja, E., Lengyel, M., Madarász, G., Osváth, M., Sárossy, B., Varga, K., Váradi, T., Prószéky, G., & Ligeti-Nagy, N. (2025). Chatpuli: Enhancement to the first Hungarian conversational model. *Annales Mathematicae et Informaticae*, 61, 261–274. <https://doi.org/10.33039/ami.2025.10.010>

Réka Dodé

Zijian Győző Yang

Flóra Földesi

Enikő Héja

ELTE Research Centre for Linguistics, Budapest, Hungary, {dode.reka, gyozyang.zijian, foldesi.flora, eniko.heja}@nytud.elte.hu

Prompt-Based Metadata Extraction with Llama, Qwen and Gemma on Hungarian Scientific Corpora

Keywords: Hungarian Scientific Text, Metadata Extraction, Prompt Engineering, Llama model, Qwen model, Gemma model

I Introduction

Structured and interoperable metadata have always been essential for the description, identification, retrieval, and preservation of documents. Contemporary publishing workflows, born-digital document creation practices, repository infrastructures, and metadata standards, however, increasingly support the creation, management, and exchange of such metadata in machine-readable and interoperable formats. Standards and frameworks such as Dublin Core, MARC, MODS, DataCite, and repository interoperability protocols enable scientific documents to become more accessible, searchable, and reusable across digital library systems and academic databases.

Despite these developments, a significant portion of scientific literature still exists in forms where metadata are incomplete or inconsistent. This is especially true in the case of retrospectively digitized documents, such as legacy collections, scanned journals, conference proceedings, and older institutional publications. Although digitization technologies and OCR systems have significantly improved access to textual content, OCR processing still faces considerable challenges, particularly in the case of older and lower-quality documents. Scanned texts, typewritten documents, damaged pages, and poor print quality often result in character recognition errors, structural distortions, and incomplete text extraction. These issues not only hinder full-text search capabilities but also make the automatic generation of reliable metadata considerably more difficult, thereby affecting efficient retrieval, indexing, long-term preservation, and integration into repository systems.

The metadata extraction approaches have largely relied on rule-based systems, regular expressions, layout analysis, or supervised machine learning models trained on manually annotated corpora so far (Councill et al., 2008, Tkaczyk et al., 2015, Lopez, 2009). Although these methods can achieve high precision in controlled environments, they often require extensive domain-specific engineering and may generalize poorly across document types, layouts, languages, and OCR quality variations. Scientific texts present additional challenges because of their heterogeneous structures, multilingual terminology, references, tables, and inconsistent formatting conventions.

Recent advances in large language models (LLMs) open new possibilities for metadata extraction and document understanding tasks (i.e. Lamba et al., 2025). Instruction-tuned and conversational LLMs can process semantically complex texts and perform information extraction tasks through natural language prompting. Unlike traditional extraction pipelines, prompt-based approaches allow the extraction task to be formulated in flexible natural language instructions without the need for extensive task-specific retraining.

This paradigm may be especially beneficial for low-resource and morphologically rich languages such as Hungarian.

In this study, we investigate the applicability of three open-source large language models, Llama 3.3, Qwen 2.5, and Gemma 3, for prompt-based metadata extraction from Hungarian scientific corpora. The research focuses on the extraction of key bibliographic and descriptive metadata elements, including titles, authors, abstracts, keywords, and publication-related information.

We compare the behaviour and performance of the three models under different prompting strategies designed to examine how task complexity and the choice of in-context examples influence metadata extraction.

Beyond comparing model performance, the study also explores whether prompt-based LLM approaches can contribute to metadata enrichment workflows in repositories and digital libraries, particularly in contexts where legacy documents and OCR-derived texts remain prevalent. By examining the capabilities, robustness, and limitations of Llama, Qwen, and Gemma under varying prompt configurations, we aim to contribute to the broader discussion on the integration of generative AI technologies into scholarly information management and metadata creation workflows.

2 Methods

2.1 The test dataset

The construction of the test dataset consisted of several stages. First, 200 documents were selected from the REAL repository¹ with the assistance of librarian experts. In the next step, preliminary annotations were generated automatically using GPT-4 (OpenAI, 2023). These pre-annotations facilitated the manual annotation process by providing initial metadata candidates and structural suggestions.

Following the automated pre-annotation stage, specialists from the Library and Information Centre of the Hungarian Academy of Sciences (MTA KIK) manually reviewed and corrected the annotations. For evaluation and comparison, the resulting annotated data were converted into well-formed JSON format.

2.2 Experiments

The experiments were conducted using three open-source large language models: Llama 3.3 70B (Meta AI, 2024), Qwen2.5 72B (Yang et al., 2024), and Gemma 3 27B (Gemma Team, 2025). All models were evaluated in prompt-based metadata extraction scenarios in order to compare their ability to process Hungarian scientific texts and generate structured metadata outputs.

The models were instructed to return the extracted metadata in structured JSON format to support automatic processing and evaluation. If no information corresponding to a given metadata label could be identified in the document, the models were instructed to return N/A for that field. The generated outputs were subsequently converted into well-formed JSON format for comparison with the annotated data.

The initial aim was to extract a broad range of bibliographic and descriptive metadata elements from the documents. The original metadata schema contained 27 fields: Hungarian title, foreign-language title, subtitle, author surname, author given name, affiliation, abstract, keyword, journal title, volume, issue number, article identifier, ISSN, title of the host publication, editor surname, editor given name, place of publication, publisher, series title, series number, ISBN, first page number, last page number, year of publication, DOI, URL, and license information.

However, preliminary observations indicated that requesting such a large number of metadata elements simultaneously was not an optimal strategy. We therefore designed the subsequent experiments to investigate how reducing the number of requested metadata fields affected the quality and consistency of the generated outputs.

The experiments started with a configuration containing 14 metadata fields: HUNGARIAN TITLE, FOREIGN LANGUAGE TITLE, SUBTITLE, AUTHOR'S LAST NAME, AUTHOR'S FIRST NAME, ABSTRACT, KEYWORDS, JOURNAL TITLE, VOLUME, ISSUE, YEAR OF PUBLICATION, ISSN, DOI, and LICENSE. The number of requested fields was then progressively reduced, with the smallest configuration containing nine fields: HUNGARIAN TITLE, AUTHOR'S LAST NAME, AUTHOR'S FIRST NAME,

¹ <https://real.mtak.hu/>

KEYWORDS, JOURNAL TITLE, YEAR OF PUBLICATION, ISSN, DOI, and LICENSE. In the smallest configuration, FOREIGN LANGUAGE TITLE, SUBTITLE, ABSTRACT, VOLUME, and ISSUE were therefore omitted compared with the 14-field configuration. This experimental setup allowed us to examine whether reducing the scope of the extraction task improved the accuracy, stability, and structural consistency of the model outputs.

In addition to varying the number of requested metadata fields, we examined the influence of the few-shot examples included in the prompts. The same sequence of progressively reduced metadata configurations was tested with two different sets of example documents. In the first condition, the prompts contained one pair of example documents, while in the second condition, these were replaced with a different pair of examples, with the corresponding metadata configurations otherwise kept unchanged. This design allowed us to examine whether model performance was influenced not only by the number of requested metadata fields but also by the specific examples provided in the prompt.

Overall, 10 prompt configurations were evaluated for each model, combining progressively reduced metadata sets with two different pairs of few-shot examples. This experimental design enabled us to investigate both the effect of reducing the scope of the extraction task and the sensitivity of the models to the choice of in-context examples.

3 Results

For the evaluation process, a comparison script was developed to automatically compare the generated metadata outputs with the manually annotated reference labels.

The script calculates three complementary types of agreement between the generated and reference metadata values assigned to individual labels.

First, exact matches are identified, requiring the generated value to correspond exactly to the reference value. For example, for a title label, the generated title must be identical to the reference title.

Second, order-independent matches are calculated for metadata fields containing multiple values, such as author names, where the ordering of the authors is disregarded and agreement is assessed based on the presence of the same elements.

Third, for longer text-based outputs, such as abstracts, Levenshtein distance is used to quantify the degree of difference between the generated and reference texts.

In addition to these measures, the evaluation process categorizes different types of discrepancies, making it possible to identify recurring extraction problems and structurally inconsistent outputs.

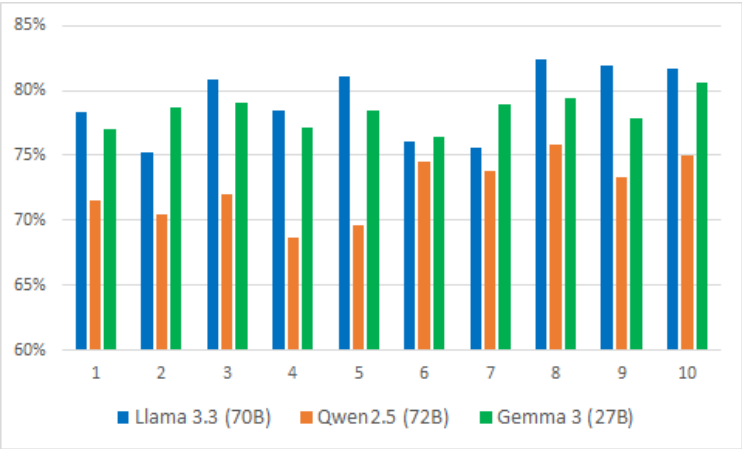


Figure 1: Comparison of the three LLMs across different prompts

Figure 1 shows that the Llama model achieved the best overall performance, whereas Qwen performed the least well. Interestingly, the Gemma model, which is considerably smaller (27B parameters compared with 70B for Llama), was still able to achieve competitive results. Furthermore, the prompt engineering experiments indicate that reducing the number of requested metadata labels generally improved model performance. Prompt sets containing fewer metadata labels (Prompts 3, 4, 5, 8, 9, and 10) tended to produce better results than those requiring the extraction of a larger number of metadata fields.

The results also reveal a modest tendency related to the choice of few-shot examples. All three models achieved slightly better results when provided with the second pair of example documents. Although the observed differences were relatively small, this pattern suggests that extraction quality is not entirely independent of the examples included in the prompt. The choice of in-context examples may therefore influence model performance even when the extraction task and the requested metadata fields remain otherwise unchanged.

The preliminary experiments suggest that prompt-based large language models can support metadata extraction workflows for Hungarian scientific documents; however, extraction quality remains highly dependent on prompt design, document structure, OCR quality, and the complexity of the targeted metadata fields.

References

- Councill, I., Giles, C. L., & Kan, M.-Y. (2008, May). ParsCit: An open-source CRF reference string parsing package. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*. European Language Resources Association (ELRA). <https://aclanthology.org/L08-1291/>
- Lamba, M., Peng, Y., Nikolov, S., Layne-Worthey, G., & Downie, J. (2025). Metadata enrichment of long text documents using large language models. *Proceedings of the Association for Information Science and Technology*, 62, 990–994. <https://doi.org/10.1002/pra2.1327>
- Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. 5714, 473–474. https://doi.org/10.1007/978-3-642-04346-8_62
- OpenAI. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774. <https://arxiv.org/abs/2303.08774>
- Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). Cermine: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4), 317–335. <https://doi.org/10.1007/s10032-015-0249-8>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv. <https://doi.org/10.48550/arXiv.2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. <https://arxiv.org/abs/2307.09288>
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., ... Qiu, Z. (2025). Qwen3 technical report. <https://arxiv.org/abs/2505.09388>

Flóra Földesi
Zsófia Ludányi
Noémi Ligeti-Nagy

ELTE Research Centre for Linguistics, Budapest, Hungary, {foldesi.flora, ludanyi.zsofia, ligeti-nagy.noemi}@nytud.elte.hu

Large Language Models as Intelligent Gateways for Rule-Based Spelling Consultancy: A Case Study on Hungarian Orthography helyesiras.mta.hu spelling consultancy portal

Keywords: Hungarian orthography, spelling consultancy, intent classification, query routing, target extraction, hybrid language technologies

I Introduction

The Hungarian spelling portal helyesiras.mta.hu is a long-established language-technology service developed by the Hungarian Research Centre for Linguistics. The portal provides consultancy for a broad range of orthographic problems, including compound-word spelling, numeral expressions, date normalization, hyphenation, name lookup and orthographic correction. Unlike general-purpose spell checkers, the system is organized around a collection of specialized tools, each implementing a particular subset of the rules of Hungarian orthography. (Ludányi and Domonkosi, 2023)

The architecture of the portal reflects the nature of Hungarian spelling itself. Although many orthographic phenomena exhibit regular patterns, correct spelling often depends on detailed linguistic rules, lexical conventions and numerous exceptions. Consequently, the service evolved as a collection of rule-based and lexicon-based components that provide deterministic and explainable answers. Such systems offer a high degree of reliability within their respective domains, but they require users to identify the type of problem they wish to solve and select the appropriate tool. (Ludányi, 2023)

Over the years, usage logs and user feedback have revealed recurring difficulties (Ludányi, 2026). Many users formulate their questions in natural language and are uncertain which spelling category their problem belongs to. In practice, incorrect tool selection is a frequent source of unsuccessful interactions. At the same time, recent advances in conversational artificial intelligence have substantially changed user expectations. Rather than navigating a collection of specialized tools, users increasingly expect to communicate with language technologies through free-form questions and chatbot-like interfaces.

This development raises a broader methodological question. Should traditional rule-based language technologies be replaced entirely by large language models, or should language models be used to enhance existing systems while preserving their underlying linguistic knowledge?

A fully LLM-based spelling assistant would appear attractive because modern language models can interpret natural-language questions, identify relevant information and generate fluent explanations. However, several considerations argue against relying exclusively on generative models in this domain. First, spelling consultancy requires a level of orthographic expertise that may not be consistently represented in the training data of general-purpose language models, especially for languages with more limited digital resources such as Hungarian. Second, many spelling tasks require exact reproduction of user-provided expressions, whereas language models may occasionally alter character sequences during generation. Third, spelling advice is expected to be deterministic and reproducible, while the outputs of probabilistic language models may vary across contexts and model versions. Finally, rule-based spelling systems provide transparent reasoning grounded in explicit linguistic principles, whereas the internal decision processes of large language models remain largely opaque.

For these reasons, we investigate a hybrid architecture that combines the strengths of both approaches. Rather than replacing the existing spelling tools, we use a large language model as an intelligent gateway to the underlying rule-based infrastructure. In this setting, the language model

is responsible for interpreting the user’s question, selecting the appropriate spelling component and identifying the expression that should be forwarded to that component. The linguistic analysis and final spelling decision remain the responsibility of the specialized backend tools.

From a machine-learning perspective, this architecture transforms the problem into a joint routing and information-extraction task. The model must simultaneously identify the intended spelling category and extract the relevant target expression from the user’s question. This differs from traditional intent classification problems because successful processing requires both correct classification and precise identification of textual spans. The task therefore combines aspects of intent detection, slot filling, structured generation and information extraction.

Recent instruction-tuned large language models have demonstrated strong performance on structured prediction tasks, suggesting that a single model may be capable of performing both operations simultaneously. In this study, we compare two alternative modelling paradigms: the multilingual sequence-to-sequence model mT5 and the instruction-tuned decoder-only model Qwen3-8B. Using a manually curated dataset of Hungarian spelling questions, we evaluate their ability to support a hybrid spelling consultancy architecture and investigate the practical trade-offs between rule-based and neural approaches in a real-world language-technology application.

2 Data and Methods

The dataset was assembled from real user questions submitted to the spelling portal. The original collection contained 14,636 manually annotated questions distributed across nine categories: *abc*, *datum*, *egyeb*, *elvalasztas*, *helyese*, *kulegy*, *nevereso*, numeral and off-topic. The data were divided into training, validation and test partitions, and the training set was balanced through up- or down-sampling.

The task was formulated as structured generation. Given a natural-language question, the model generated a JSON object containing (1) the appropriate spelling tool and (2) the target expression extracted from the question.

We first trained mT5 (Raffel et al., 2020), a multilingual sequence-to-sequence model frequently used for text-to-text tasks. We subsequently fine-tuned Qwen3-8B (Team, 2024), an instruction-tuned decoder-only language model, using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Both models were trained to generate JSON output directly from user questions.

Performance was evaluated using tool-selection accuracy, target-extraction F1, JSON exact match and an overall score combining routing and extraction quality.

3 Results

The strongest mT5 configuration achieved 95.2% tool-selection accuracy and 80.5% target-extraction F1. While routing performance was already strong, target extraction remained challenging, particularly in categories involving compound-word boundaries and morphologically complex expressions. We also observed occasional modifications of character sequences in extracted targets, which is problematic in spelling-consultancy applications where exact reproduction of user input is often required.

The Qwen3-8B model substantially improved performance across all evaluation metrics. On the same test set it achieved 98.0% tool-selection accuracy, 94.8% target-extraction F1 and 94.4.

Table 1: Comparison of the mT5 and Qwen3-8B models.

Metric	mT5	Qwen3-8B
Tool Accuracy	95.2%	98.02%
Target F1	80.46%	94.84%
JSON Exact Match	76.83%	94.44%
Overall Score	0.87	0.96

In addition to strict JSON evaluation, we manually assessed extraction quality from a functional perspective. Under this criterion, the model extracted the correct keyword in 1240 of 1260 test

questions (98.4%), indicating that many formally incorrect outputs remained suitable for downstream processing.

4 Discussion and Conclusion

The results demonstrate that instruction-tuned decoder-only models are highly effective for Hungarian spelling routing and target extraction. The largest improvements were observed in extraction quality, suggesting that such models are particularly capable of preserving exact textual spans while simultaneously performing classification.

We also observed that Qwen3-8B appeared more robust to annotation inconsistencies and minor dataset modifications than mT5, whose performance fluctuated more strongly across training iterations.

Finally, we evaluated an 8-bit quantized version of Qwen3-8B and observed only negligible performance degradation, making practical deployment feasible despite infrastructure constraints.

The study shows that a single instruction-tuned language model can successfully replace a traditional multi-stage routing pipeline and provide near-production-level performance for Hungarian spelling consultancy.

References

- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR.
- Ludányi, Z. (2023). A helyesírási tanácsadó portál használata az anyanyelvi nevelésben: Esettanulmányok a portál eszközeinek megfelelő használatára. In B. Istók, G. Lőrincz, T. Török, & P. Baka L. (Eds.), *A helyesírás-tanítás aktuális kérdései. az sje tdk magyar nyelv és irodalom tanszéke mellett működő variológiai kutatócsoport xii. nemzetközi tudományos szimpóziumának tanulmánykötete* (pp. 55–65). Selye János Egyetem. https://pf.ujs.sk/documents/books/IB_LG_TT_BLP_A%20helyesiras%20tanitas%20aktualis%20kerdesei%202023.pdf
- Ludányi, Z. (2026). *A nyelvi tanácsadás nemzetközi és hazai gyakorlatai. különös tekintettel a nyelvtudományi kutatóközpontra* [[Language Consulting in Hungary and Beyond, with Special Regard to the Hungarian Research Centre for Linguistics]]. Szépirodalmi Figyelő Alapítvány.
- Ludányi, Z., & Domonkosi, Á. (2023). Language consulting and language management from the perspective of the Hungarian language consulting service. *Taikomoi kalbotyra*, 20, 74–88. <https://doi.org/10.15388/Taikalbot.2023.20.6>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.
- Team, Q. (2024). *Qwen technical report* (tech. rep.). arXiv preprint.

Barbara Heinisch

Eurac Research, Bolzano, Italy, barbara.heinisch@eurac.edu

Specialised language and large language models: Terminology as pain point in LLM-produced translations of university texts in German?

Keywords: language for specific purposes, corporate language, terminology annotation

I Introduction

Large language models (LLM), such as ChatGPT, Gemini, Mistral and Claude generally perform well with general language, e.g. product descriptions, newspaper articles and blog posts. However, apart from the fact that LLMs only cover about 2-3 per cent of the world's languages (Bella et al., 2024), their capabilities are also limited when it comes to specialised language (Van der Yeught, 2016), such as that used in medicine, law, education and health; even in high-resource languages, such as German.

This is a paper related to the application of LLMs in the creation and translation of specialised texts, with a focus on the terminology contained in these texts and is part of the UniTermGPT (*University terminology in German in the age of ChatGPT*) project. The research aim of UniTermGPT is to assess how ChatGPT copes with university terminology in translation tasks between German and English (just by means of prompt design, without retrieval-augmented generation (RAG) (Lewis et al., 2020) or terminology-augmented generation (TAG) (Lackner et al., 2025)).

The research question of this sub-study is: Which text types in specialised domains (in this case the university domain) are particularly challenging in translations generated by ChatGPT? The hypothesis is that texts containing a higher amount of university terminology will result in a higher number of terminology errors in the LLM output.

In the language industry, specialised language (or language for specific purposes) is defined as “natural language used in communication between experts in a domain and characterized by the use of specific linguistic means of expression” (ISO, 2019). Terminology, according to the ISO standard on *Terminology work and terminology science* is considered a “set of designations and concepts belonging to one domain or subject” (ISO, 2019). Examples of terms in the domain of higher education are “transcript of records”, “bachelor’s programme” or “European Credit Transfer and Accumulation System (ECTS)”. It is important to note that the literature on terminology theory reflects different schools of thought, which define terminology in slightly varying ways. For this reason, this paper adopts the industry-oriented definition of terminology.

System-bound terminology refers to concepts that are tied to a specific institutional, legal or historical system and therefore cannot be transferred directly to another system (Sandrini, 2018). In law, concepts are shaped by national legal traditions and socio-political contexts, making direct equivalence across systems impossible (Chiocchetti, 2021). As a result, terminology must be analysed in relation to the system it belongs to avoid false equivalence in translation and communication. Similarly, culture-bound terminology reflects culturally specific (and changing) ways of organising knowledge and conceptualising reality (Temmerman & van Campenhoudt, 2014). While law is the most prominent example of system-bound terminology, other domains, including higher education and university systems, are also characterised by strong system specificity (Ralli et al., 2007).

This study focuses on university texts, understood as written materials produced by higher education institutions. These texts include a wide range of text types, from highly standardised and legally binding documents, such as curricula and university statutes, to strategic documents, technical instructions and more informal (student-facing) or marketing texts. They also address different target groups, including university staff, students and external audiences. Due to their diversity in text type and degree of specialisation, university texts are particularly suitable for testing different prompts in LLM-based research. In particular, the study examines whether ChatGPT handles less terminology-rich texts, such as marketing materials, more effectively than

highly specialised texts containing domain-specific, system-bound and/or institution-specific terminology.

2 Method

This study adopts a qualitative research design based on manual analysis and manual terminology annotation of LLM-generated translations. The analysis is grounded in user prompts entered into ChatGPT, while the resulting translations were manually annotated according to the UniTermGPT terminology annotation guidelines (Heinisch, 2026). The prompt design was informed by *The Prompt Report: A Systematic Survey of Prompting Techniques* (Schulhoff & et al., 2024). In line with *The Prompt Report's* findings, few-shot prompting using exemplars was rejected, as the selection, quality and sequence of examples may substantially influence LLM output and thus reduce methodological consistency. Likewise, automatic prompt optimisation or prompt selection methods were not applied, since the study aims to reflect authentic user translation practices rather than developer-oriented optimisation strategies. Consequently, the prompts mirror the type of prompts commonly used by ChatGPT users for translation tasks. The study used ChatGPT 5 (GPT-5.5) and the prompt structure for the current ChatGPT model as suggested by OpenAI Developers (2026): As the translation task is rather straightforward and does not require the provision of much additional information, the prompts only contained the role, the goal and the success criteria. Therefore, the prompt for a translation from English into German was the following: "You are a professional translator for German and English at the University of Vienna. You are responsible for ensuring that the University of Vienna presents itself professionally in both German and English. Translate this [text type] from English into Austrian German for [target audience], in line with the Austrian higher education system. The translation must comply with the University's corporate language guidelines. Ensure that the entire text is translated, that the style is appropriate to the type of text, and that the terminology used corresponds to that of the University of Vienna."

The task consisted of bidirectional translation between German and English. For this purpose, excerpts of approximately equal length were selected from different university-related text types produced by the University of Vienna. All documents were available in both German and English. These served as reference texts or translations. Excerpts from these texts were translated in both directions (German- English and English-German). The selected text types were curricula, postdoctoral job advertisements, a development plan description, a marketing brochure, statutes, a web page providing information for prospective students and a university glossary, which was considered particularly terminology-rich and therefore especially challenging.

Although several frameworks for assessing translation quality and readability exist, including MQM (2024), few approaches focus specifically on terminology. MQM (Lommel et al., 2024) considers terminology as one quality aspect among many, whereas the ACTER terminology annotation guidelines (Rigouts Terryn, 2021) constitute, to the best of the author's knowledge, the only dedicated terminology annotation guidelines currently available (in English). While the ACTER guidelines are highly detailed and precise, their application requires considerable effort due to the differentiation between numerous annotation categories. For this reason, more intuitive and practice-oriented terminology annotation guidelines were developed for the broader UniTermGPT project (Heinisch, 2026).

The LLM-generated translations were manually annotated by an university and translation expert specialising in university terminology using these practice-oriented guidelines. Annotation was carried out by a single expert, as no second annotator with comparable domain expertise was available at the time of the study. Although terminology represents only one aspect of languages for specific purposes (alongside phraseology, stylistic, and syntactic features (ISO, 2019)), terminology remains a central challenge in the preparation and translation of specialised texts with LLMs. The present study therefore focuses specifically on terminology as a key indicator of translation quality in university contexts.

In contrast to quantitative evaluation metrics such as BLEU (Papineni et al., 2002), qualitative analysis enables a more comprehensive assessment of translation quality, particularly regarding terminology. It allows for the identification of cases in which a term is semantically appropriate but does not correspond to the institutionally preferred terminology of the University of Vienna. Such deviations may involve only minor orthographic or lexical differences, for example *fachlich einschlägig* (relevant to the subject) instead of the preferred form *facheinschlägig*, or the use of *dissertation* instead of the institutionally preferred term *doctoral thesis*.

3 Findings and outlook

The results indicate that ChatGPT generally handled university terminology, including proper names surprisingly well, particularly in translations into English. Institutional designations such as faculty names or administrative units, such as *StudienServiceCenter* (SSC) (Studies Service Center), *StudienServiceStelle* (SSS) (Studies Service Unit) and *Studienprogrammleitung* (SPL) (Directorate of Studies) were often reproduced correctly, including consistent spelling conventions and the appropriate use of acronyms. Even when German terms remained untranslated in the English source text, the model generated plausible translations for terms such as *Promotion* (doctoral graduation) and *Sponsion* (master's graduation). Another notable finding was the high degree of terminological consistency throughout individual texts, even in cases where the selected terminology did not correspond to the University of Vienna's preferred terms. For example, terms such as "degree programme" or "Master's programme" were used consistently across entire documents, and stylistic conventions such as the use of British English (as mentioned in the University of Vienna's Style Guide) emerged without explicit prompting.

At the same time, the analysis revealed recurring difficulties with highly specialised and systembound terminology, especially in the area of study administration and legal language. Some terms such as *Abgang von der Universität* (withdrawal from university), *abweichende Prüfungsmethode* (alternative mode of assessment), *prüfungsimmanent/nicht-prüfungsimmanent* (continuous assessment/noncontinuous assessment) or *Lehrveranstaltungsleiter*in* (lecturer) were translated word by word rather than providing either appropriate or institutionally preferred equivalents. Similar patterns appeared in strategic documents, where the terms such as *strategische Schwerpunkte* (strategic priorities) or *Nachwuchswissenschaftler*innen* (early-stage researchers) were appropriate in the English version but not according to institutional terminology preferences. In the reverse translation direction (English-German), terminology was likewise mostly accurate, although some legal and administrative terms caused problems. Beyond terminology, ChatGPT also struggled with German gender-inclusive language, producing grammatically or typographically inconsistent forms. In a text on alternative modes of assessment, the German translation reads: "Kann Ihre/r Prüferin bzw. Lehrveranstaltungsleiter*in Ihren Grund nicht anerkennen ...", combining the preferred gender-inclusive form (*Lehrveranstaltungsleiter*in*) with an unconventional juxtaposition of the masculine and feminine forms without punctuation marks (*Ihre/r Prüferin*, which could alternatively read *Ihr/e Prüfer/in*).

Overall, the findings suggest that ChatGPT performs well with institutional terminology and terminological consistency, but encounters difficulties with highly specialised and institution-specific terminology.

The results of this pilot study therefore suggest that the hypothesis that ChatGPT performs better on less specialised texts, such as marketing materials, due to their lower terminology richness, can only be partially confirmed. While more specialised text types, such as curricula and statutes, contain a higher amount of terminology and therefore allow greater potential for deviations and errors, the translations across text types were generally readable and mostly employed either the preferred terminology of the University of Vienna or otherwise appropriate terminology. However, beyond terminology, some translations (particularly the German version of the job advertisement) appeared stylistically clumsy and lacked fluency.

This study has several limitations: It focuses exclusively on ChatGPT, on university texts and on translation between German and English, which limits generalisability to other models, domains

and language pairs. Although RAG (retrieval-augmented generation) and TAG (terminology-augmented generation) offer promising avenues for improving specialised translation, both require substantial infrastructure, including up-to-date terminology resources, API (application programming interface) access and carefully selected reference materials. The annotation was conducted by a single expert, which precludes inter-annotator agreement analysis. Future research should include multiple annotators and consider semi-automatic approaches where reference translations are available. As a pilot study, no statistical testing was performed, with the analysis instead centred on qualitative terminology annotation.

4 Conclusion

Overall, the findings of this pilot study only partly support the hypothesis that ChatGPT performs more effectively on less specialised university texts. Although highly terminology-rich documents, such as curricula and statutes, showed a greater potential for variation in terminology use, the generated translations were generally comprehensible and terminologically appropriate across all text types. In many cases, ChatGPT correctly applied either the preferred terminology of the University of Vienna or acceptable equivalents. Nevertheless, the analysis also revealed limitations beyond terminology, particularly regarding stylistic fluency and the handling of German gender-inclusive language.

The paper is novel in its qualitative investigation of LLM-generated translations in the university domain, combining elements from terminology studies, translation studies and natural language processing (NLP) research, thereby highlighting the potential for cross-disciplinary innovation within the humanities.

Its impact on the research community lies in its focus on specialised language within a narrowly defined institutional domain, while simultaneously comparing different text types. Furthermore, the study offers a qualitative alternative to the current dominance of quantitative evaluation methods in NLP research by foregrounding terminology use, institutional (or corporate) language and contextual appropriateness. The societal impact of the study lies in its relevance for universities, translators and public institutions increasingly using LLMs for multilingual communication. By identifying the strengths and limitations of LLMs in handling specialised and institution-specific terminology, the study contributes to more reliable and precise communication in higher education and other domains.

References

- Bella, G., Helm, P., Koch, G., & Giunchiglia, F. (2024). Tackling language modelling bias in support of linguistic diversity. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3630106.3658925>
- Chiocchetti, E. (2021). Terminology as a societal resource. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 27(1), 110–139. <https://doi.org/10.1075/term.00060.chi>
- Heinisch, B. (2026). Developing an annotation scheme for terminology evaluation in large language model output for different german language varieties in the university domain [Conference held 21–22 September 2026]. *4th International Conference: AI-Generated Texts in Romance and Germanic Language Varieties (AI-ROM 2026)*, 10–12. <https://tu-dresden.de/gsw/slk/romanistik/ressourcen/dateien/das-institut/Professuren/prof-dr-anna-maria-de-cesare-greenwald/plakate/AI-ROM-IV-Book-of-abstracts.pdf/at/download/file>
- ISO. (2019). ISO 1087:2019 Terminology work and terminology science—Vocabulary.
- Lackner, A., Vega-Wilson, A., & Lang, C. (2025). Terminology augmented generation: A systematic review of terminology formats for in-context learning in llms. *3rd International Conference on Multilingual Digital Terminology Today (MDTT) 2025*, 1–9.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledgeintensive nlp tasks. In H. Larochelle, M. Ranzato, & R. e. a. Hadsell (Eds.), *Advances in neural information processing systems* (pp. 9459–9474). Curran Associates, Inc.
- Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Marazzato Sparano, R., Foresi, M., Innis, J., Han, L., & Nenadic, G. (2024). The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. In M. Martindale, J. Campbell, &

- K. e. a. Savenkov (Eds.), *Proceedings of the 16th conference of the association for machine translation in the Americas* (volume 2: Presentations) (pp. 75–94). Association for Machine Translation in the Americas.
- MQM. (2024). MQM (Multidimensional Quality Metrics). <https://themqm.org/the-mqm-full-typology/>
- OpenAI Developers. (2026). Prompt guidance: Gpt-5.5 prompting guide. <https://developers.openai.com/api/docs/guides/prompt-guidance>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation [40th Annual Meeting, 7–12 July 2002]. In P. Isabelle (Ed.), *Proceedings of the 40th annual meeting on association for computational linguistics – ACK'02* (p. 311). Association for Computational Linguistics.
- Ralli, N., Stanizzi, I., & Wissik, T. (2007). Der Bologna-Prozess und die Folgen für die italienische und österreichische Universitätsterminologie. In N. Niemelä & E. Lehtinen (Eds.), *Vakki-symposium xxvii* (pp. 9–19).
- Rigouts Terryn, A. (2021). ACTER Terminology Annotation Guidelines. <http://hdl.handle.net/1854/LU-8503113>
- Sandrini, P. (2018). Terminology work in different domains: Legal terminology. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Language for special purposes. an international handbook* (pp. 509–521). De Gruyter Mouton.
- Schulhoff, S., & et al. (2024). The prompt report: A systematic survey of prompting techniques. <https://doi.org/10.48550/arXiv.2406.06608>
- Temmerman, R., & van Campenhoudt, M. (Eds.). (2014). *Dynamics and terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication*. John Benjamins Publishing Company.
- Van der Yeught, M. (2016). Concepts and frameworks in English for specific purposes. *ASp*, 69, 41–63. <https://doi.org/10.4000/asp.4788>

Berenike Herrmann

Computational Literary Studies, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany, berenike.herrmann@uni-bielefeld.de

Andreas Witt

Digital Linguistics, Leibniz Institute for the German Language (IDS), Mannheim, Germany, witt@ids-mannheim.de

Textification. On the role of text as an interface in the age of LLMs¹

Keywords: large language models, textification, machine learning, text as data representation

1 Introduction

This paper introduces the term and the concept textification. This concept allows to model and to describe the states and the processes of modern infrastructures in the age of Large Language Models. Across disciplines, these processes include transformation of data-analytic and machine learning processes into language-mediated, semantically interpreted workflows, and transformation of data. As a central layer in human-machine interaction, text takes on the role of a method for describing, interpreting, generating, and structuring knowledge and may serve as a means for humans to understand certain states when machines process data. Viewed from this angle, textification is a way to make processing in the times of AI and ML more accessible and understandable.

2. What is Textification

'Text' — broadly understood as written natural language — has had a pivotal role in the history of human thought as a mode of knowledge transmission. In the humanities and social sciences, text has been also the main instrument for abstraction, categorization and analysis, as well as argumentation and critique. In the natural sciences and engineering, by contrast, instrumentation has required more formalization, with programming languages and formal architectures to model and compute information as data².

After the spreading of Large Language Models, text becomes more and more relevant also in natural and life sciences and in engineering — and emerges as a mediator for datafication of the SSH. Textification is emerging as a new paradigm: Text, represented in digital form, takes on a new role as input as well as output for data processing: text is used to generate information and better texts, as well as pictures or videos (prompting), to create and optimize computer programs (vibe programming), to carry out DNA sequencing, or to process medical multi-modal data and so on. But, text is also generated by machines for example to describe pictures and videos, to summarize longer texts, to transcribe speech, to translate texts, and to provide potential medical diagnoses.³

It is evident that textification has implications for research data infrastructures. While text has long been predominately a medium for knowledge generation in/by humans as well as one of transfer between humans (e.g., by printed books, their digital simulations as PDFs), it is now a

¹ First published as: Berenike Herrmann & Andreas Witt (2026), "Textification. On the role of text as an interface in the age of LLMs", in M. Wynne & T. Kontino (eds.), *CLARIN Annual Conference Proceedings 2026*, pp. 71–75. <https://doi.org/10.5281/zenodo.22916661>

² The term text denotes quite different concepts. In the context of information technology, text refers to textual data, that is, a sequence of human-readable characters which, although readable by humans, is primarily designed for processing by computers. This understanding of text is exemplified by the many MIME types beginning with `text/`, such as, `text/html` or `text/javascript` (cf. Freed and Borenstein, 1996). Data denotes a "reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing", which can be processed by humans or by automatic means. ISO-IEC-2382-1, 2015

³ It is important to note that textification means the data-agnostic interaction with any type of data and information and is thus not limited to textual foundation models.

key technique for data manipulation. Textification means understanding and curating written language to connect different kinds of actors (human and non-human), as well as types of knowledge, information and data (and maybe even wisdom, see Rowley, 2007). Among the main challenges to infrastructure are implementing and evaluating ‘textual operations’ in domain-specific research repertoires. Important questions concern the domain-specific use of benchmarks – depending on the domain, these can be deeply textified or at the fringes of textification.

3 A semi-formal description of textification

3.1 Text and Textuality

Text is a stretch of written (or transcribed spoken) natural language that is interpretable by humans and machines. More specifically, text is a typically discrete, meaningful, and interconnected sequence of symbols and characters, such as words or other character arrangements that can be regarded as a unit (Horstmann, 2007).¹

In research infrastructures, text is represented (or representable) in digital form, that is, readable both by humans and machines. Text is thus data “in the form of characters, symbols, words, phrases, paragraphs, sentences, tables, or other character arrangements, intended to convey a meaning, and whose interpretation is essentially based upon the reader’s knowledge of some natural language” ISO-IEC-2382-1, 2015. By contrast to ISO-IEC-2382-1, 2015 we thus delimit text to such sequences that can be understood as natural language. In textified research infrastructures, different dimensions of textuality may become relevant, including a text’s internal interconnectedness (linearity, hierarchical nature, coherence/cohesion, grammar and meaning, complexity and scope (from sub-word level to comprehensive supra-segmental ensembles), communicative dimensions (human and non-human actors, intentionality and goals), as well as pragmatic and cultural constraints, including acceptability or informationability, or formal and medial representation (spoken, written, diagrammatic; digital encoding such as UTF-8, ASCII.)

3.2 Textification and Data Processing

Textuality cannot be reduced to the text-as-data compressed and stored within the LLMs, as ‘text’ is intuitively understandable by humans. The current ‘chatbot’ versions of LLMs typically operate with foundation models² that are typically text-based and instruction-tuned (IT)³. Thus, both the input and the output of LLMs are commandable by natural language.

We know that human language is neither innate nor ‘statistical noise’ – but that it comes in distributional patterns that carry and compresses information. The emergent character of meaning and grammar from usage may not only drive the transfer learning on neural networks that is at the core of LLMs (Millière, 2024), but enable textification: text seems to have an optimum degree of informational complexity to compress information, but also to connect humans and machines, being more accessible to humans than other knowledge representation formats (e.g. Knowledge Graphs or the base-2 numeral system). Tasks include domain-specific prompt engineering (systemic, beyond the individual prompt), but also LLM Agents (where textual operations manage computational tool use via language models), Natural Language to SQL (NL2SQL – translation from natural language to database queries, incorporating ML pipelines), Automated model selection (AutoML, where text is the control interface for engineering models for machine learning), or RAG (combining information retrieval and LLMs, a component within textified systems).

¹ As the concept of ‘text’ is culturally and historical variable (see Scherner, 1996), a prototype-approach (Winko, 2008) requires only minimal assumptions.

² A foundation model is “any model that is trained on broad data that can be adapted to a wide range of downstream tasks”(Bommasani et al., 2022, p.1).

³ IT is also termed supervised fine-tuning (SFT) and denotes the additional training of LLMs using (INSTRUCTION, OUTPUT) pairs, where INSTRUCTION denotes the human instruction for the model, and OUTPUT denotes the desired output that follows the INSTRUCTION (Zhang et al., 2024, p.1)

4 Text was there: Textification from and of the Humanities

With the advent of generative LLMs, tasks such as writing, reading, and analyzing texts in the domains of culture, science, law, and politics (e.g., policies, legal documents, speeches) seem to be increasingly delegable to machines. Some see reason to declare a 'crisis of the Humanities'. We, however, argue that given the ongoing datafication and computerization of the world, textification not only connects the Humanities to the modern data paradigm, but also draws on their insights on 'text', as the medium for the generation, transmission, and critique of knowledge:

(1) Language does not just reflect 'reality', or is shaped by contextual, social and historical forces, but rather shapes them as well in a reciprocal relationship. (2) The constitutive and gradable uncertainty of linguistic meaning rests on the distribution of text in contexts, which regulates vagueness and precision of meaning. (3) Natural language is self-contained in that it comprises its own metalanguage, which allows for metacommunication and methodology (Harris, 2002; Jakobson, 1985). Using the same set of symbols and the same generative logic for communication (generation and transmission of information) as for metacommunication (method), the recursive character of text appears to offer a needed flexibility for abstraction, redundancy, and compression of information.

While it is domain-general, textification has the potential to provide an environment for data-work that is very close to habitual practices of reading and writing in the SSH. Users may not even need to acquaint themselves with technical software and editors, learn code, or use formal databases. Textification in the SSH may thus be a radical potentialization of the "Aufschreibesysteme" (Kittler, 1990) based on vast amounts of data, and offering powerful inference and reasoning systems. It is a non-trivial challenge of textification especially in the SSH that the inherent data-nature of the conducted research can get out of sight, with the empowering as well as problematic dimensions of postdigitality (Jandrić et al., 2018, p. 893). One of the challenges for building textified infrastructures is thus to equip users with data / AI literacy: the ability to understand, use, evaluate, and ethically navigate tools and infrastructure that make use of LLMs and are thus data-driven and data-oriented (cf. Southworth et al., 2023). Textified data literacy means data awareness as well as evaluation understood as (1) domain competency potentially including formal benchmarking of LLMs, and (2) standards for quantifying and ML operations (F1 scores for classifiers, inferential or Bayesian statistics etc.).

5 Applications of Textification

5.1 Textification – Data Beyond Text

Because it provides intuitive access to compute and data, textification plays a pivotal role in the ongoing datafication both in the SSH (e.g., Computational Literary Studies, Digital History, or Computational Social Science) as well as within STEM (e.g. Molecular Chemistry and Biology). Across disciplines, LLMs are being fused with non-textual data, creating for example genomic language models (gLM), Audio Large Language Models (AudioLLMs), or Multimodal AI more generally. All those data are manipulable through text and interfaces: the advances of textification thus include text summarization and generation as well as image, video and sound description and generation; text, image, video and sound annotation and analysis; code generation; natural language interfaces to relational databases; DNA sequencing, and much more.

5.2 Textification and Data Infrastructures

In the context of contemporary data infrastructures, textification calls for new approaches in both usage and building of infrastructure. Rather than treating text merely as a representational layer, infrastructures increasingly need to accommodate text as a primary interface through which data are accessed, processed, and interpreted. This shift affects not only how users interact with resources, but also how infrastructures are conceived, implemented, and sustained over time.

Text-driven computational applications already demonstrate how text can function as a central interface across heterogeneous data types. The automatic description of images, the generation of images from textual prompts, and the transformation of spoken language into written form

(transcription) demonstrate that diverse modalities are increasingly translated into text in order to enable uniform processing, interpretation, and reuse. Infrastructures, should support and standardize these processes as part of their core functionality.

This implies that text should be treated as a primary operational layer rather than as a secondary representation. From a philosophy-of-science perspective, one might say that text should be modeled and implemented also as a metalanguage. The growing practice of generating programming code from textual prompts indicates that text can also serve as a control mechanism for computational processes — a development that infrastructures should explicitly accommodate and support. Similarly, in Computational Literary Studies and Corpus Linguistics, established practices such as annotation can be increasingly automated, allowing non-annotated resources to be enriched on the basis of textual input, that is, through natural-language descriptions of the desired annotation. Taken together, these developments suggest that textification should be addressed proactively as a guiding principle for the further evolution of — language centered — data infrastructures.

6 Textification as conceptual reorientation

Textification is not only a technical challenge, but a conceptual reorientation. It is the transformation of data-analytic and machine learning processes into language-mediated, semantically interpreted data workflows. As a central layer in human-machine interaction, text takes on the role of a method for describing, interpreting, generating, and structuring knowledge from and as various types of data. With domain-general and domain-specific applications, textification may thus serve as a means for humans to understand and access Large Language Models and their derivatives in state-of the art infrastructures.

References

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022). On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). <https://doi.org/10.48550/arXiv.2108.07258>
- Freed, N., & Borenstein, N. S. (1996). Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. <https://doi.org/10.17487/RFC2046>
- Harris, Z. S. (2002). The background of transformational and metalanguage analysis. In B. E. Nevins (Ed.), *Current issues in linguistic theory* (pp. 1–15, Vol. 228). John Benjamins Publishing Company. <https://doi.org/10.1075/cilt.228.03har>
- Horstmann, S. (2007). Text. In K. Weimar, H. Fricke, K. Grubmüller, & J.-D. Müller (Eds.), *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte* (pp. 594–597). de Gruyter.
- ISO-IEC-2382-1. (2015). *Data, Information Technology — Vocabulary* (1.0). IEC.
- Jakobson, R. (1985). Metalanguage as a Linguistic Problem. In *Volume vii contributions to comparative mythology* (pp. 113–121). De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110855463.113>
- Jandrić, P., Knox, J., Besley, T., Ryberg, T., Suoranta, J., & Hayes, S. (2018). Postdigital science and education. *Educational Philosophy and Theory*, 50(10), 893–899. <https://doi.org/10.1080/00131857.2018.1454000>
- Kittler, F. A. (1990). *Discourse Networks 1800*. Stanford University Press, 1990.
- Millière, R. (2024). Language Models as Models of Language (arXiv:2408.07144). <https://doi.org/10.48550/arXiv.2408.07144>
- Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2), 163–180. <https://doi.org/10.1177/0165551506070706>
- Scherner, M. (1996). Text: Untersuchungen zur Begriffsgeschichte. *Archiv für Begriffsgeschichte*, 39, pp. 103–160.
- Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. 4, 100127. <https://doi.org/10.1016/j.caeai.2023.100127>
- Winko, S. (2008). Textualitätsannahmen und die Analyse literarischer Texte. *Zeitschrift für germanistische Linguistik*, 36(3), 427–443. <https://doi.org/10.1515/ZGL.2008.035>
- Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2024). Instruction Tuning for Large Language Models: A Survey (arXiv:2308.10792). <https://doi.org/10.48550/arXiv.2308.10792>

Sahil Kale

University of California, Los Angeles, CA, USA, sahilrkale@ucla.edu

Analysing Temporal Self-Knowledge Signals in Large Language Models

Keywords: large language models, self-knowledge calibration, uncertainty estimation

Large language models (LLMs) are increasingly expected to estimate uncertainty in their own responses to support retrieval triggering, adaptive compute allocation, selective abstention, and tool use. In this work, we study whether confidence signals measured at different temporal stages of reasoning encode distinct forms of self-knowledge. Inspired by meta-memory theory, we distinguish between anticipatory pre-solve confidence (*Feeling-of-Knowing*, FOK) and retrospective post-solve confidence (*Judgement-of-Learning*, JOL). Across frontier LLMs and reasoning domains, we find that pre-solve confidence is consistently weakly calibrated, while post-solve confidence generally improves reliability despite substantially higher inference cost and strong domain dependence. We further show that hallucinated responses maintain persistently elevated confidence even after reasoning generation, suggesting important limitations in current LLM self-monitoring abilities. Motivated by these findings, we outline ongoing work investigating whether stronger retrospective confidence signals can be distilled into cheaper anticipatory selfknowledge representations using internal model logits.

I Introduction

Recently, large language models are being deployed in systems that require reliable estimates of their own uncertainty to trigger verification or additional support, including retrieval-augmented generation, tool-using agents, and adaptive reasoning pipelines (Cheng et al., 2025; Li et al., 2024). In these settings, models are expected not only to generate answers, but also to assess whether those answers are likely to be correct, a capability often framed as *self-knowledge* or *metacognition* (Kale & Nadadur, 2025; Ma et al., 2025). However, recent work shows that LLM confidence remains poorly calibrated under many reasoning settings, with models frequently exhibiting overconfidence and hallucination-prone certainty (Kale & Dhimi, 2025; Lin et al., 2022).

Most prior work treats confidence as a static signal accompanying the final generated output. In contrast, we study whether confidence changes across different temporal stages of reasoning. Drawing inspiration from classical meta-memory research (Koriat, 1997; Nelson, 1990), we distinguish between two complementary forms of confidence: anticipatory pre-solve confidence, analogous to *Feeling-of-Knowing* (FOK), and retrospective post-solve confidence, analogous to *Judgement-of-Learning* (JOL). FOK measures a model’s estimated likelihood of solving a task correctly before reasoning begins, while JOL measures confidence in a completed reasoning trajectory and final answer. Our work is thus guided by two questions: *RQ1*: Which temporal confidence signal, pre-solve FOK or post-solve JOL, better aligns with answer correctness to support downstream decisions? *RQ2*: Do FOK/JOL confidence patterns correspond to distinct failure types such as hallucinations, reasoning failures, or knowledge gaps?

We evaluate these temporal confidence signals across frontier LLMs and multiple domains spanning mathematical reasoning, logical reasoning, and factual recall. Our findings show that the inexpensive pre-solve confidence is consistently weakly calibrated, while post-solve confidence generally provides better estimates but with substantially higher inference cost. We additionally observe that hallucinated responses maintain elevated confidence even after reasoning generation, suggesting that reasoning traces may sometimes reinforce rather than correct false certainty.

Beyond analysing these signals descriptively, our broader research agenda investigates whether stronger retrospective confidence signals can supervise improved anticipatory self-knowledge representations. Motivated by the calibration gap between FOK and JOL, we outline ongoing

work exploring whether post-solve confidence behaviours can be distilled into cheaper pre-solve signals using internal logits and hidden-state representations.

2 Temporal Self-Knowledge Framework

We operationalise temporal self-knowledge using a three-stage prompting pipeline illustrated in Figure 1. First, the model receives only the input question and estimates the likelihood of solving it correctly before reasoning begins, producing a pre-solve *Feeling-of-Knowing* (FOK) confidence signal. Second, the model generates an unconstrained answer with chain-of-thought reasoning. Third, the model is shown its generated answer and estimates the probability that the answer is correct, producing a retrospective *Judgement-of-Learning* (JOL) confidence signal. Because our evaluation includes closed-source frontier models accessible only through APIs, we use prompting-based verbal confidence estimates as the primary self-knowledge signal.

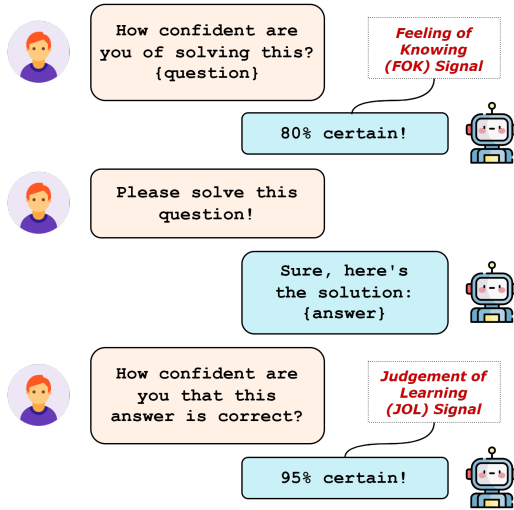


Figure 1: Temporal confidence framework.

We evaluate this framework using GPT-5.4 and Claude Sonnet 4.6 across three domains spanning mathematical reasoning, logical reasoning, and factual recall. For mathematical reasoning, we use 500 samples from the Mathematics Stack Exchange split of RealMath (Zhang et al., 2025). Logical reasoning is evaluated using 400 samples from RiddleBench (Halder et al., 2025), while factual recall is evaluated using 750 randomly sampled questions from TriviaQA (Joshi et al., 2017). All experiments use deterministic decoding with temperature set to 0.

For each problem instance, we obtain a pre-solve confidence estimate, a generated answer, and a postsolve confidence estimate. Generated answers are evaluated against benchmark reference solutions to obtain correctness labels. We then evaluate confidence quality using Expected Calibration Error (ECE) and Brier Score (Naeini et al., 2015), alongside AUROC to measure discrimination between correct and incorrect generations. We additionally analyse inference cost trade-offs and confidence dynamics across manually annotated error categories including hallucinated knowledge, reasoning failures, and knowledge gaps.

3 Results and Findings

Post-solve confidence generally provides stronger epistemic alignment, pre-solve confidence is not calibrated for usage. As seen in Table 1, across most settings, retrospective post-solve confidence (JOL) exhibits lower calibration error and stronger discrimination than anticipatory pre-solve

confidence (FOK), particularly in reasoning-heavy domains. The consistently weaker calibration performance of the cheaper FOK signal suggests that anticipatory self-knowledge remains insufficiently reliable for downstream decisions such as routing, retrieval triggering, or adaptive compute allocation.

Reasoning domains exhibit substantially higher calibration instability than factual recall. Both models remain notably miscalibrated on mathematical and logical reasoning despite moderate discrimination performance, while TriviaQA produces comparatively smaller temporal shifts between FOK and JOL confidence.

Surprisingly, the transition from anticipatory to retrospective confidence is not uniformly beneficial. Claude Sonnet 4.6 on logical reasoning exhibits degraded calibration after reasoning generation, with ECE increasing from 0.341 to 0.472. This suggests that extended reasoning traces may sometimes reinforce false certainty rather than improve epistemic self-monitoring.

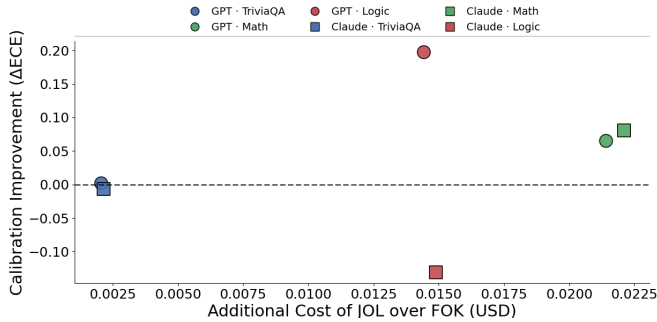


Figure 2: Calibration improvement versus additional inference cost for JOL relative to FOK.

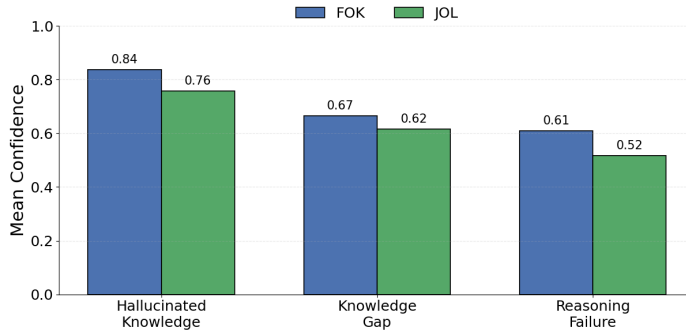


Figure 3: Mean FOK and JOL confidence across annotated error categories.

Table 1: Calibration and discrimination performance across domains and confidence stages

Model	Domain	Pre-Solve (FOK)			Post-Solve (JOL)		
		ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑
GPT-5.4	TriviaQA	0.195	0.209	0.567	0.194	0.203	0.606
	Math	0.381	0.366	0.719	0.315	0.330	0.751
	Logic	0.483	0.469	0.679	0.286	0.313	0.690
Claude Sonnet 4.6	TriviaQA	0.071	0.119	0.656	0.077	0.114	0.703
	Math	0.188	0.231	0.500	0.107	0.140	0.791
	Logic	0.341	0.316	0.767	0.472	0.456	0.695

In terms of cost, JOL generally improves calibration in structured reasoning domains, but factual recall tasks show only marginal gains despite substantially higher token usage. The cost-benefit analysis estimates the trade-off between the additional inference cost required to obtain post-solve confidence (JOL) relative to pre-solve confidence (FOK) and the calibration improvement achieved, as seen in Figure 2. FOK cost includes only the initial confidence estimation prompt, while JOL includes both answer generation and post-answer confidence estimation. Post-solve confidence becomes more informative only as reasoning complexity increases and can justify the additional inference budget in structured problem-solving settings. However, in some settings like factual recall, additional reasoning increases inference cost with no improvement in calibration.

Error-type analysis further reveals that hallucinated responses maintain persistently elevated confidence both before and after reasoning generation. After manual annotation, hallucinations remain highly confident showing overconfident internal representations that remain stable throughout answer generation, while reasoning failures exhibit lower, unstable confidence patterns, as seen in Figure 3. Knowledge gaps occupied an intermediate regime, where models often began uncertain but occasionally increased confidence after producing an answer.

4 Ongoing Work

The observed calibration gap between anticipatory and retrospective confidence signals motivates an important follow-up question: can stronger post-solve confidence behaviours be distilled into pre-solve self-knowledge representations to obtain a cheap yet well-calibrated self-knowledge signal? Our ongoing work investigates whether retrospective confidence signals can supervise improved anticipatory confidence estimation using internal logits and hidden-state representations prior to reasoning generation. Preliminary experiments span factual recall and reasoning-heavy domains including mathematics and logical reasoning, with the goal of obtaining pre-solve confidence estimates that better approximate the reliability of post-solve self-assessment without incurring the cost of full reasoning generation.

5 Conclusion

We introduced a temporal framing of self-knowledge in large language models by separating anticipatory pre-solve confidence from retrospective post-solve confidence. Across frontier LLMs, post-solve confidence generally provides stronger epistemic alignment but at substantially higher and sometimes unjustified inference cost. Our findings further suggest that hallucinated responses remain persistently overconfident even after reasoning generation. Overall, the results indicate that confidence-aware downstream systems should account for domain and cost when using temporal LLM uncertainty signals.

References

- Cheng, J., Ghate, K., Hua, W., Wang, W. Y., Shen, H., & Fang, F. (2025, July). REALM: A dataset of real-world LLM use cases. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: ACL 2025* (pp. 8331–8341). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.437>
- Halder, D., Saji, A., Jayakumar, T., Puduppully, R., Kunchukuttan, A., & Dabre, R. (2025). Riddlebench: A new generative reasoning benchmark for llms. <https://arxiv.org/abs/2510.24932>
- Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.
- Kale, S., & Dhama, D. S. (2025). Knowrl: Teaching language models to know what they know. <https://arxiv.org/abs/2510.11407>
- Kale, S., & Nadadur, V. (2025, May). Line of duty: Evaluating LLM self-knowledge via consistency in feasibility boundaries. In T. Cao, A. Das, T. Kumarage, Y. Wan, S. Krishna, N. Mehrabi, J. Dhamala, A. Ramakrishna, A. Galystan, A. Kumar, R. Gupta, & K.-W. Chang (Eds.), *Proceedings of the 5th workshop on trustworthy NLP (TRUSTNLP 2025)* (pp. 127–140). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.trustnlp-main.10>

- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Li, J., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., Lin, Y., Xu, B., Ren, B., Feng, C., Gao, Y., & Huang, H. (2024, August). Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11116–11141). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.599>
- Lin, S., et al. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.
- Ma, Z., Yuan, Q., Wang, Z., & Zhou, D. (2025, November). Large language models have intrinsic metacognition, but need a good lens. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 3460–3477). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.171>
- Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2901–2907.
- Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.). Academic Press. [https://doi.org/https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/https://doi.org/10.1016/S0079-7421(08)60053-5)
- Zhang, J., Petrucci, C., Nikolić, K., & Tramèr, F. (2025). Realmath: A continuous benchmark for evaluating language models on research-level mathematics. <https://arxiv.org/abs/2505.12575>

Noémi Ligeti-Nagy

Mátyás Osváth

ELTE Research Centre for Linguistics, Budapest, Hungary, {ligeti-nagy.noemi, osvath.matyas}@nytud.hu

Does Answer Format Matter? Multiple-Choice vs. Open-Ended Evaluation of Cultural Knowledge in Hungarian LLMs

Keywords: cultural knowledge evaluation, LLM benchmarking, Hungarian NLP, multiple-choice vs. open-ended evaluation, LLM-as-judge

I Introduction

More and more studies ask whether LLMs really know things that are specific to a culture, and not only the language. Benchmarks such as CulturalBench (Chiu et al., 2025), BLENd (Myung et al., 2024), SANSKRITI (Maji et al., 2025), BLUCK (Kabir et al., 2025), CARTE (Almeida Carneiro et al., 2026) and Global MMLU (Singh et al., 2025) show large gaps between high-resource and low-resource cultures. Global MMLU, for example, finds that 28% of MMLU questions need culturally sensitive knowledge, mostly Western, and that model rankings change when we look only at these questions. This work shows that cultural knowledge in LLMs is uneven and often overestimated. It says less about how we measure this knowledge, and whether the measurement itself changes the result.

Almost all of these benchmarks, including the cultural module of HuGME, use multiple-choice questions. This is practical, because one letter can be checked against a key without any human or automatic judgment. But there are known problems. Models are sensitive to the order and form of the options and can have position biases (Robinson et al., 2023). They can also find the answer by elimination instead of recall (Myrzakhan et al., 2024). In CulturalBench, the same facts asked in two formats gave a difference of about 27 points for GPT-4o (Chiu et al., 2025). Other studies that make open-ended versions of closed benchmarks report similar effects (Liu et al., 2024; Tanwar et al., 2025; Yoo et al., 2026): without options, not only the scores change, but also which items and models look strong.

Our question is simple: when a benchmark says that a model “knows” a Hungarian cultural fact, is this knowledge, or partly test-taking skill? In practice: how much of an ABCD score is left when we remove the options?

HuGME (Hungarian Generative Model Evaluation) tests language skills, world knowledge and safety behaviour of LLMs in Hungarian (Ligeti-Nagy et al., 2025). Its cultural module was written in Hungarian by domain experts, not translated. Until now it was only used in ABCD format. We add a parallel open-ended version of *all* 2,641 questions, and compare what models can choose with what they can say on their own. The method is not specific to Hungarian. It can be used to check, for any language, whether a “cultural knowledge” benchmark measures knowledge or test-taking.

2. Data and method

2.1 ABCD datasets and results

The dataset has 2,641 multiple-choice questions in 16 categories, each written and reviewed by domain experts. The biggest categories are Literature (402 items), Music (255), Live-action film (242), Folk customs (233) and Food and drink (218); the smallest is Radio (34). Each item has a question, four options and one correct letter. The questions need Hungarian knowledge, for example about Hungarian films, books, inventors and folk traditions.

Table 1 shows the ABCD results for six models. Accuracy follows general model strength more than Hungarian-specific training. The Hungarian PULI-chat is about as good as the general Qwen3-8B (45.7% vs. 47.7%), and Racka-4B is 12 points below its base model Qwen3-4B and close to chance (25%). Geography is the hardest category for five of the six models (22.3% on average, so around chance), and inventions and scientists is the easiest (60.6%).

Table 1: ABCD accuracy per model: overall, best and worst of 16 categories. Chance level is 25%.

Model	Org.	Overall	Best category	Worst category
gpt-4-0613	OpenAI	60.1%	Inventions (84.1%)	Geography (21.1%)
Qwen/Qwen3-8B	Qwen	47.7%	Food/drink (70.2%)	Geography (25.3%)
NYTK/puli-chat	NYTK	45.7%	Inventions (61.9%)	Geography (23.2%)
Qwen/Qwen3-4B	Qwen	40.2%	Food/drink (59.2%)	Geography (24.2%)
gpt-3.5-turbo	OpenAI	39.1%	Inventions (60.3%)	Geography (19.1%)
elte-nlp/Racka-4B	ELTE NLP	28.0%	Inventions (41.3%)	Radio (17.7%)

2.2 Open-ended version

We remove the four options and ask each question in Hungarian with a short instruction. The 2,641 items have three answer types: *entity* (1,105; a name, place, date or term), *short answer* (759; a short fact) and *explanation* (776; a longer answer). Each item has a gold answer. Entity items also have accepted aliases, and the other items have a rubric with required elements and critical errors. Models had a system prompt and a limit of about 8k generated tokens. The Qwen3 models ran in their default thinking mode, and we score only the text after the thinking block.

Entity items are scored by string matching. We normalise the answer (lower case, no accents or punctuation) and compare it with the gold answer and its aliases. An answer is *strict-correct* if it is equal to an alias, starts with an alias, or contains an alias and has at most 12 tokens. A *lenient* score also accepts stem matches (to handle Hungarian suffixes, e.g. *Budapest* and *Budapesten*) and different word order (*Munkácsy Mihály* and *Mihály Munkácsy*). *Short-answer and explanation items* are judged by an LLM (gpt-4o). It gets the question, the reference answer and the rubric, and gives one label: correct, partially correct, incorrect or uncertain. “Uncertain” is at most 1.4% for every model except Racka-4B. Our main measure is *strict accuracy* (only “correct”), because ABCD is also all-or-nothing. The lenient score also counts partially correct answers. Confidence intervals are 95% bootstrap intervals over items. We have not yet checked the gpt-4o judge against human ratings, so the judged results are provisional.

3 Results

3.1 Open-ended accuracy

Open-ended questions are much harder than ABCD questions (Table 2). Even GPT-4 gets only 18.8% strictly (34.8% leniently), and the Qwen3 models are close to zero. GPT-3.5 and PULI-chat are not clearly different (their intervals overlap). Within a model family, the bigger model is better for every answer type. Entity items give the lowest strict score for four of five models, but this also reflects how strict string matching is (the lenient entity scores are 1.1–2.0 times higher).

3.2 ABCD compared with open-ended

Table 3 shows three things. First, every model is much worse without options. This fits the old finding that recognition is easier than recall (Mandler, 1980). The gap is 30–44 points, and the largest for Qwen3-8B (44.4) and GPT-4 (41.3).

Second, the formats agree on categories, but not on models. Mean ABCD and mean open-ended accuracy of the five models are strongly correlated across the 16 categories (Spearman $\rho = 0.79$). The ranking of models is only weakly kept ($\rho = 0.30$, $n = 5$; we do not claim significance). GPT-3.5 moves from fifth place to second. On ABCD it is about 1 point behind Qwen3-4B and 8.7 points behind Qwen3-8B. Open-ended, it is 4.6 times better than Qwen3-4B and 2.8 times better than Qwen3-8B (9.2% vs. 2.0% and 3.3%; the intervals do not overlap). PULI-chat is 6.6 points ahead of GPT-3.5 on ABCD, but they are tied open-ended (8.0% vs. 9.2%). So the ABCD advantage of the Hungarian-focused model does not appear when models must produce the answer. GPT-3.5 gets the smallest help from the options (30 points, against 38–44 for the others).

Table 2: Table 2: Open-ended accuracy (%) on all 2,641 questions: strict accuracy by answer type and overall, 95% interval of the overall strict score, and lenient overall score.

Model	Entity (1105)	Short (759)	Expl. (776)	Overall		
				strict	95% CI	lenient
GPT-4	16.7	19.9	20.5	18.8	17.3–20.2	34.8
GPT-3.5-turbo	7.2	11.7	9.5	9.2	8.1–10.3	22.5
PULI-chat	8.5	8.8	6.3	8.0	6.9–9.0	18.7
Qwen3-8B	1.6	5.4	3.6	3.3	2.7–4.0	9.0
Qwen3-4B	0.9	3.8	1.7	2.0	1.5–2.5	5.6
Racka-4B*	2.2	–	–	–	(1.4–3.1)	–

*Entity items only, scored on the first line of the raw output (see §3.3).

Racka-4B has no usable final answer for most items, so the judged types cannot be scored.

Table 3: Table 3: ABCD and strict open-ended accuracy (%) on the same 2,641 questions, with ranks among the five models with complete results. “Chance-corr.” is the ABCD score rescaled so that 25% becomes 0 and 100% stays 100.

Model	ABCD	Chance-corr.	Open	ABCD–Open	ABCD rank	Open rank
GPT-4	60.1	46.8	18.8	41.3	1	100%
Qwen3-8B	47.7	30.3	3.3	44.4	2	400%
PULI-chat	45.7	27.6	8.0	37.7	3	300%
Qwen3-4B	40.2	20.3	2.0	38.2	4	500%
GPT-3.5-turbo	39.1	18.7	9.2	29.8	5	200%
Racka-4B	28.0	4.0	2.2†	–	–	–

Third, open-ended scoring separates categories that look similar on ABCD (Table 4). Advertising (35.9% on ABCD) and animated film (38.5%) are above chance, but open-ended accuracy is only 1.4% and 2.2%. Models can recognise the right slogan or cartoon among four options, but almost never produce it. Geography is at chance level on ABCD, but it is not the lowest category open-ended (3.8%).

Table 4: Table 4: Mean accuracy (%) of the five scored models by category, ABCD and strict open-ended, sorted by ABCD.

Category	n	ABCD	Open	Category	n	ABCD	Open
Inventions	63	64.4	21.0	Literature	402	45.6	6.2
Food/drink	218	61.2	12.8	Television	209	44.1	5.6
Folk customs	233	60.1	8.3	Live-action film	242	42.8	6.4
Nature	127	55.0	12.8	Animated film	156	38.5	2.2
Fine arts	82	53.2	11.7	Theatre	75	38.1	8.5
Sport	158	49.9	14.7	Advertising	58	35.9	1.4
Architecture	134	46.6	12.1	Radio	34	31.8	7.1
Music	255	46.0	7.4	Geography	194	22.6	3.8

3.3 Racka-4B and its base model

Racka-4B is a continual-pretraining adaptation of Qwen3-4B on Hungarian-heavy data. It is not instruction-tuned or aligned (Csibi et al., 2026), so we can compare it with its base model. On ABCD, Racka-4B (28.0%) is 12.2 points below Qwen3-4B (40.2%) and close to chance. Taken alone, this would suggest that the Hungarian adaptation damaged the model’s cultural knowledge.

Open-ended generation shows a different problem. Racka-4B often does not end its answers. 56% of its generations reach the 8k-token limit, often in repetition loops or by repeating the

prompt. For 87% (2,287 of 2,641) there is no final answer after we remove the thinking block. With standard parsing it scores 0% on entity items, and the gpt-4o verdicts on the other types (885 “uncertain”) are not meaningful. This is a format problem, not necessarily a knowledge problem. So we also score the *first line* of the raw output on entity items, where scoring is only string matching. On these 1,105 items, Racka-4B gets 2.2% (95% CI 1.4–3.1) and Qwen3-4B gets 0.9% (0.4–1.5). Racka-4B alone gets 21 items right that Qwen3-4B misses, and the reverse is true for 7 items (exact McNemar $p \approx 0.013$).

So the two formats point in opposite directions: ABCD says the adaptation lost knowledge, and open-ended entity scoring says it gained a little. We do not take either as the final answer. Both models are near zero open-ended, so the difference is small in absolute terms. The first-line rule is also generous for a model that does not stop. On the ABCD side, we cannot check whether the 12-point gap comes from Racka-4B not answering with a single letter, because we do not have the ABCD generations. What we can say is that a non-aligned adaptation model can look much worse or a little better than its base model, depending on the format, and that in open-ended evaluation it loses mainly because it does not stop, not because it lacks knowledge.

4 Discussion

Format matters in three ways.

First, the options help every model a lot. The gap between ABCD and open-ended accuracy is 30–44 points for all models, and even after correcting for chance, ABCD accuracy is 2–10 times the open-ended accuracy (Table 3). For example, Qwen3-8B gets 47.7% on ABCD but only 3.3% on the same questions without options. ABCD therefore overestimates how much Hungarian cultural knowledge small and mid-sized models can produce. The best open-ended score is 18.8%.

Second, the help is not equal for all models, so the ABCD ranking is not a ranking of knowledge. GPT-3.5 and Qwen3-4B are tied on ABCD, but differ by a factor of 4.6 open-ended. PULL-chat’s ABCD lead over GPT-3.5 becomes a tie. Possible reasons are that some models are better at eliminating wrong options, or that they follow the answer format more reliably. We cannot separate these, because we do not have item-level ABCD generations. The agreement between formats on categories ($p = 0.79$) shows that ABCD still shows which topics are hard. But for geography, where all models are at chance level, ABCD cannot tell the models apart.

Third, open-ended evaluation has its own format requirements: the model must stop, and the answer must be extractable. Racka-4B fails this for most items, and the Qwen3 models in thinking mode hit the token limit for 30 (4B) and 17 (8B) items. So neither format is free of format effects, and comparing base, adapted and aligned models needs control of decoding (chat template, stop sequences, token limit, answer extraction). This matters especially for lightweight adaptation models like Racka-4B. We recommend that cultural benchmarks (i) report both a closed-choice and an open-ended score, (ii) report how many generations are empty or cut off, (iii) check answer-format compliance in ABCD, and (iv) check automatic scoring on single items, because a scoring error can change the conclusions as much as a change of format.

Limitations. The gpt-4o judge, used for 58% of the open-ended items, is not yet validated against human ratings. Entity scoring is based on strings and can reject correct paraphrases. The lenient scores show the size of this effect (the lenient overall score is 1.9–2.8 times the strict score, but it also counts partial answers), but it is only a rough estimate. We used one prompt per answer type and one generation per item. ABCD results are only available as category totals, so we cannot yet compare the two formats item by item, or check ABCD answer-format compliance. The Racka-4B comparison depends on the first-line rule, and it is not fully controlled, because continual pretraining also changed the training data.

5 Contributions and next steps

This work (i) extends a Hungarian cultural benchmark with an open-ended version of all 2,641 items in 16 categories; (ii) gives a scoring method for open-ended cultural questions (alias

matching for entities, rubric-based LLM judging for the other types); (iii) shows that ABCD accuracy strongly overestimates open-ended knowledge, keeps the difficulty of categories, but not the ranking of models; and (iv) shows that a same-base comparison of a non-aligned adaptation model gives opposite results in the two formats, and that open-ended evaluation fails there mainly on stopping and answer extraction. Next, we plan to validate the judge with human ratings, compare the two formats item by item, check ABCD format compliance, and run more models.

References

- Almeida Carneiro, S., Xypolopoulos, C., Fei, X., Zhang, Y., & Vazirgiannis, M. (2026). CARTE: A Benchmark for Mapping Language Model Knowledge Across France. arXiv:2606.01995.
- Chiu, Y. Y., et al. (2024). CulturalBench: A Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. arXiv:2410.02677.
- Csibi, Z., Gortka, B. Gy., Nagy, K., Nemeskey, D. M., Sallai, M., Simonyi, A., Szekeres, A. M., & Palkó, G. (2026). Racka: Efficient Hungarian LLM Adaptation on Academic Infrastructure. *Proc. XXII. Hungarian Computational Linguistics Conference*. Also arXiv:2601.01244.
- Kabir, D., Mahim, M. R. C., Shafayat, S., Sadik, A., Ahmed, A., Kim, E., & Oh, A. (2025). BLUCK: A Benchmark Dataset for Bengali Linguistic Understanding and Cultural Knowledge. arXiv:2505.21092.
- Ligeti-Nagy, N., Madarász, G., Földesi, F., Lengyel, M., Osváth, M., Sárossy, B., Varga, K., et al. (2025). HuGME: A benchmark system for evaluating Hungarian generative LLMs. In *Proc. 4th Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 385–403. ACL.
- Liu, Z., Xie, T., & Zhang, X. (2024). Evaluating and Mitigating Social Bias for Large Language Models in Open-ended Settings. arXiv:2412.06134.
- Maji, A., Kumar, R., Ghosh, A., Anushka, & Saha, S. (2025). SANSKRITI: A Comprehensive Benchmark for Evaluating Language Models' Knowledge of Indian Culture. arXiv:2506.15355.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87(3), 252–271.
- Myrzakhan, A., Bsharat, S. M., & Shen, Z. (2024). Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. arXiv:2406.07545.
- Myung, J., et al. (2024). BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. arXiv:2406.09948.
- Robinson, J., & Wingate, D. (2023). Leveraging Large Language Models for Multiple Choice Question Answering. In *Proc. ICLR 2023*.
- Singh, S., et al. (2025). Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proc. ACL 2025, Vol. 1*, pp. 18761–18799.
- Tanwar, E., Chatterjee, A., Saxon, M., Albalak, A., Wang, W. Y., & Chakraborty, T. (2025). Do You Know About My Nation? Investigating Multilingual Language Models' Cultural Literacy Through Factual Knowledge. In *Proc. EMNLP 2025*, pp. 14956–14979.
- Yoo, H., Cho, W. I., Kim, G., & Han, J. (2026). From National Curricula to Cultural Awareness: Constructing Open-Ended Culture-Specific Question Answering Dataset. arXiv:2601.04632.

Adam Majczyk
Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, {adam.majczyk, maciej.ogrodniczuk}@ipipan.waw.pl

Retrieval-Augmented Generation System for the Polish PressMint Corpus

Keywords: Polish historical newspapers corpus, LLM, retrieval augmented generation

I Introduction

This paper presents a local, citation-backed Retrieval-Augmented Generation (RAG) system tailored to the Polish historical press materials in the PressMint project – spanning nearly 110 years (1830–1939) and comprising over 1.5 million words.

Historical newspapers represent a rich, chronologically anchored record of societal evolution, making them ideal resources for diachronic research, computational history and distant reading methodologies. Despite these advantages, historical press collections present many technical bottlenecks for natural language processing (NLP) and information retrieval (IR): heterogeneous content, OCR noise, archaic orthography and non-standard layouts.

The European PressMint initiative (Erjavec et al., 2026) addresses these challenges by compiling a multilingual, comparable, annotated and interoperable collection of historical newspaper corpora across 17 countries. The Polish contribution to PressMint (Ogrodniczuk et al., 2026) encompasses three complementary subcorpora spanning nearly 110 years (1830–1939): two 19th-century datasets from the Microcorpus of Nineteenth-Century Polish (rXIX; Bilińska et al., 2016, 2018) covering short press notes and journalistic articles, alongside the InterWar Corpus representing complete newspaper issues published in the restored Second Polish Republic (1918–1939).

While standardized Text Encoding Initiative (TEI) encodings and automatic linguistic annotations facilitate corpus linguistic queries, end-user interactive knowledge discovery remains constrained by traditional concordance or keyword-based search. Large Language Models (LLMs) offer intuitive, conversational natural-language interfaces. However, their deployment in scholarly contexts is severely limited by hallucination and the lack of verifiability. To bridge this gap, this paper adapts the local, citation-backed Retrieval-Augmented Generation (RAG) architecture, originally engineered for the Digital Library of Old Polish Ephemeral Prints (CBDU; Gruszczyński & Ogrodniczuk, 2014; Ogrodniczuk & Gruszczyński, 2011), to the Polish PressMint dataset.

2 RAG Architecture Adaptation

The RAG pipeline adapts the self-contained, open-source stack designed for CBDU (Majczyk & Ogrodniczuk, 2026). All components run locally as Docker containers and need no external services at query time. This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>. The system is deployed as a separate instance with its own vector index, so PressMint and CBDU documents are never mixed in retrieval. The AnythingLLM¹ as the chat interface and tool orchestrator, Qdrant² as the vector store, two textembeddings-inference³ instances for embedding and reranking and vLLM⁴ for serving the LLM. The system is deployed as a separate instance with its own vector index, so PressMint and CBDU documents are never mixed in retrieval. The AnythingLLM patches developed for CBDU are reused unchanged. They force a retrieval call on every user turn in agent mode, add GPU cross-encoder reranking to the Qdrant provider and display human-readable citation titles.

¹ <https://anythingllm.com>

² <https://qdrant.tech>

³ <https://github.com/huggingface/text-embeddings-inference>

⁴ <https://github.com/vllm-project/vllm>

2.1 From TEI to retrievable documents

The indexed data comprises the 4,337 TEI texts of the current Polish PressMint release: 400 FXIX samples (200 each of short press texts and journalistic texts) and 3,937 InterWar articles from complete issues of *Kurjer Wileński*. Each non-empty text is ingested as a single document, which is later split into chunks for retrieval. The 52 texts with an empty `<body>` are skipped, leaving 4,285 documents. Text is taken from the `<p>` elements of `<body>`. Conversion artefacts are normalised: heading markers (`##`) are removed, `{{...}}` omission marks are replaced by a bracketed ellipsis (`[...]`) and the braces around other `{{...}}` spans (foreign or emphasised words, partly illegible words) are dropped. Bibliographic metadata comes from `<sourceDesc>/<bibl>`: journal and article titles, publisher, place of publication, issue number, publication date, holding library and the URI of the digitised scan, complemented by the genre from `<catRef>` and the TEI identifier of the text (the root `xml:id`). The URI field may hold a bare URL, a URL followed by a descriptive title, two alternative URLs or no URL at all. We therefore extract the scan URL by pattern matching, preferring library scans over Wikisource transcriptions. Where the journal title is missing, we recover it from the remaining descriptive text, which reduces the number of documents without one from 782 to 427.

2.2 Using the linguistic annotation

The linguistically annotated version of each text (.ana.xml) is used to enrich its metadata. From the named-entity layer we collect persons, locations and organisations, discarding the miscellaneous class (mostly dates and agency abbreviations) and entities without a nominal token, such as derived adjectives tagged as locations, initials and abbreviations. Entities are normalised with the lemma layer: single-word names are replaced by their lemmas (genitive *Mickiewicza* → *Mickiewicz*; *WASZYNGTON* → *Waszyngton* ‘Washington’) and multi-word names are represented by a nominative occurrence where available (*Liga Narodów* ‘League of Nations’; about a third of multi-word names, otherwise the first form found), since concatenated lemmas do not form valid names. Within each document, the normalised entities are ranked by their number of mentions and the most frequently mentioned ones are added to the document’s label, capped per type at the corpus’s 90th percentile of distinct entities per document (20 persons, 15 locations and 8 organisations), which leaves each list complete for about 90% of documents and all three lists complete for 81.5%. The entities thus inform both retrieval and the model’s view of what an article covers. Of the 4,285 non-empty documents, 4,226 (99%) receive at least one entity. The system prompt marks these fields as automatic annotations to be used as hints and verified against the article text.

2.3 Metadata-aware chunking

As in CBDU, metadata reaches the LLM in two ways, but the PressMint TEI headers and annotation make it considerably richer: in CBDU the inline label, prepended to every OCR page, carried only the dating and, for 155 of the 315 records, a scholarly commentary, whereas here it also includes the journal, issue, publisher, place of publication, genre, article title, named entities and the TEI identifier. First, a bracketed label, e.g. [*Czasopismo: Kurjer Wileński, nr 14 | Data: 19 stycznia 1932 r. | Wydawca: ... | Miejsce wydania: ... | Gatunek: ... | Tytu: ... | Osoby: ... | Miejsca: ... | Organizacje: ... | ID: ...*] (‘Journal: Kurjer Wileński, no. 14 | Date: 19 January 1932 | Publisher: ... | Place of publication: ... | Genre: ... | Title: ... | Persons: ... | Locations: ... | Organisations: ... | ID: ...’), is prepended to the first paragraph of each document. It is attached to the paragraph rather than placed on a separate line, so the text splitter cannot cut it off as a near-empty chunk. Dates are given in the source’s own wording – when only the year is known, the label states this explicitly, so the model does not invent a day. Second, each document’s title (date, journal where known, article title) is repeated by AnythingLLM in the header of every chunk, so date and journal reach chunks of longer articles that the label never reaches. Documents are split into chunks of 1,500 characters with a 20-

character overlap and embedded with BAAI/bge-m3 (1,024-dimensional dense vectors), giving an index of 10,592 chunks.

2.4 Retrieval, generation and citations

Retrieval parameters are taken over unchanged from the CBDU deployment, where the top-75 setting was validated in a retrieval evaluation. For each query, the 250 nearest chunks by cosine similarity are filtered by a similarity threshold, reranked with the BAAI/bge-reranker-v2-m3 cross-encoder and the top 75 enter the context. Generation uses Qwen/Qwen3.5-9B with a 131k-token context window and thinking mode disabled, served by vLLM. This is the main departure from the CBDU deployment, which used the 35B-parameter mixture-of-experts model Qwen/Qwen3.6-35B-A3B-FP8. The smaller model was chosen because of the GPU memory available on the shared machine hosting the system and the larger one can be substituted without re-indexing. The model runs in agent mode, which allows several retrieval calls per question and is prompted to issue multiple reformulated searches for broad or comparative questions. The workspace is configured with a similarity threshold of $\theta = 0.4$ and a temperature of 0.05, but in AnythingLLM these apply only to its non-agent chat and query modes, which perform a single retrieval per message: in agent mode the retrieval tool uses the default threshold of 0.25 and generation requests carry no temperature, so vLLM's default of 1.0 applies. The system prompt, adapted from the (also Polish) CBDU version, restricts answers to the retrieved context. Its corpus-specific facts (document count, date range, composition) are computed at indexing time. It also asks the model to distinguish the publication date of an issue from the date of the reported event and to treat OCR errors and historical spelling (e.g. *Kurjer* for modern *Kurier* 'courier', *nietylko* for *nie tylko* 'not only') as properties of the source rather than correct them. Each document's label carries its TEI identifier (e.g. `PressMint-PL_1932-01-19-19-stycznia-1932-r-KW-4`), which the model is instructed to cite verbatim alongside the journal, date and article title. Unlike the scan URL, which is shared by all articles of the same issue, the identifier points to exactly one text in the corpus. Each retrieved chunk is also linked to the persistent URL of its digitised scan (mostly Polona and the Jagiellonian Digital Library, plus regional and university digital libraries; 177 documents have none), shown as a clickable citation next to the answer, so statements can be checked against the corpus text and, where available, the original page.

3 Conclusion

This paper has presented the adaptation of a local, citation-backed RAG system for the Polish subcorpora of the European PressMint project. By combining standardized TEI encodings, linguistic annotations, dense retrieval with cross-encoder reranking and an agentic local LLM architecture, the system enables interactive natural-language querying of the Polish PressMint dataset, providing a pilot implementation of a future solution which could integrate more data and languages.

Future extensions will incorporate multi-modal processing for historical press illustrations and expand the RAG index across all 15 partner languages in the PressMint ecosystem. A systematic evaluation of retrieval and answer quality, as carried out for CBDU, is also left for future work.

Acknowledgments

The submission was supported by (1) the PressMint CLARIN Flagship Project, (2) part of the investment: CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024–2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/VVK/01.

References

- Bilińska, J., Derwojedowa, M., Kwiecień, M., & Kieraś, W. (2016). Mikrokorpus polszczyzny 1830-1918 [EN: Microcorpus of Nineteenth-Century Polish]. *Komunikacja Specjalistyczna/Communication for Special Purposes*, (11), 149–161.
- Bilińska, J., Kwiecień, M., & Derwojedowa, M. (2018). Microcorpus of Nineteenth-Century Polish. In E. Fuß, M. Konopka, B. Trawiński, & U. H. Waßner (Eds.), *Grammar and corpora 2016* (pp. 377– 387). Heidelberg University Publishing. <https://doi.org/10.17885/heiup.361.c4712>
- Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., & Rigau, G. (2026). PressMint: Towards interoperable corpora of historical newspapers. *Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026)*, 1–5. <https://doi.org/10.63317/55w7588oukim>
- Gruszczyński, W., & Ogrodniczuk, M. (2014). Digital Library 2.0: Source of Knowledge and Research Collaboration Platform. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1649–1653). European Language Resources Association. <https://aclanthology.org/L14-1164/>
- Majczyk, A., & Ogrodniczuk, M. (2026). Evaluating the Use of Generative LLMs for Intralingual Diachronic Translation of Middle-Polish Texts into Contemporary Polish. *Proceedings of the 28th International Conference on Asia-Pacific Digital Libraries (ICADL 2026)*.
- Ogrodniczuk, M., Czerski, D., & Pawłowski, A. (2026). The Polish PressMint corpus. *Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026)*, 6–10. <https://doi.org/10.63317/3rqajugehtb3>
- Ogrodniczuk, M., & Gruszczyński, W. (2011). Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage. *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 27–33. <http://www.aclweb.org/anthology/W11-4105>

Alexey Matyushin

Institute of Languages for Professional Purposes, Sechenov University, Moscow, Russia,
matyushin_a_a@staff.sechenov.ru

Daria Ryzhova

School of Linguistics, HSE University, Moscow, Russia, dryzhova@hse.ru

Anastasiya Klenova

School of Linguistics, HSE University, Moscow, Russia, aaklenova@hse.ru

Adelina Makarova

School of Linguistics, HSE University, Moscow, Russia, aimakarova_3@hse.ru

Lada Starostina

Department of Pediatrics, Sechenov University, Moscow, Russia,
starostina_l_s@staff.sechenov.ru

LLM-Assisted Serbian–Russian Lexicography: The Case of Medical Terminology

Keywords: large language models, lexicography, medical terminology, Serbian-Russian dictionary, dictionary development workflow

I Introduction

The automation of lexicographic work has long relied on corpus-based evidence, automatic extraction, and post-editing workflows, in which the lexicographer evaluates and revises machine-proposed material rather than composing each entry from scratch. Recent advances in large language models (LLMs) have reinforced this approach, since generative systems can now produce draft definitions, equivalents, collocations, and illustrative examples in immediately readable form, while still requiring systematic expert validation. Lew (2023), for instance, found that GPT-4 could generate COBUILD-style definitions of acceptable quality, yet struggled to produce reliable illustrative examples. Recent studies have therefore shifted from asking whether LLMs can generate dictionary-like text to asking which parts of such output are lexicographically reliable, verifiable, and usable (de Schryver, 2023; Jakubiček & Rundell, 2023). A Serbian case study presented at eLex 2025 frames LLM use as promising but not autonomous, noting persistent issues with sense disambiguation, usage examples, and definitions (Stanković et al., 2025).

For bilingual lexicography, the problem is even more complex. Besides describing meanings, dictionary entries must establish appropriate translation equivalents, distinguish near-equivalents, indicate usage constraints, and provide examples that are natural in both source and target languages. This is particularly important for Serbian–Russian lexicography, where the close relationship between the languages may obscure differences in meaning, style, collocability, and domain-specific usage. Recent work on the new Serbian–Russian dictionary has emphasized the need to revise both the word list and the entry structure for bilingual purposes, drawing on corpora of modern Serbian to verify frequency and usage contexts, and explicitly addressing lexical equivalence and polysemy (Dragičević et al., 2024).

Medical terminology provides a productive test case for LLM-assisted Serbian–Russian lexicography because it lies at the boundary between specialized language and general vocabulary. Some medical terms often circulate outside strictly professional discourse, appearing in patient communication, public health information, media texts, and everyday speech. At the same time, medical vocabulary is lexicographically demanding: it includes internationalisms, Greco-Latin forms, adapted borrowings, competing native and foreign terms, abbreviations, eponyms, synonymy, and differences between professional and lay usage. These characteristics make medical terminology a particularly interesting domain for evaluating how effectively LLM-generated drafts can support Serbian–Russian lexicographic work, particularly whether the proposed entries and examples are accurate, appropriately scoped for a general contemporary dictionary, and suitable for expert post-editing.

2 Materials and Methods

The study was designed as a two-stage evaluation of LLMs for drafting entries for a contemporary Serbian–Russian dictionary (Dragičević et al., 2024). The design of Stage 1 draws on Matyushin (2025): 24 models (20 local in Q4 quantization and 4 cloud full-precision; 4–671B parameters) were prompted in Serbian Cyrillic to generate ten dictionary entries for a single Serbian medical term following a predefined structure: headword in Cyrillic, part-of-speech (POS) label, and Russian equivalent. A separate prompt then asked each model to generate an example of the headword's use in a medical context. Outputs were scored using a custom scale: lexicographic components (entry correctness, POS adequacy, and translation accuracy) were each rated 0–1, while examples were assessed separately for Serbian (grammar 0–2, orthography 0–1, relevance 0–1) and Russian (grammar 0–2, orthography 0–1, translation accuracy 0–1, relevance 0–1, naturalness 0–5). This component-based scoring distinguished models that followed the requested structure from those that produced formally correct but lexicographically unreliable output.

The highest-scoring models advanced to Stage 2, where each was provided with 30 medicine-related terms: 15 monosemous and 15 polysemous frequent items from the Russian National Corpus, excluding body-part names. Each model was prompted to draft usage examples in Serbian and a corresponding Russian translation. Generated Serbian examples were evaluated by one of the authors with C1–C2 competence in Serbian in terms of grammar, orthography, and relevance. The Russian parts of the output were evaluated for grammar, orthography, translation accuracy, relevance, and naturalness by another author (a linguist; Evaluator 1) alongside the third author (a medical professional; Evaluator 2). Statistical analysis was performed using R version 4.4.2 (R Core Team, 2024).

One of the methodological choices concerned the Serbian script. Although Serbian is commonly written in both Cyrillic and Latin scripts, we focused on Cyrillic as a deliberately more challenging condition. Since Cyrillic is considerably less widespread internationally than the Latin script, we assumed that Cyrillic-script Serbian would be less extensively represented in multilingual LLM training data and might therefore present greater difficulties for language identification and generation. Accordingly, we expected some script-related errors to be less pronounced if the same task were performed in Latin-script Serbian. However, script was not treated as an experimental variable in the present study, and this expectation therefore remains to be tested directly.

3 Results and Discussion

Of the 24 models tested in Stage 1, a clear performance gradient emerged. The lowest-scoring models—Llama 3.1 (mean 0.50), Llamax 3 (0.50), and Phi-3 Medium (1.38)—produced structurally incomplete or incoherent output. For the remaining models, lexicographic components were generally correct when present (mean 0.71–0.83 on a 0–1 scale across four evaluators), suggesting that LLMs demonstrate sufficient structural compliance with the task. The critical weakness lay in illustrative examples, and specifically in Serbian. The mean score in this task was 1.50/4.0, with a median of 0.00, indicating that approximately half of the models produced no valid Serbian examples at all.

The qualitative notes of the other evaluators independently corroborated this pattern: they noted Russian-to-Russian paraphrases, missing Serbian output, language contamination (Ukrainian, English, German, and Chinese words or phrases), template-like repetitive phrasing, and transliteration in place of translation. The occurrence of Ukrainian is particularly relevant to the choice of script: some models appeared to rely partly on orthographic cues when identifying the input language, producing another Cyrillic-script Slavic language rather than Serbian. By contrast, contamination by Latin-script Slavic languages such as Czech or Polish was not observed. This pattern suggests that the use of Cyrillic may indeed make language identification more challenging for some models, although a direct Cyrillic–Latin comparison would be required to establish a script effect.

Russian example quality was considerably more robust and was evaluated by all four raters. The mean Russian score was 6.29/10 (median 7.50, N = 96 ratings). Nine models scored consistently above 8.0 and advanced to Stage 2 (Figure 1A). Notably, model size did not significantly predict performance (Pearson $r = 0.352$, $p = 0.092$, $N = 24$), with the 27B TranslateGemma-27b nearly matching the 671B DeepSeek-V3.1. Inter-rater agreement on Russian example quality was good (mean Pearson $r = 0.777$ across all six evaluator pairs, range 0.678–0.901), with the strongest convergence on grammar and orthography and the greatest divergence on naturalness judgments.

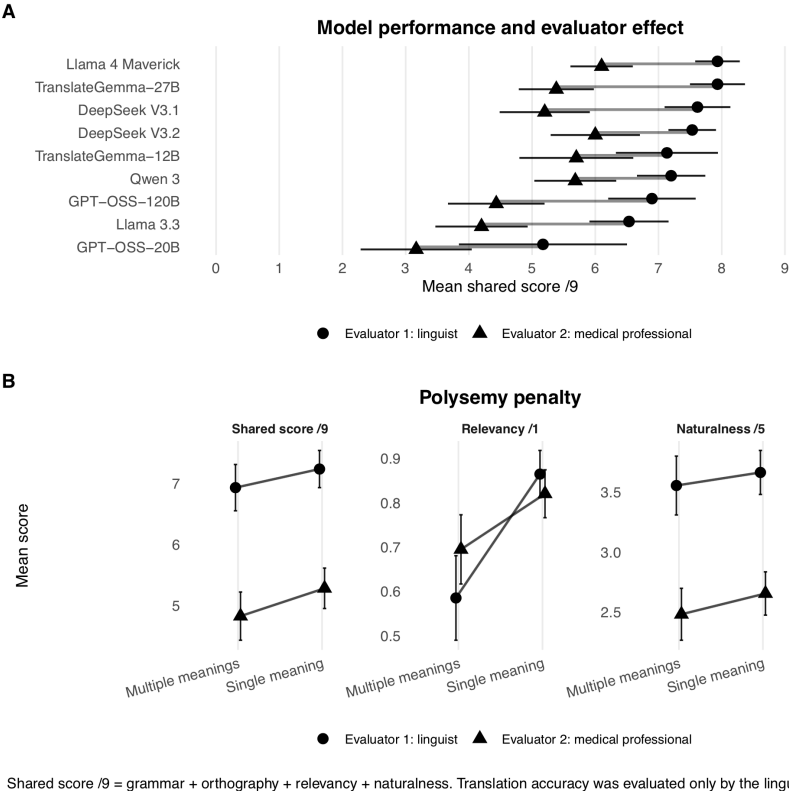


Figure 1: (A) Model performance on the shared 9-point scale; (B) effect of polysemy on mean score.

Stage 2 confirmed that the best-performing LLMs can generate Serbian–Russian example sentences that are often fluent and structurally usable as draft material, but the statistical results also show that this usability is uneven and requires expert control. On the shared 9-point scale (Figure 1A), translation accuracy excluded, the mean scores assigned by Evaluator 1 were substantially higher than those of Evaluator 2, suggesting that domain-specific assessment identifies weaknesses that are less apparent in a general linguistic or lexicographic evaluation. Agreement between Evaluator 1 and Evaluator 2 was moderate overall, with the strongest relationship observed for the shared composite score ($r = 0.67$) and naturalness ($r = 0.58$).

Consistent with the Stage 1 results, the models performed relatively well on more formal criteria such as grammar and orthography, but less reliably on relevance, naturalness, and translation accuracy. The mean grammar scores assigned by Evaluator 1 and Evaluator 2 were 1.80/2 and 0.89/2, respectively; orthography scores were high for both evaluators, with means of 0.94/1 and 0.84/1. Relevance was closer between evaluators, with means of 0.75/1 for Evaluator 1 and 0.77/1 for Evaluator 2, but naturalness showed a larger difference (3.62/5 versus 2.59/5).

The effect of polysemy further clarifies the models' main weakness (Figure 1B). Monosemous terms received higher scores than polysemous terms from both evaluators. On the shared 9-point scale, Evaluator 1 scores fell from 7.23/9 for single-meaning terms to 6.93/9 for polysemous terms, while Evaluator 2 scores fell from 5.28/9 to 4.82/9. The clearest drop was in relevance: Evaluator 1's relevance scores declined from 0.86/1 for monosemous terms to 0.58/1 for polysemous terms, while Evaluator 2's fell from 0.82/1 to 0.69/1. This indicates that models often generate fluent sentences but fail to convey the intended medical sense when the Serbian headword also has a broader or non-medical meaning.

Finally, frequency did not meaningfully predict example quality. Correlations between log corpus frequency and shared score were negligible ($r \approx 0.06$ for Evaluator 1's shared score and $r \approx 0.01$ for Evaluator 2's), suggesting that the models' success in the task does not depend straightforwardly on how frequently a headword might appear in the training data.

The Serbian-side evaluation added an important corrective to the Stage 2 results. Llama 4 Maverick and TranslateGemma-27b were found to be the most reliable models for Serbian example drafting, with mean scores of 3.92/4 and 3.84/4, respectively, and no zero-score outputs. Criterion-level results showed that the main weakness of the evaluated models in Serbian examples was relevance (mean 0.67/1). This weakness was especially notable for polysemous terms: single-meaning terms scored 3.22/4, compared with 2.91/4 for multiple-meaning terms, and the relevance score dropped from 0.78/1 to 0.50/1.

4 Conclusion

Overall, this study shows that current-generation LLMs can usefully assist in drafting bilingual Serbian–Russian dictionary examples for medical headwords. Their most appropriate role is as generators of candidate material within an expert-supervised lexicographic workflow. The models' main strength lies in producing fluent, formally acceptable examples, especially for monosemous and semantically straightforward headwords. On the other hand, sense selection, plausibility, relevance, and translation accuracy, especially for polysemous or culturally specific medical terms, remain substantially weaker. These shortcomings partly reflect the fact that Serbian and, to a lesser extent, Russian remain under-resourced languages in current LLM training data, which limits the models' ability to handle nuanced lexicographic tasks in these languages reliably. Despite these limitations, the study sheds light on both the strengths and weaknesses of LLM-assisted example generation for this language pair and may offer practical guidance for both general and specialized lexicography.

References

- de Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. <https://doi.org/10.1093/ijl/ecad021>
- Dragičević, R., Korolkova, M. D., Ryzhova, D. A., Shapich, Y. L., & Yakushkina, E. I. (2024). A new large serbian-russian dictionary: A general conception and problems of lexicographic description. *Voprosy Leksikografii – Russian Journal of Lexicography*, 32, 43–60. <https://doi.org/10.17223/22274200/32/3>
- Jakubiček, M., & Rundell, M. (2023). The end of lexicography? can ChatGPT outperform current tools for post-editing lexicography? In I. Kosem, M. Jakubiček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible lexicography. proceedings of the eLex 2023 conference* (pp. 518–533).
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), 704.
- Matyushin, A. (2025). LLMs as tools for drafting ad hoc pharmaceutical glossaries [November 20–21, 2025]. *Proceedings of the First International Workshop on Language and Language Models (WoLaLa 2025)*.
- R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Stanković, R., Stijović, R., Škorić, M., & Krstev, C. (2025). The dictionary of contemporary serbian language (RSSJ): Advanced automation and other challenges. In I. Kosem, M. Jakubiček, M. Medved', K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. proceedings of the eLex 2025 conference* (pp. 59–75).

Anna Mikó

Department of Cognitive Science, BME, Budapest, Hungary, miko.anna@edu.bme.hu

Anna Babarczy

ELTE Research Centre for Linguistics, Budapest, Hungary, babarczy.anna@nytud.hu

Epistemic vigilance and the effect of LLM-based assistance on stress during cognitively demanding tasks: Dissociating state anxiety from perceived workload

Keywords: cognitive offloading, AI chatbots, LLM hallucination, mental workload, state anxiety, AI reliability

I Introduction

The integration of large language model (LLM)-based assistants into cognitively demanding educational and professional environments has fundamentally altered the nature of human-machine interaction. Unlike passive information retrieval systems, LLMs actively participate in problem formulation and solution generation, functioning as what Kosmyna et al. (2024) term “intelligent cognitive partners.” EEG evidence from the same research group demonstrates that prefrontal cortex activity-associated with executive function and complex decision-making is significantly reduced during LLM-assisted task performance compared to unassisted conditions. This reduction may provide short-term relief from mental effort; however, it also raises substantive concerns about the accumulation of “cognitive debt” and the gradual erosion of autonomous critical thinking (Kosmyna et al., 2024). Throughout this paper, we use stress as an umbrella term encompassing both the affective component of state anxiety (STAI-S) and the perceived-workload component (NASA-TLX, particularly the Frustration and Performance subscales); where the two dissociate, we report them separately.

The theoretical case for stress reduction through LLM assistance rests on cognitive offloading theory, which predicts that delegating mentally demanding sub-tasks to external tools (e.g. calculator, AI) frees internal working memory resources and reduces load-related anxiety (Eysenck et al., 2007; Risko & Gilbert, 2016). However, LLMs introduce a qualitatively distinct and largely unexamined source of uncertainty: hallucination as the confident and fluent generation of factually incorrect or internally inconsistent content (Ji et al., 2023). Because hallucinated outputs are linguistically indistinguishable from accurate ones, users cannot rely on surface-level plausibility cues to detect errors, placing an additional monitoring demand on the user that runs parallel to the primary cognitive task.

The degree to which users manage this effectively depends critically on trust calibration, the alignment between their subjective confidence in the LLM and the actual reliability of the system (Lee & See, 2004). Miscalibration in either direction carries costs: overtrust leads to uncritical acceptance of erroneous outputs, while undertrust generates hypervigilant monitoring that exhausts attentional resources (Hoff & Bashir, 2015). A particularly under-researched question concerns what happens psychologically when users encounter LLM errors unexpectedly when initial overtrust is abruptly disrupted. Theory predicts an acute cognitive re-evaluation response that elevates load and subjective stress (Hoff & Bashir, 2015), yet direct experimental evidence for this effect during active task performance under time pressure remains scarce.

2 Methods and Data

A three-group, between-subjects online experiment was conducted in March 2026 via the Gorilla Experiment Builder, a platform specifically designed for behavioral research that supports automated randomization, strict response-time enforcement, and millisecond-accurate event logging. Participants (N = 123; 66.7% female; average age 27) were recruited from university and professional networks using snowball sampling and randomly assigned to one of three conditions: (1) a no-AI control group (n = 42) without AI assistance; (2) a reliable-AI group (n = 36) receiving a consistently accurate LLM-generated suggestion alongside each task; or (3) an

fallible-AI group ($n = 45$) receiving hallucinated - deliberately incorrect suggestions in 60% of trials. Participants in both AI conditions were not informed about the nature or accuracy of the suggestions. ChatGPT was selected as the basis for both task construction and AI suggestion generation, reflecting its dominant usage among the Hungarian student population ((GKID Research & Consulting & UNIside, 2024). Full debriefing was provided upon completion.

State anxiety was assessed before and midway through task completion using the state subscale of the State-Trait Anxiety Inventory (STAI-S; Spielberger, 1970), capturing transient anxiety as a function of situational stressors. Perceived workload was measured post-task via the NASA Task Load Index (NASA-TLX; (Hart & Staveland, 1988)), with particular focus on the Performance and Frustration subscales. Participants in both AI conditions additionally completed five custom Likert-scale items assessing perceived cognitive offloading, awareness of potential AI errors and reliance on AI under uncertainty. All analyses were conducted in R Studio running R 4.6.0. Logistic mixed effects models were fitted to analyse group differences in accuracy at the cognitive tasks, and to test differences in participants' willingness to accept AI suggestions.

3 Procedure

Participants started with 3 three unscored practice items with no time limit then completed 10 cognitively demanding tasks within 45-second response windows. The tasks were designed to engage working memory and insight-based reasoning, drawing on established cognitive paradigms including the Wason selection task, matchstick arithmetic equations and classical insight problems (e.g., the surgeon problem). Each item offered four response options: one clearly incorrect distractor, two plausible but logically flawed alternatives, and the correct answer. Task content and response options were generated using a few-shot prompting method with ChatGPT to ensure coherent formatting and realistic presentation of AI suggestion. This asymmetric design was chosen to model ecologically plausible hallucination rates while maintaining sufficient within-condition variability to observe behavioral responses to both correct and incorrect suggestions. AI suggestions were presented as inline text alongside each task item, mimicking the appearance of a real-time AI assistant. The experimental session comprised nine sequential sections: informed consent, pre-task STAI-S, practice tasks, cognitive tasks 1–5, mid-task STAI-S, cognitive tasks 6–10, AI experience items (AI conditions only), NASA-TLX, and debriefing.

4 Results

The effects of AI assistance accuracy on task performance were analysed using a logistic mixed-effects model. Relative to the no-AI group, participants in the reliable-AI group showed significantly higher odds of a correct response ($\beta = 1.33$, $SE = 0.25$, $z = 5.28$, $p < .001$, $OR = 3.78$), while the fallible-AI group did not differ significantly from controls ($\beta = 0.39$, $SE = 0.23$, $z = 1.72$, $p = .086$, $OR = 1.48$; Figure 1), although the effect size suggests their odds of a correct response may also have been higher.

To determine whether fallible-AI participants actively filtered suggestions rather than overrelying on them, compliance rates were compared between hint types. The model revealed that participants were substantially more likely to comply with correct hints than hallucinated ones ($\beta = 2.74$, $SE = 0.71$, $z = 3.88$, $p < .001$, $OR = 15.43$), with estimated compliance probabilities of .75 and .16, respectively, that match the compliance rate in the reliable-AI group (.75). This pattern indicates active epistemic vigilance and critical filtering rather than automation bias.

State anxiety (STAI-S) was analysed using a linear mixed-effects model. There was no significant overall change from baseline to mid-task ($\beta = 0.62$, $SE = 0.50$, $t(124) = 1.25$, $p = .215$), and the timepoint $\times$ group interaction was also non-significant ($F(2, 124) = 0.92$, $p = .402$). Descriptively, the no-AI group showed the largest anxiety increase (+1.57 points), while AI-assisted groups showed minimal change (+0.19 and +0.11), but these differences were not statistically reliable.

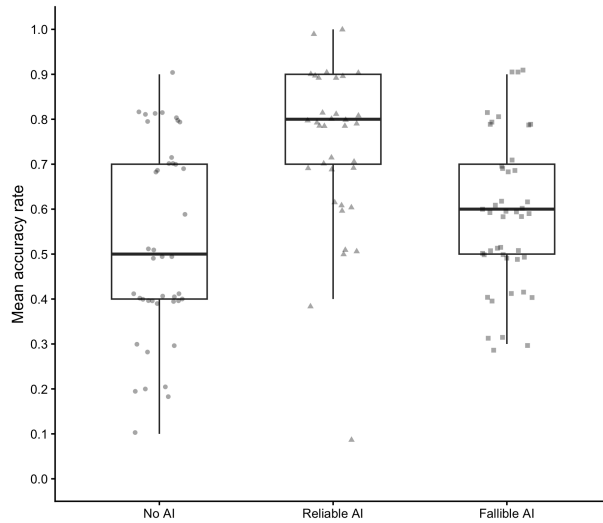


Figure 1: Mean individual accuracy rates on the problem-solving task organised by group.

Of the NASA-TLX subscales, only perceived performance showed a significant group effect ($F(2, 121) = 3.38, p = .037$). Tukey-adjusted comparisons revealed that the reliable-AI group rated their performance significantly better than the no-AI group ($MD = 13.69, SE = 5.54, p = .039$), consistent with their objectively higher accuracy. The fallible-AI group showed a descriptively similar but non-significant improvement ($p = .138$), and the two AI groups did not differ from each other ($p = .788$). Mid-task anxiety correlated significantly with perceived performance ($r(122) = .250, p = .005$), an association driven primarily by the fallible-AI group (Figure 2).

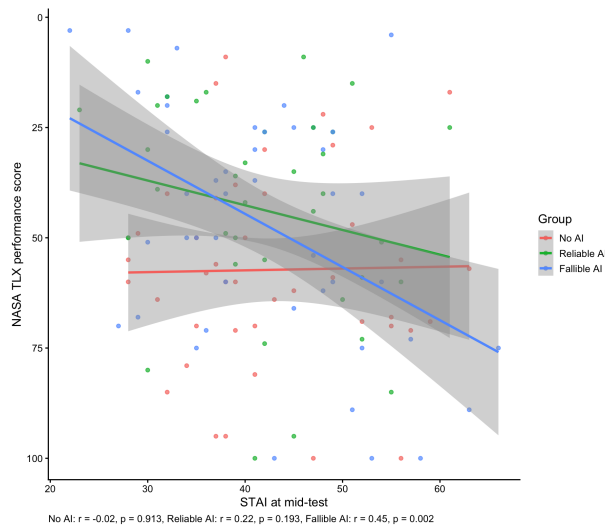


Figure 2: Correlation between state anxiety at mid-test and the NASA-TLX performance subscale by group. Note that lower performance scores indicate better perceived performance; the y-axis is reversed for ease of interpretation.

Mann-Whitney tests on subjective AI experience items confirmed that reliable-AI participants perceived significantly greater cognitive offloading ($U = 256, p = .030$), while fallible-AI participants

more strongly endorsed awareness that the AI might be wrong ($U = 543$, $p = .027$). No statistically significant differences were observed for the remaining three items (all others $p \geq .3$).

Unreliable LLM assistance neither buffers nor amplifies subjective stress during cognitively demanding performance under time pressure - yet it subtly inflates perceived performance competence. These findings reveal a critical dissociation: accurate LLM assistance improved task accuracy in a manner consistent with intelligent offloading (Wu & Yu, 2023), but failed to translate into measurable reductions in anxiety or overall workload. The persistent monitoring demand inherent to LLM interaction - even in the accurate condition, where users reported perceiving offloading yet not experiencing relief - sustains cognitive engagement at a level that offsets the expected stress-reducing benefits of delegation, consistent with the vigilance cost of automation.

The absence of automation bias likely reflects the STEM-educated, AI-familiar profile of the current sample. As habitual LLM reliance increases across broader populations with lower AI literacy, epistemic vigilance may weaken and compliance with hallucinated outputs may rise - with significant implications for both performance and wellbeing ((Hoff & Bashir, 2015; Hornberger et al., 2023). These findings carry direct practical implications: interface design that foregrounds AI reliability-through uncertainty signaling or confidence calibration - may be necessary to realize the stress-reducing potential of LLM assistance.

References

- Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, 7(2), 336–353. <https://doi.org/10.1037/1528-3542.7.2.336>
- GKID Research & Consulting & UNIsid. (2024). *Digital student survey: The impact of generative AI on higher education in hungary* (Research report). GKID.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). North-Holland.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence*, 5, 100165. <https://doi.org/10.1016/j.caeai.2023.100165>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Kosmyna, N., Liu, C., Hauptmann, R., Barber, S., & Maes, P. (2024). Is your brain on AI? neuroscientific study of LLM use in human–computer interaction [arXiv preprint, <https://arxiv.org/abs/2506.08872>].
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Wu, R., & Yu, Z. (2023). Do AI chatbots improve students' learning outcomes? evidence from a metaanalysis. *British Journal of Educational Technology*, 55(1), 10–33. <https://doi.org/10.1111/bjet.13334>

Alexandr Parahonco

USM, Chisinau, Moldova, alexandr.parahonco@math.usm.md

Victoria Bobicev

UTM, Chisinau, Moldova, victoria.bobicev@ia.utm.md

Maria Carp

RACAI, Romanian Academy, Bucharest, Romania, maria@racai.ro

Verginica Barbu-Mititelu

RACAI, Romanian Academy, Bucharest, Romania, vergi@racai.ro

From Metadata to Synthetic Legal Text: A Two-Stratum Romanian Generation Pilot for ADAMo

Keywords: natural language processing, text corpus, text generation, LLM

I Motivation and research question

This work investigates how Romanian language varieties affect the generation and detection of AI-written text. ADAMo aims to enrich CoRoLa with Romanian texts from the Republic of Moldova and build a synthetic corpus for studies of language variation and AI-text detection (Barbu Mititelu et al., 2025). Its prompts are to follow CoRoLa metadata so that generated and human-authored material can be compared by topic and genre. MOROCO previously provided a news-domain corpus from the two countries (Butnaru & Ionescu, 2019). Synthetic-data research emphasises that generation must be followed by curation and evaluation (Long et al., 2024).

We ask *which parts of a constrained generation task are controlled by metadata-informed instructions, and which require independent validation?* We examine a broad legal collection and a smaller paired pilot of explanatory constitutional-law articles. The latter shares a thematic frame across countries but requires distinct institutions, statutes, and case law. Neither collection is used here to train a detector or measure language-variety differences.

2 Prompting Methodology and the Rationale for Strict Typing

2.1 Methodological context

The prompting guide, developed inside the team, treats each generated text as the output of a documented corpus cell rather than as an unconstrained request for prose. This design is relevant to CoRoLa 2.0, where corpus enrichment is accompanied by revised metadata and an additional syntactic annotation layer; the published overview also describes safeguards intended to keep AI-generated material distinct from web-collected corpus texts (Irimia et al., 2026). ADAMo, in turn, studies AI-generated texts as a separate condition for detection across Romanian varieties (Barbu Mititelu et al., 2025). The prompt therefore needs to preserve distinctions among text type, genre, domain, source, language variety, and generation condition. These are not interchangeable labels: each supports a different corpus query and a different comparison.

We identified all 121 files in the broader collection as Moldova-context texts: 107 title-only generations and 14 paraphrases; 21 concern constitutional law and 100 administrative law. The accompanying XML files are metadata records. Directory and XML labels identify several model families, including ChatGPT, Gemini, Gemma, GPT-OSS, MiniMax, Nemotron, Copilot, and Manus. Tasks, genre labels, and lengths differ, and a complete matched prompt set is unavailable. We use these files to document corpus-building scope, not model-performance differences.

The paired pilot links ten constitutional-law themes across the two contexts, including EU integration, referendums, elections, language policy, and rights. Each topic index denotes *thematically parallel*, not identical, tasks: EU accession concerns Moldova; EU membership and the status of EU law concern Romania. Country-specific anchors and terms prevent jurisdictional conflation but also mean that national variety is not the sole experimental variable.

The paired archives contain 20 UTF-8 encoded articles generated with `gemma4:31b-cloud`. The notebook specifies a temperature of 0.2, a top-p value of 0.2, and thinking mode enabled, and it sends the instruction as a user-role message.

2.2 Structure of the prompt

The guide assembles one prompt as a single message from ordered, reusable components. It describes the procedure as having six blocks, although the actual labels run from [0] to [6]; this section follows the seven numbered components in the guide.¹

[0] Role

A fixed instruction, copied identically across runs, defines the model as an author writing in Romanian as spoken in Romania or Moldova. It requests only the text, requires a title, disallows a preamble or commentary, and prohibits web search, source citation, and copying from accessible texts. Keeping this block invariant is intended to limit avoidable variation in the basic task framing.

[1] Task

The selected `DocumentType` is translated into a natural-language writing task (for example, a `Newspaper article` becomes a press article, while `inProceedings` becomes a conference paper). The model is thus told what kind of document to produce, not merely given a topic.

[2] Frame

Metadata provide the document type, domain and subdomain, source and source name, and year. The guide specifies fallback or omission rules for empty fields. It also distinguishes the year assigned to the generated text in this frame from historical years that belong to the subject matter and should instead appear in the content instructions.

[3] Form

Values from the data sheet specify approximate word count, paragraph count, sentence length, sentencelength variation, grammatical person, and dominant tense. These values are copied without rounding or discretionary adjustment, so that the requested form is an explicit input rather than an informal stylistic suggestion.

[4] Content

The topic is combined with three or four named-entity anchors and, where relevant, three or four specialist terms. The guide asks that anchors be integrated naturally rather than treated as a checklist; not every anchor must appear. Terminology is normally omitted for genres such as imaginative writing and memoirs, where specialist terms are not expected.

[5] Genre constraints

A complete, style-specific constraint block is inserted from the selected `DocumentTextStyle`. For example, the `Law` block requests conventional references to legal provisions (article, paragraph, subparagraph), complex sentences, no first person, and no didactic explanation. Other blocks encode the relevant conventions of journalistic, scientific, administrative, imaginative, memoir, and other texts; a blog post additionally receives a direct-address instruction.

[6] Closure

A final invariant instruction asks the model to produce only the text, starting with its first sentence. It closes the output contract and discourages headings such as “Here is the requested text” or other material that is not part of the generated document.

¹ The discrepancy is retained here rather than silently resolved: the guide says “six blocks” but enumerates [0]–[6].

The guide also defines run-level records: text identifier, model name as displayed, interface, date, prompt version, task and corpus-cell identifiers, all data-sheet values, regeneration count, human intervention, achieved length, and free-text observations. Recording these fields connects the reusable prompt template to the particular output actually produced.

2.3 Purpose of strict typing

Here, *strict typing* means that prompt-construction fields have explicit meanings, controlled value sets, and, where necessary, cross-field constraints. It does not mean that the model is required to return JSON: the generated document remains free-form natural language. The metadata scheme constrains such fields as `DocumentType`, `DocumentTextStyle`, `Source`, `DocumentTextDomain`, and the domain-dependent `DocumentTextSubDomain`. For example, a legal text is encoded as `Law/Society/Law`; an administrative text as `Administrative/Society/Administration`; and an imaginative text has no domain or subdomain. Such rules are intended to reduce semantically incompatible combinations caused by inconsistent wording.

This typing serves four methodological purposes. First, it improves *consistency*: repeated categories are represented with the same labels and capitalization, reducing category drift across prompt authors and batches. Second, it supports *comparability*: when the same ordered slots and fixed role instruction are used, intended differences in topic, genre, variety, or model can be specified explicitly instead of being entangled with ad hoc prompt wording. Third, it enables *validation and retrieval*: invalid or incomplete combinations can be detected before generation, and corpus records can later be filtered by document type, style, domain, source, or variety. Fourth, it strengthens *reproducibility*: a prompt version and its typed inputs can be linked to the model, interface, date, regeneration history, and any human edits.

Strict typing controls the specification, not the truth of the generated content or the model's compliance. In particular, the guide's prohibition on web search and source citation does not verify legal claims; it makes independent factual review more important. Nor does the existence of a common template prove that every archived output received exactly the same prompt. For this reason, the guide should be used together with per-run prompt storage, metadata validation, and a separate review of legal grounding and language quality. In the present pilot, the guide supplies the methodological specification, while the available XML sidecars do not preserve the full prompt and generation parameters for each output.

3 Preliminary observations

The additional collection has 121 distinct text files. Every article in the paired pilot has a distinct title and six body paragraphs, showing complete adherence to these two structural requirements. Eight of the 20 body texts—four in each national context—fall below the prompt's 650-word minimum. This result is stable across the two word-count methods used in the working audit, although exact word totals vary slightly with tokenisation. Thus, a highly regular paragraph structure does not guarantee compliance with a minimum length instruction.

The targeted legal check identifies a different problem. One Moldova-context article invokes Constitutional Court Judgment No. 11 of 23 September 2025 but presents it mostly as a general statement about defence rights and access to justice (Constitutional Court of Moldova, 2025). The ruling specifically concerned a legislative omission in the procedure for requests to hear co-accused persons from separate proceedings as defence witnesses. The Romania-context article invokes Constitutional Court Decision No. 670 of 25 November 2025 as if it substantively strengthened defence guarantees; the Court in fact dismissed the constitutional challenge as inadmissible (Constitutional Court of Romania, 2025). In both examples, the institution and decision number appear, but the legal question or procedural outcome is blurred. These two examples demonstrate a possible mismatch between *anchor inclusion* and *claim fidelity*. They do not show how often that mismatch occurs across all 20 texts.

The outputs are therefore useful as a pilot of prompt behaviour, but they are not yet validated legal articles. In particular, a fluent passage should not be treated as evidence that a case was interpreted correctly. The same distinction matters when comparing the two contexts: jurisdiction-specific entity names may be generated correctly even when the substance remains generic or legally incomplete. The broader collection adds a second caution: an administrative-law paraphrase, a legal exposition, and a journalistically framed article serve different tasks and should not be pooled into one undifferentiated model score.

4 Scaling and validation protocol

The verified inventory already exceeds 100 documents, but a larger *matched* experiment still requires new controls. One possible next stage is ten themes $\times$ two national contexts $\times$ five independent outputs, yielding 100 comparable articles under one specified model and prompt version. Further models can then be added as separate conditions with the same topics and evaluation rules. This is a proposed design, not a claim that the 121 additional files satisfy that balanced matrix. Repeated outputs would permit examination of variation within the same prompt as well as differences between contexts and models.

Before scaling, each task needs a short, versioned factual dossier drawn from authoritative records. It should state the relevant institution, legal instrument, date, procedural status, and precise holding where a case is mentioned. The prompt, model identifier and version, decoding parameters, generation date, tool calls, retries, and human edits should be preserved with each output.

Evaluation should then use separate dimensions for formal compliance, topical relevance, jurisdictional specificity, fidelity to the factual dossier, journalistic register, repetition, and Romanian language quality. Independent Romanian-proficient reviewers, with legal expertise for case-specific assertions, should annotate the texts blind to the generating model where feasible, record disagreement, and adjudicate disputed claims against primary sources. The resulting data would support descriptive comparisons and later downstream experiments while keeping synthetic, human-authored, and post-edited material distinct. Crucially, legal-domain synthetic texts alone cannot establish how people actually write in either country; representative human-authored sources remain necessary for linguistic comparison.

5 Conclusion

This work documents 141 unique legal-domain texts generated through two different sets of workflows and presents a focused audit of a 20-text, two-context journalistic subset. Its strongest current result is the separation between readily controlled *surface form* and *less reliable factual grounding*: all paired-pilot outputs meet the six-paragraph requirement, eight miss the minimum body length, and selected court-case passages lose legally consequential detail.

References

- Barbu Mititelu, V., Bobicev, V., Alexei, V., Braniște, R., Caftanator, O., Carp, M., et al. (2025). *Automatic Detection of AI-Generated Texts from Moldova and Romania (ADAMo): A Project Presentation*. ConslLR 2025 paper supplied by the authors.
- Butnaru, A., & Ionescu, R. T. (2019). MOROCO: The Moldavian and Romanian Dialectal Corpus. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 688–698.
- Constitutional Court of Romania. (2025). Decision No. 670 of 25 November 2025.
- Constitutional Court of the Republic of Moldova. (2025). Judgment No. 11 of 23 September 2025.
- Irimia, E., Barbu Mititelu, V., Ion, R., Păiș, V., Mitrofan, M., & Tufiș, D. I. (2026). CoRoLa Version 2.0: Corpus Enrichment and a New Annotation Level. *Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora*, 84–90.
- Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *Findings of the Association for Computational Linguistics: ACL 2024*, 11065–11082.

Nicolae Rusu

UTM, Chisinau, Moldova, nicolae.rusu@iis.utm.md

Victoria Bobicev

UTM, Chisinau, Moldova, victoria.bobicev@ia.utm.md

Daniela Istrati

UTM, Chisinau, Moldova, daniela.istrati@ia.utm.md

Evaluation of Semantic and Visual Coherence in AI-generated Journalistic Content

Keywords: generative models, generated text, generated images, generated content detection

I Introduction

This paper presents a comparative analysis of original news media content and content generated using large language models. Recent artificial intelligence models can automatically generate text and images, and this capability is transforming the way informational content is created, summarized, and illustrated. In the realm of online journalism, the use of such models raises a key issue: whether the generated content may retain an article's general theme, or it risks losing factual details, editorial context or visual elements that lend documentary value to a genuine news story. In this work, the general task consisted in comparing original press articles and photographs with automatically generated ones in order to observe the similarities, differences, and limitations of the used models. The study included the semantic analysis of texts, the analysis of text-image relationships, and the automatic classification of human written versus AI generated texts. The work primarily addresses the first two areas: semantic text analysis and the analysis of multimodal coherence between text and image. The study aims to evaluate whether generated texts and images can serve as a "double" of the actual journalistic article. Here, "double" refers not to an exact copy, but to an alternative representation that preserves the theme and general meaning. The research seeks to answer three questions: how semantically close the generated texts are to the originals; how visually and conceptually close the generated images are to the actual photographs; and to what extent the textual and visual results are coherent within the context of an initial article.

2 Related Work

Recent advances in generative models development raised the necessity of reliable detection of synthetic news content as a combination of generated language and imagery. Schlee et al. Schlee et al., 2025 introduced SyN24News, a corpus derived from N24News using ChatGPT and Midjourney, and developed an interpretable multimodal classifier to distinguish real from synthetic articles. Their study provides a close precedent for constructing corresponding original and generated news materials. Our study addresses a different question: how closely generated texts and images resemble their original counterparts on the base of a collection of Romanian-language news articles from Moldova.

For textual evaluation, Reimers and Gurevych Reimers and Gurevych, 2019 introduced Sentence-BERT, which represents sentences as embeddings that can be compared using cosine similarity. This provides a semantic perspective alongside measures of lexical overlap. However, similarity and faithfulness require separate consideration. Through human evaluation of abstractive summaries, Maynez et al. Maynez et al., 2020 found substantial unsupported content and reported that entailment-based measures correlated better with faithfulness than standard metrics. These findings motivate our use of complementary textual measures while placing limits on what similarity scores alone can establish about factual preservation.

Research on news imagery also distinguishes visual relevance from correspondence to a particular photograph. Heitz et al. Heitz et al., 2025 describe the NewsImages 2025 task, which compares

image retrieval and generation for news articles and specifies human assessment of image fit and relevance. Hessel et al. [2021] proposed CLIPScore for evaluating image–caption compatibility, but found it less effective for news captions requiring contextual knowledge. Together, these studies motivate examining different aspects of visual correspondence. Our analysis compares original and generated images through CLIP image embeddings and structural similarity, alongside the textual comparisons; it does not apply CLIPScore or directly measure text and image alignment.

3 Research Methodology

The study uses a dataset of 141 online news articles from the Moldovan media space, with the original text and the original image, each paired with an LLM-generated text, and a text-to-image generated picture. The research evaluates the ability of generative models to preserve semantic meaning, thematic vocabulary, and multimodal coherence. Text analysis relies on Sentence-BERT, TF-IDF character ngrams, Jaccard similarity, BLEU, and Bag-of-Words cosine similarity. Image analysis relies on CLIP embeddings, cosine similarity, SSIM, L2 distance, FID, UMAP, and t-SNE.

3.1 Dataset

The analyzed dataset contains 141 online news articles collected from the unimedia.md archive and 141 generated texts. Each real article is accompanied by a metadata file with the following elements: source link, creation date, title, author, topic, summary.

For each collected article, a photography was downloaded as well and the generated texts were based on this photography. GPT-5.1 Instant was used for the text generation and the prompt sounded as follows: Generate a news article based on this image, in approximately 140 words. Metadata for generated articles contained link to the original news, source image, model used for the generation and the prompt.

Additionally, an image was generated for each article based on the textual content. From a methodological standpoint, this workflow allows for the comparison of two types of transformations: original text to generated text and original text to generated image.

It should be noted that the prompt strongly influences the length and specificity of the generated text. The results show that the generated texts are considerably shorter than the original texts, suggesting that the model worked mostly as a summarizer and paraphraser.

3.2 Experiment Methodology

The analysis of the texts was conducted using several complementary metrics. Sentence-BERT was used to estimate the semantic similarity between the original and generated texts. Character n-gram TF-IDF measured proximity in terms of textual form, Jaccard measured vocabulary overlap, BLEU measured the matching of lexical sequences, and Bag-of-Words evaluated the word frequency distribution ignoring the word order. The simultaneous use of these metrics is important because two texts may be semantically close but lexically and structurally very different. For the images, CLIP visual embeddings were used, calculated separately for the real photograph and the generated image. The cosine similarity between the embeddings served as a measure of the conceptual closeness of the two images. SSIM was used to measure structural similarity, while L2 distance quantified the separation between the vector representations. An FID-style score was calculated globally to compare the distributions of real and generated embeddings. Additionally, the distributions were projected into two dimensions using UMAP and t-SNE to visually observe whether the real and generated images formed distinct clusters or overlapped.

4 Results

Table 1 presents the results for the text comparison. The average SBERT score of 0.478 indicates moderate semantic closeness between the original and generated texts. While the model successfully preserves the general topic of many articles, it does not guarantee informational

fidelity. The discrepancy between SBERT and BLEU scores is significant: even when the general meaning is similar, the exact wording differs almost entirely. This confirms that the generated texts are paraphrases or summaries rather than verbatim reproductions of the articles. The Jaccard score averages only 0.076, indicating minimal vocabulary overlap. However, the Bag of Words Similarity average of 0.435 suggests that key terms or general thematic areas are preserved to some extent. In other words, the model does not use the same specific words but tends to retain the article’s core thematic elements.

Indicator	Mean	Std. Dev.	Min.	Max.
SBERT Similarity	0.478	0.157	0.077	0.808
TFIDF Character Similarity	0.289	0.086	0.094	0.594
Jaccard Similarity	0.076	0.027	0.0046	0.205
BLEW	0.005	0.006	0.000	0.051
Bag of Words similarity	0.435	0.112	0.186	0.651

Table 1: Summary of textual metrics for for real and generated article pairs.

Indicator	Mean	Std. Dev.	Min.	Max.
CLIP cosine similarity	0.538	0.139	0.174	0.872
SSIM	0.168	0.072	0.023	0.507
Real embedding norm	0.205	0.825	8.443	12.066
Generated embedding norm	9.158	0.768	7.710	12.350
Embedding L2 distance	9.243	1.496	5.123	12.939
FID-style global	52.358	-	52.358	52.358

Table 2: Synthesis of visual metrics for real and generated image pairs.

Table 2 contains the metrics for the comparison of real and generated images. The CLIP cosine similarity averages 0.538, indicating a moderate conceptual closeness between the real photograph and the generated image. In many cases, the text-to-image model produces an image belonging to the same general category such as a street, an accident, a political meeting, a building, a person. However, SSIM averages only 0.168, showing that the concrete visual structure differs. The model does not reconstruct the photograph but instead generates a new image inspired by the description.

CLIP cosine similarity indicates moderate visual-semantic proximity; SSIM values suggest low structural similarity. Real embedding norm demonstrate stable representations for real images and generated embedding norm represent similar orders of magnitude. Embedding L2 distance values show large distance between real and generated pairs. FID style global values suggest that the distributions are close in terms of domain but not identical.

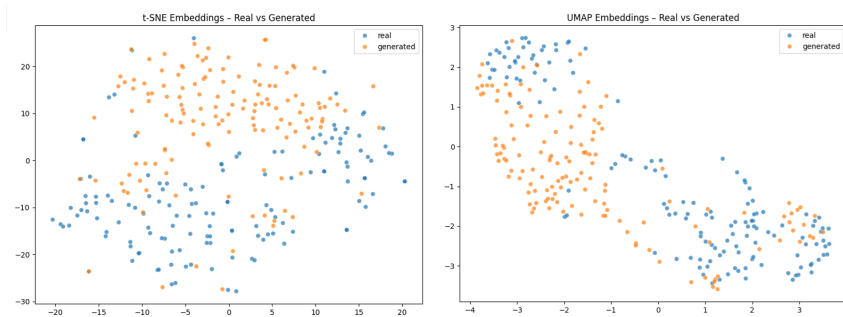


Figure 1: t-SNE (left) and UMAP (right) projection of visual embeddings for real and generated images

Figure 1 presents the t-SNE and UMAP projection of visual embeddings for real (blue dots) and generated (orange dots) images. Both the t-SNE and UMAP projections show that they form distinct clusters in the feature vector space. Although they are not completely isolated they do not overlap completely. This indicates clear differences in visual distribution and style between real journalistic photographs and AI-generated synthetic images. This visible separation in t-SNE and UMAP demonstrates that visual embeddings calculated via CLIP contain sufficient discriminative information to distinguish between authentic and artificially generated visual content and these metrics can be successfully used in an automated classification model.

5 Conclusion

The general conclusion is that AI models are useful for summarizing, paraphrasing, and generating generic images, but cannot replace authentic journalistic content in contexts where accuracy, sourcing, visual identity, and editorial accountability are essential. For academic and research purposes, the presented workflow is valuable because it offers a reproducible framework for multimodal evaluation and can be extended to the automated detection of artificially generated content.

References

- Heitz, L., Rossetto, L., Kille, B., Lommatzsch, A., Elahi, M., & Dang-Nguyen, D.-T. (2025). NewsImages in MediaEval 2025 – comparing image retrieval and generation for news articles. *Working Notes Proceedings of the MediaEval 2025 Workshop*. <https://2025.multimediaeval.com/paper1.pdf>
- Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. <https://doi.org/10.18653/v1/2021.emnlp-main.595>
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP/IJCNLP)*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Schlee, M., Kant, G., Ehrling, C., Säfken, B., & Kneib, T. (2025). Decoding synthetic news: An interpretable multimodal framework for the classification of news articles in a novel news corpus. *Artificial Intelligence Review*, 58, 302. <https://doi.org/10.1007/s10462-025-11188-9>

Rodrigo Santos

Luís Gomes

Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal, {rsdsantos, luis.gomes}@fc.ul.pt

Afonso Vaz

Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal,

fc64505@alunos.ciencias.ulisboa.pt

António Branco

Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal, antonio.branco@di.fc.ul.pt

RAG vs. Fine-Tuning for Chatbot Enhancement: Empirical Study with a Real-World Public Services Use Case

Keywords: Large Language Models, Retrieval-Augmented Generation, Sovereign AI, European Portuguese, Chatbot

I Introduction

Governments are increasingly deploying AI-powered chatbots to handle citizen queries about public services around the clock. These deployments typically rely on proprietary, cloud-hosted LLMs operated by a small oligopoly of global technology providers. Such a dependency raises serious concerns about digital and cultural sovereignty, data privacy, and long-term autonomy, particularly for applications that deal with sensitive citizen information (Branco et al., 2026).

The alternative, adapting open on-premises LLMs to the target domain, raises the practical question of how to best specialise a general-purpose model. Two methodologies dominate the literature. Fine-tuning, which updates model weights on domain data, and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which augments an unchanged model with knowledge from a database retrieved at inference time. Both have complementary strengths, yet a rigorous empirical comparison in a real governmental deployment scenario is largely absent from the literature.

This paper fills that gap by reporting experiments conducted on both fine-tuning and RAG for a public services chatbot. Specifically, we use Gervásio 8B (Santos et al., 2024), a fully open LLM specialised for European Portuguese, as the base model, and evaluate progressively fine-tuned variants alongside RAG-enhanced configurations, and compare all systems against the commercial solution currently in production at gov.pt.¹

Our main contributions are: (1) a report on a fine-tuning pipeline for a government chatbot domain, including details on the data used; (2) a direct comparison of fine-tuning, RAG, and their combination under controlled evaluation conditions; and (3) evidence that open sovereign AI achieves performance competitive with, and in some metrics superior to, a proprietary production system.

2 Data

2.1 Training data

Two training datasets were prepared for fine-tuning.

The first, `govpt2025`, consists of texts scraped from gov.pt, a national public services web portal.² It contains 6,204 documents with over 160,000 paragraphs, totalling more than 1.5 million words.

The second, `dialogs-govpt2025`, consists of real dialogues between citizens and the virtual assistant from the gov.pt web portal.¹ The raw data was anonymised before being shared with

¹ <https://gov.pt>

² Collected on October 2025.

the research team. It contains approximately 3,500 dialogues in Portuguese (over 5,000 turns). After quality filtering — removing dialogues with typos, formatting errors, out-of-domain questions, or incorrect automatic responses — 1,686 usable turns remained (about 32% of the original corpus).

2.2 Development and test data

The development set consists of 88 turns (65 after deduplication) constructed manually in 2024 to support development and testing of the commercial chatbot.

The test set was constructed independently by our research team and contains approximately 300 turns. Half are direct questions; the other half are verbose paraphrases of those questions enriched with contextual information, as observed in real user queries. Verbose versions were drafted with the assistance of a frontier LLM and then manually edited. The gold answers are excerpts from official portal texts. Following best practice, this set was never used during training or development.

3 Models and fine-tuning

3.1 Base models and baselines

Our primary base model is Gervásio 8B (G8B), a fully open (parameters, licence, and distribution) 8-billion parameter decoder LLM specialised for European Portuguese (Santos et al., 2024). It is representative of the class of open models that can realistically be deployed on-premises by a public institution.

Two baselines are defined: (1) G8B, the unmodified Gervásio 8B model, and (2) G8B RAG, the same model operating in RAG mode over the textual content of the gov.pt web portal.

3.2 Fine-tuned models

Two fine-tuned model variants were produced, all using PEFT (Parameter Efficient Fine-Tuning) (Hu et al., 2022) with causal language modelling: (i) G8B-cpt is obtained by continued pre-training of G8B on govpt2025, adapting the model to the vocabulary and content of the portal. (ii) G8B-ift is obtained by applying instruction fine-tuning to G8B-cpt using dialogs-govpt2025 (clean), using a zero-out (Touvron et al., 2023) causal language modelling objective.

Additionally, G8B-ift RAG integrates G8B-ift into the same RAG pipeline used for the G8B RAG baseline, combining the benefits of both fine-tuning and retrieval.

4 Evaluation

4.1 Metrics

Two metrics are used, both computed via an LLM-judge (Zheng et al., 2023) approach using a state-of-the-art frontier model accessed through its API. To account for the stochastic nature of LLM generation, each model runs 3 times per example, and the judge evaluates each run 3 times, yielding 9 values per example. The reported score is the mean of these 9 values.

Semantic similarity measures how close the response of the model is to the gold answer on a scale of 0 (no similarity) to 5 (full equivalence).

Accuracy measures the proportion of questions answered correctly, where correctness is a binary CORRECT/INCORRECT judgement by the LLM-judge (without reference to the gold answer). Calibration against manual inspection revealed an uncertainty of approximately 16% for this metric, making it less stable than semantic similarity.

¹ Collected from early 2025 to January 2026.

4.2 Results

Table 1 reports all model scores on both the development set and the test set.

Table 1: Evaluation results across all models on development and test sets. *Results taken from (Branco et al., 2026). “—” indicates not evaluated on this set. Sim. = semantic similarity; Acc. = accuracy.

Phase	Model	Dev set		Test set	
		Sim. [0-5]	Acc. %	Sim. [0-5]	Acc. %
Phase 1	G8B-ift RAG	2.46	33.85	3.89	39.50
	G8B-ift	2.75	40.51	2.08	22.37
	G8B-cpt	2.29	25.98	2.11	17.58
Baselines	G8B RAG	2.77	44.10	4.00	63.55
	G8B	2.08	17.26	1.41	6.39
Phase 2	G70B RAG*	—	—	4.03	69.10
	G70B*	—	—	2.02	14.16
Prod. system	GOVPT	—	—	3.66	75.11

5 Discussion

5.1 Effect of Fine-tuning

Fine-tuning on portal texts (G8B-cpt) already delivers a meaningful gain over the unmodified base model. Test set accuracy rises from 6.39% to 17.58% and semantic similarity from 1.41 to 2.11. A second round of instruction fine-tuning on real user dialogues (G8B-ift) further improves accuracy to 22.37%. Taken together, the fine-tuning pipeline increases accuracy by more than 3.5× compared to the base model, confirming that domain and task adaptation are both meaningful and cumulative.

5.2 Effect of RAG

The comparison between G8B and G8B RAG is striking. Adding RAG raises semantic similarity from 1.41 to 4.00 and accuracy from 6.39% to 63.55% on the test set. RAG alone, applied to the same base model without any fine-tuning, dramatically outperforms all fine-tuned non-RAG variants. This is consistent with the nature of the test set, whose gold answers are drawn directly from portal texts, the very documents indexed in the RAG knowledge base.

When RAG is added on top of the fine-tuned model (G8B-ift RAG), the test set semantic similarity rises to 3.89, though accuracy (39.50%) remains below the simpler G8B RAG baseline (63.55%). This suggests that fine-tuning may shift the model’s generation style in ways that occasionally conflict with strict retrieval-grounded answering.

These results point towards the creation of a dataset that tries to boost retrieval-grounded answering, for example, a dataset where a retrieved ground truth is coupled to the question and answer pairs of the dialogs-govpt2025 dataset.

Furthermore, when looking at the results obtained by Branco et al., 2026 with a 70-billion parameter Gervásio model (Santos et al., 2024) in RAG mode (G70B RAG), they demonstrate a semantic similarity of 4.03 and an accuracy of 69.1%, surpassing the results of the 8B model family across the board and approaching the results of the commercial system. The gain from RAG is equally pronounced at this scale.

5.3 Comparison with the production system

The production commercial system scores 3.66 in semantic similarity and 75.11% in accuracy. G70B RAG already surpasses it in semantic similarity (4.03 vs. 3.66). The accuracy gap (69.1% vs. 75.1%) corresponds to fewer than 8 questions out of ≈300 in the test set, and lies within the calibration uncertainty of approximately 16% for this metric. These results provide empirical

evidence that sovereign open AI solutions are technically competitive with, and on some measures superior to the proprietary system currently serving millions of citizens.

5.4 Discussion

Several observations deserve special attention. First, the test set proves harder than the development set for fine-tuned models without RAG (e.g. G8B TD: 40.5% on dev vs. 22.4% on test), but easier for RAG-equipped models (G8B RAG: 44.1% on dev vs. 63.6% on test). This inversion reflects the fact that test-set gold answers are text excerpts from the portal, making RAG retrieval especially effective, while the development set was constructed differently.

Second, data quality matters greatly. Of the 5,281 turns in the dialogue corpus, only 1,686 (32%) were usable after filtering. Investing in data curation yields cleaner fine-tuning signal and should be prioritised.

Third, the LLM-judge accuracy metric is less stable than semantic similarity ($\approx 16\%$ calibration uncertainty), and results based on it should be interpreted with appropriate caution.

6 Conclusion

We have presented a systematic empirical comparison of fine-tuning and RAG for adapting open LLMs to a government public services chatbot. Our findings show that: (i) fine-tuning consistently and cumulatively improves domain performance, raising accuracy by more than 3.5x; (ii) combining the fine-tuned model with RAG produces even larger gains; (iii) standalone RAG performs better than any of the previous combinations, pointing towards the need for more appropriate fine-tuning algorithms and training data while using RAG; and finally, (iv) a 70B open model with RAG matches or exceeds a commercial proprietary system on both evaluation metrics, at a lower cost and with full on-premises sovereignty.

Future work should focus on the creation of a dataset that tries to boost retrieval-grounded answering, expanding the dialogue training corpus as more real-user data accumulates, and fine-tune a 70B parameter model to understand if the conclusions obtained throughout the paper transfer to a bigger model.

All in all, these results provide a strong empirical foundation for the adoption of sovereign open AI solutions in public-service applications.

References

- Branco, A., Gomes, L. M. S., Santos, R., Santos, E., Silva, J. R., Marques, N., & Rodrigues, M. (2026). Sovereign AI-based public services are viable and affordable. *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, 8235–8245.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *lcl*, 1(2), 3.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459–9474.
- Santos, R., Silva, J., Gomes, L., Rodrigues, J., & Branco, A. (2024). Advancing generative ai for portuguese with open decoder gervásio pt. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, 16–26.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36, 46595–46623.

Bence Sárossy

ELTE Research Centre for Linguistics, Budapest, Hungary, sarossy@nytud.hu

Augmenting Fake News Corpus with Information from LLMs

Keywords: fake news, LLM, glossing

I Introduction and motivation

The spreading and volume of fake news are a growing threat to societies all around the world. They endanger especially democracies, where pieces of information can be spread without control due to freedom of speech, even if they are emitted by obscure sources. Fake news always existed, yet in the digital age creators of fake news have a way more opportunity and tools to disseminate false information than before. Distinguishing between fake news and real news is a great endeavour, which still needs to be developed. To make distinction between real and fake news, many fake news detectors were introduced, that make use of certain linguistic features and classify text as fake news using machine learning. In this paper we present a novel approach to use for processing fake news, which based on glossing named entities with background information from LLM's. With these information contrasted with the information on named entities in the fake news articles themselves we can get a deeper knowledge on fake news, which can have relevance in fake news detection as well.

2 Defining fake news

A first task is to define what counts as fake news in the context of corpus building. In public discussion, the term is often used broadly and inconsistently (Lazer et al., 2018; Wardle & Derakhshan, 2017). For corpus purposes, however, the inclusion criteria must be explicit. The goal is to make the material identifiable and usable for more focused analytical work.

We define fake news primarily based on the intent of the author, whose goal is – as opposed to authors of real news – not to inform but to deceive the public (Jack Grieve, 2023).

As source of fake news we used a list of fake news portals collected by a fact-checking website, *Urban Legends* (Urban Legends, 2026). This list includes websites that publish clickbait articles on medical, lifestyle, and tabloid topics, as well as conspiracy theories and propaganda. Although the fake news list also featured sites focused on spiritual and occult content, we did not categorize these as fake news, as such topics fall into a grey area of personal and religious belief.

3 Data collection and processing

We collected the data by scraping the following websites. They mostly contain sitemaps that made scraping easy: budapest-news.com, erdekessegek.net, mivilagunk.com, otletdivak.hu, rejtelyekszigete.com, kimondottigazsagok.com, titkokszigete.hu, tudnodkell.info, vilagunk.hu, tudasfaja.com, oroszhirek.hu, napitema.com, napiufo.hu, hirextra.hu, frissvilag.com.

The total number of tokens for these websites is 30 million, which is considerably low compared to traditional press websites; yet, it is big enough to carry out comparative analyses based on them.

The technical processing of these materials includes the extraction of the article text from the webpage, removal of boilerplate elements, filtering of non-Hungarian or non-relevant pages, and documentlevel deduplication. These steps are necessary because web-based press materials often contain repeated blocks, navigational content, recommendation panels, and unstable page structure.

4 Research perspectives

From a linguistic point of view, fake news texts are interesting because they may differ from other press materials in recurring ways (Rashkin et al., 2017). These differences may appear in headline structure, lexical choice, evidentiality, evaluation, the treatment of named entities, or rhetorical

strategies of urgency and threat. Such questions are difficult to study without a corpus that stores the material in a structured and searchable form.

In the present research we used LLM for named entity glossing of fake news articles with background information and contrasted these with information on the named entities found in fake news articles themselves.

As for future possible research, we foresee the use of LLMs in annotating fake news articles for persuasion techniques and loaded language (for other languages see (Da San Martino et al., 2019; Piskorski et al., 2023), (Huang et al., 2023)).

The glossed corpus could be a useful tool for research in the area of political science. Revealing intentions for manipulation of the public leads to more transparent public discourse and can serve as an important method for defending democratic values.

5 Conclusion

This paper has outlined the some methods for procseeing fake news texts with LLMs. The aim is to open new ways for fake news detection, which has enormous social and political significance. Although the collected fake news corpus is relatively small, its size is already sufficient for comparative analyses.

References

- Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5640–5650. <https://doi.org/10.18653/v1/D19-1565>
- Huang, K.-H., McKeown, K., Nakov, P., Choi, Y., & Ji, H. (2023). Faking fake news for real fake news detection: Propaganda-loaded training data generation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14571–14589. <https://doi.org/10.18653/v1/2023.acl-long.815>
- Jack Grieve, H. W. (2023). *The language of fake news*. Cambridge University Press. <https://doi.org/10.1017/9781009349161>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., & Nakov, P. (2023, July). Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, (pp. 3001–3022). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.169>
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. <https://doi.org/10.18653/v1/D17-1317>
- Urban Legends. (2026). *Urban legends – legendák, hoaxok, álhírek tényellenőrzése*. Retrieved April 17, 2026, from <https://www.urbanlegends.hu/>
- Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (tech. rep.). Council of Europe. <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>

Marina Sokolova Zolotilova

University of Illinois at Chicago, Chicago, USA, msokolopez@gmail.com

Natalia Moskvina

Autonomous University of Barcelona, Barcelona, Spain, natalia.moskvina@uab.cat

Nayara Mirio e Silva

University of Barcelona, Barcelona, Spain, nayara.mirio.silva@alumni.usp.br

Code-Switching as a Parsing Cue: Humans vs. LLMs

Keywords: code-switching, parsing cues, human processing, LLM processing, RC attachment resolution

1 Introduction

The paper examines the processing effect of code-switching (CS) in humans and large language models (LLMs) and tests whether CS imposes a prosodic break that serves as a processing cue for humans and whether it is a strong parsing cue for machines. We use ambiguous relative clauses (RCs) (1) and test if place of CS affects sentence processing and ambiguity resolution.

- (1) Bill saw [DP the friend of the neighbor [RC that was drinking coffee]]
a. [DP [DP the friend of the neighbor] [RC that was drinking coffee]] (*HA, the friend was drinking coffee*)
b. [DP the friend of [DP the neighbor [RC that was drinking coffee]]] (*LA, the neighbor was drinking coffee*)

The place of CS is motivated by the Implicit Prosody Hypothesis (IPH, Fodor 2002) summarized in (2).

- (2) Bill saw [DP the friend of the neighbor [RC that was drinking coffee]]
a. [DP [DP the friend of the neighbor] pause [RC that was drinking coffee]] (*HA*)
b. [DP the friend pause of [DP the neighbor [RC that was drinking coffee]]] (*LA*)

Thus, languages favoring high attachment (HA) have a prosodic break before the RC, whereas a break before the preposition is associated with low attachment (LA) languages. The assumptions of IPH have been extended and tested in processing experiments showing that manipulating the placement of a prosodic break affects RC resolution. In auditory experiments, an HA preference has been observed when a prosodic break is placed before the RC in, for example, English, thus supporting the generalized predictions of the IPH (Guzzo et al., 2021; Fromont et al., 2017). We extend this work to silent reading and to bilingual contexts involving humans and LLMs, using CS as an implicit prosodic cue that shapes RC parsing.

2 Theoretical background

Studies with ambiguous RCs began with experiments in English which led to the assumption that LA was a universally preferred parsing strategy: it is cognitively less costly to attach a node to the most recently processed one (Late Closure; Frazier & Fodor, 1978). This assumption was challenged by Cuetos and Mitchell (1988), who showed a consistent preference for HA among native speakers of Spanish. Subsequent work established cross-linguistic differences in RC attachment, with speakers of languages such as Norwegian and English preferring LA, and speakers of languages such as Spanish and Russian preferring HA (see Fodor, 2002; Grillo and Costa, 2014). However, the source of these preferences remains unclear, and even less is known about machine processing.

One account (Gibson et al., 1996) proposes that languages differ in their governing parsing principles. Languages with richer morphology may follow Predicate Proximity (favoring HA), whereas others follow Recency (Late Closure, favoring LA). Later work shows that in the

absence of explicit cues, human parsers rely on semantic and pragmatic information to resolve RC ambiguity (MacDonald et al., 1994; Gilboy et al., 1995; Hemforth et al., 2015). In contrast, LLMs show little sensitivity to such information, although they replicate cross-linguistic attachment preferences (e.g., LA in English, HA in Spanish) (Scheinberg et al., 2025; Amouyal et al., 2025; Lee et al., 2024, 2025).

The Implicit Prosody Hypothesis (IPH; Fodor, 2002), which forms the basis of the present study, proposes that prosodic structure guides parsing even in silent reading. It predicts that a prosodic break before the RC favors HA, whereas a break within the noun phrase favors LA. This hypothesis has been supported by auditory studies manipulating prosodic boundaries but remains difficult to test in silent reading (Fromont et al., 2017; Goad et al., 2020).

We propose that code-switching provides a natural way to elicit such implicit prosodic effects. Specifically, we hypothesize that the place of CS induces a prosodic break at the point of language change, affecting RC parsing as predicted by the IPH in human subjects. For LLMs, which lack prosodic representation, language change may instead function as a structural segmentation cue.

We put forward the following Research Questions (RQs):

Human processing:

Does the place of code-switching affect RC parsing as predicted by the IPH?

LLM parsing:

Does CS function as a parsing cue (and therefore affect RC parsing as predicted by the IPH)?

Does model behavior parallel human sensitivity to prosodic effects?

3 Experiment and Results

The design manipulates the position of CS, favoring HA (3a) vs. LA (3b). In the human experiment, CS is kept in the L1→L2 direction across language pairs to limit the number of experimental variables, keep the experiment at a manageable length, and avoid making the purpose of the study too apparent to participants. Although CS direction is therefore held constant in the human experiment, the study includes a wider range of language pairs, allowing us to examine whether the effects generalize across languages. Unlike the human experiment, the LLM experiment includes CS direction as an additional variable, with both English→Spanish and Spanish→English tested. Because LLMs are not subject to the same constraints on experiment length and participant awareness, we could vary CS direction while limiting the language pairs to English and Spanish, the two most extensively studied languages in this context. In both experiments, baseline attachment preferences are established using unilingual sentences, and RC attachment preference is measured based on the generated response.

Humans. A 1×2 design contrasts HA- vs. LA-cueing CS (16 items), with a unilingual baseline (16 items): 32 target items in total and 48 fillers. Example stimuli are given in (3).

(3) Design Humans, *italics marks a language change*

RC attachment by place of CS

- a. Bill arrested the friend of the neighbor *that was drinking coffee* (HA by CS)
- b. Bill arrested the friend of the neighbor *that was drinking coffee* (LA by CS)

RC attachment by Language (16 sentences)

- c. Eng: Bill arrested the friend of the neighbor that was drinking coffee (LA by Language)
- d. Rus: Ivan arestoval druga sosedu kotoryj pil kofe (HA by Language)
- e. Nor: Bill arresterte vennen til naboen som drakk kaffe (LA by Language)
- f. Port: O Guilherme prendeu o amigo do vizinho que estava a beber café (LA by Language)

LLMs. A 2×2 design crosses CS position (HA vs. LA) and direction (English→Spanish vs. Spanish→English) (16 items per condition), with unilingual baselines (8+8 items): 48 target items in total and 48 fillers. Example stimuli are given in (4).

(4) Design LLMs, *italics marks a language change*

RC attachment by place of CS vs. Direction of CS

- a. Bill called the friend of the neighbor *que hablaba de fútbol en el patio*
- b. Bill called the friend *del vecino que hablaba de fútbol en el patio*
- c. Juan vio al amigo del vecino ^[break] *that was talking about football on the patio*
- d. Juan vio al amigo ^[break] *of the neighbor that was talking about football on the patio*

RC attachment by Language (8 + 8 sentences)

- e. Juan vio al amigo del vecino *que hablaba de fútbol en el patio* (HA by Language)
- f. Bill called the friend of the neighbor *that was talking about football* (LA by Language)

Adult L2 speakers (upper-intermediate to advanced) participated in a self-paced reading experiment (Linger) with word-by-word presentation. RC attachment was assessed using comprehension questions. Participants included 149 adult L2 speakers (upper-intermediate to advanced) across five language groups (Norwegian–English, Russian–Norwegian, Portuguese–English, Spanish–English, Catalan(Spanish)–English), with mean ages ranging from 22 to 29 years and high proficiency scores (≈ 80 –95%). The LLM analysis includes GPT-4.1 and DeepSeek-V4-Flash, which represent distinct transformer-based LLM families with different architectural and training paradigms. Human data were analyzed with GLMMs, and LLM data with χ^2 tests comparing HA vs LA distributions.

The results show that humans generally favor LA in CS sentences (≈ 70 –30), with the placement of CS producing a modest shift ($\sim 4\%$) toward the attachment type cued by its position. This effect increases to $\sim 13\%$ in language pairs whose baseline RC attachment preferences differ across languages (e.g., Norwegian→Russian, English→Spanish), supporting the role of prosodic structure in guiding RC parsing in humans whose linguistic background includes languages with different attachment preferences (GLMM: main effect of CS position, $p < .05$; interaction with group, $p < .05$; see Figure 1).

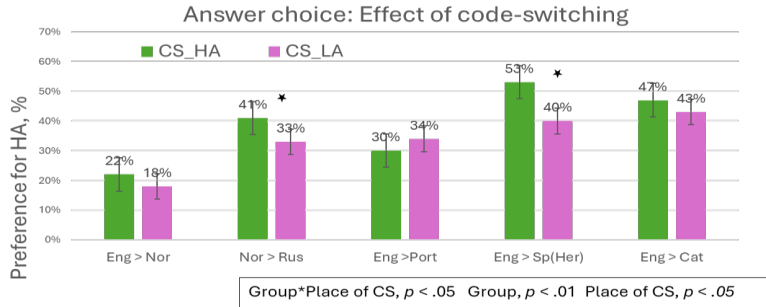


Figure 1. Sensitivity to the place of CS RC processing by humans

The LLM results reveal both convergences and divergences across models (see Figure 2). GPT-4.1 shows an effect of CS position only in Spanish→English switching, while maintaining a strong overall LA preference in English→Spanish CS and in unilingual English and Spanish ($\chi^2(1) = 486.0$, $p < .001$). DeepSeek, by contrast, shows sensitivity to the position of CS in both switching directions. In Spanish→English CS, its behavior parallels GPT-4.1 and follows the predictions of the IPH, whereas in English→Spanish CS it shows a robust effect in the opposite direction. The two models also differ in their unilingual baselines: GPT-4.1 shows near-categorical LA in both English and Spanish, whereas DeepSeek reproduces the human-like LA preference in English but shows no stable attachment preference in Spanish. DeepSeek additionally demonstrates increased instability in code-switched sentences, with instances of ungrammatical responses, such as selecting the matrix subject as the answer, becoming more frequent under CS. The proportion of glitches rises from 1.48% in unilingual conditions to 6.03% under CS, indicating that language switches increase the likelihood of anomalous responses.

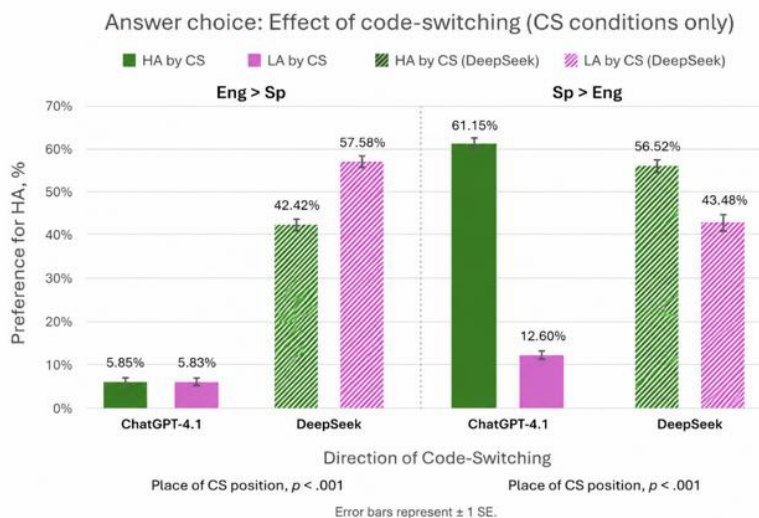


Figure 2. Sensitivity to the place of CS in RC processing by LLMs

4 Discussion and Further Work

Overall, the findings suggest that while humans with exposure to languages that differ in attachment preferences use CS as a probabilistic prosodic cue, LLMs treat language switches as strong, direction-dependent structural signals, resulting in asymmetric and inconsistent parsing behavior across language pairs.

Future work will extend this comparison by testing human participants with code-switching in both directions within the same language pairs and by evaluating a broader range of LLMs from different model families.

References

- Cuetos, Fernando, and Don C. Mitchell. (1988) Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy. *Cognition* 30(1): 73–105.
- Fodor, Janet Dean. (2002) Prosodic disambiguation in silent reading. In *Proceedings of the North East Linguistic Society (NELS 32)*, ed. Masako Hirotsu, 113–132.
- Frazier, Lyn, and Janet Dean Fodor. (1978) The sausage machine: A new two-stage parsing model. *Cognition* 6(4): 291–325.
- Fromont, Laurent A., Salvador Soto-Faraco, and Emmanuel Biau. (2017) Searching high and low: Prosodic breaks disambiguate relative clauses. *Frontiers in Psychology* 8: 96. <https://doi.org/10.3389/fpsyg.2017.00096>
- Goad, Heather, Natalia Guzzo, and Lydia White. (2020) Parsing ambiguous relative clauses in L2 English: Learner sensitivity to prosodic cues. *Studies in Second Language Acquisition* 43(1): 83–108.
- Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. (2024) Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences* 121(36): e2400917121. <https://doi.org/10.1073/pnas.2400917121>
- Kamerath, Michael, and Aniello De Santo. (2025) Do LLMs disambiguate Italian relative clause attachment? In *Proceedings of the Society for Computation in Linguistics (SCiL)*.
- Lee, So Young, Russell Scheinberg, Amber Shore, and Ameeta Agrawal. (2025) Multilingual relative clause attachment ambiguity resolution in large language models. arXiv preprint arXiv:2503.02971.
- Scheinberg, Russell, So Young Lee, and Ameeta Agrawal. (2025) Missing the cues: LLMs' insensitivity to semantic biases in relative clause attachment. *Proceedings of the Linguistic Society of America* 10(1). <https://doi.org/10.3765/plsa.v10i1.5902>

Gaurish Thakkar
Marko Tadić
Nives Mikelić Preradović

University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia, {gthakkar, marko.tadic, nmikelic}@ffzg.unizg.hr

Co-text-Aware Croatian Sentiment Annotation Using Open-Source Large Language Models

Keywords: sentiment analysis, evaluation, classification, large language models, Croatian language

I Introduction

Sentiment analysis is typically performed at the sentence level in isolation, a practice that ignores the discourse co-text¹ in which utterances occur. We present an evaluation framework for co-text-aware sentiment annotation in which each sentence is classified by an open-source large language model (LLM) under four co-text conditions: full document, co-text window, initial document sentences, and no co-text.

Sentiment analysis – the automatic classification of text as *positive*, *negative*, or *neutral* – is a foundational task in natural language processing with applications ranging from customer-feedback analysis to social-media monitoring (Liu, 2012; Pang & Lee, 2008). However, most deployed systems treat each sentence as a self-contained unit and thus struggle with utterances whose polarity can only be resolved by reference to what was said previously. Consider the phrase “It was too late”: in isolation, this is ambiguous. However, following a series of sentences describing e.g. project failures, it is unambiguously negative.

The rise of instruction-tuned LLMs has created new opportunities for co-text-aware annotation. Unlike supervised classifiers, LLMs can receive arbitrary natural-language co-text in a prompt, making them well suited to discourse-level reasoning. In this paper, we focus on the task of sentence-level sentiment classification in online movie reviews.

This paper makes three contributions:

1. new sentiment analysis **evaluation dataset in Croatian**, which is an expansion of the previous Cro-FiReDa dataset (Thakkar et al., 2023)
2. **prompt-based co-text-aware evaluation framework** in which each sentence receives the k surrounding sentences as discourse co-text during classification.
3. a QLoRA **fine-tuning study** on the best-performing model, examining prompt design effects.

2. Related work

Early sentiment analysis relied on lexicon-based (Turney, 2002) and rule-based approaches (Agić et al., 2010) before shifting to supervised machine-learning classifiers (Pang et al., 2002). Transformer-based models, notably BERT and its derivatives, subsequently set new benchmarks on sentence-level tasks (Devlin et al., 2019). Aspect-based sentiment analysis (ABSA) has extended the field to entity level labels (Pontiki et al., 2014), and recent work demonstrates that fine-tuned LLMs achieve state-of-the-art performance on ABSA sub-tasks (Šmíd & Dušek, 2026). Co-text-dependent sentiment, where the polarity of an utterance depends on prior discourse, has received comparatively little attention. Prior work by Poria et al., (2019) proposes a multi-modal model that uses con-text for conversational emotion recognition; Majumder et al., (2019) extends this to textual dialogue. However, both lines of work require task-specific fine-tuning. On the model comparison side, Lima et al., (2026) conducts an extensive benchmark of 29 open-

¹ In order to differentiate between a linguistic context and extra-linguistic context that would allow later usage of this framework with multimodal LLMs, we introduce the terms ‘co-text’ for strictly linguistic context and ‘con-text’ for extra-linguistic context such as pictures, sounds, videos, maps, etc.

source LLMs on Portuguese sentiment, finding that chain-of-thought prompting benefits larger models but that compact models are also competitive. Our study is complementary: we focus on Croatian movie reviews, prompting with co-text rather than prompting strategy variation, and provide an explicit F1 comparison across models.

3 Methodology

3.1 Four-Way Co-text Ablation Study

We evaluate sentiment classification under four co-text conditions to isolate the contribution of discourse information. Given a review, $C = (s_1, s_2, \dots, s_n)$ where s_i is the i -th sentence, the goal is to assign a sentiment label $y_i \in \{\text{positive}, \text{negative}, \text{neutral}, \text{mixed}, \text{other}\}$ to each s_i under four co-text conditions.

1. **In full co-text (FC)**, each s_i receives all the sentences from the document: $\text{co-text}(s_i) = C = (s_1, s_2, \dots, s_n)$.
2. **In co-text (CC)**, each s_i receives preceding and following sentences as a symmetric window: $\text{co-text}(s_i) = \{s_{\max(1, i-k)}, \dots, s_{i-1}, s_{i+1}, \dots, s_{\min(n, i+k)}\}$, i.e., every target sentence receives only the k surrounding sentences as framing co-text, testing whether local bidirectional context is sufficient for sentiment disambiguation.
3. **In initial co-text (IC)**, every s_i receives only the document’s first 5 sentences as global framing: $\text{co-text}(s_i) = \{s_1, \dots, s_5\}$ for all i , testing whether starting text of the document can help the classification process.
4. **In no co-text (NC)**, sentences are classified in isolation: $\text{co-text}(s_i) = \emptyset$, i.e., sentences are classified in isolation, serving as the baseline.

This design lets us ask: (i) does any co-text help over NC? (ii) does a local symmetric window (CC) outperform a fixed global framing (IC)? and (iii) what is the cost–accuracy trade-off for production use? We hypothesise that full bidirectional co-text will yield the highest F1 scores, particularly for discourse-dependent sentiment (e.g., pronouns and implicit polarity), while the no co-text baseline will struggle with ambiguous sentences lacking explicit sentiment markers.

3.2 A QLoRA fine-tuning Study

We fine-tune `google/gemma-4-31B-it` — selected on the basis of its IC and CC scores (§3.5) — using QLoRA: 4-bit NF4 quantization with double quantization, LoRA adapters injected into all linear projections. Training runs on a single NVIDIA A100-PCIE-40GB GPU. The Unsloth library (v2026.6.9) provides optimised kernels and memory management. At 4-bit, the 31B model occupies approximately 18–22 GB VRAM; the full training loop with gradient checkpointing peaks at ≈ 38 GB.

3.3 Dataset

We conducted an annotation campaign involving between 3 and 5 native Croatian speakers per data point to ensure reliability and capture inter-annotator variability. The data set was divided among multiple groups, and each annotator independently assigned sentiment labels to review sentences following detailed annotation guidelines, the final labels determined by majority voting or adjudication in cases of disagreement. To identify potentially mislabelled instances, we processed all annotated datasets using the `cleanlab` library (Northcutt et al., 2021), which computes a self-confidence label-quality score for each data point based on cross-validated predicted probabilities. Data points flagged with low confidence scores (below threshold $\tau = 0.5$) were passed through a second verification round, where annotators re-examined the original sentence in co-text and either confirmed or corrected the initial label. This two-stage quality-control procedure ensures that the final dataset minimises label noise while preserving genuine annotation ambiguity in borderline cases. The final evaluation set has 1,364 documents with 53.6k sentences. The label distribution is 12.5k positive, 6.4k negative, and 31.7k neutral, mixed 2.7k and other 115 reflecting naturally skewed sentiment. The other class — covering sarcasm,

idioms, proverbs, rhetorical questions, and meta-commentary — has no consistent surface form and represents fewer than 2% of training examples. The training / validation split is: positive 5k / 58, negative 1.9k / 92, neutral 9.2k / 115, mixed 1k / 48, and other 344 / 0. All entries retain their document and sentence IDs to ensure exact reproducibility. The dataset will be available under a permissive license in the HR-CLARIN repository¹.

3.4 Prompts

Three prompt variants were compared in both zero-shot and fine-tuning settings: a concise instruction without label definitions (P1, shown below), a definition-heavy variant (P2), and a few-shot variant with Croatian examples (P3). P1 consistently outperformed P2 (macro F1: 57.51% vs. 54.17%), suggesting that over-specifying label semantics constrains generalisation when training on labelled examples — the model learns better class boundaries from examples alone.

Prompt

System message: You are a sentiment analysis expert. You will be given a Croatian sentence along with its surrounding document context. Your task is to classify the sentiment of the TARGET sentence only, using exactly one of these labels: positive, negative, neutral, mixed, other. Respond with ONLY the label — no explanation, no punctuation. User template:
 Document context (sentences from the same document): {context}
 Target sentence to classify: "{sentence}"
 Sentiment label:

3.5 Experimental setup

We set $k = 5$ (a window of total 5 surrounding sentences). Temperature is fixed at 0.0 for deterministic, reproducible outputs. Label normalisation maps any synonymous surface form (*pos*, *good*, *LABEL 2*, etc.) to a canonical label. For models with a fine-tuned sentiment head (e.g., *classla/xlm-roberta-hr-sent*), we use direct inference, mapping output labels to our schema; context is ignored, as these models were trained on isolated sentences.

The following models were evaluated with IC and CC settings: *google/gemma-4-31B-it*, *google/gemma-3-27B-it*, *meta-llama/Llama-3.1-8B-Instruct*, *mistralai/Mistral-7B-Instruct-v0.3*, *meta-llama/Meta-Llama-3-8B-Instruct*, *Qwen/Qwen2.5-7B-Instruct*, *CohereForAll/aya-23-8B*, *google/gemma-2-9b-it*, *classla/bcms-bertic-parlasent-bcs-ter*.

Based on initial IC and CC evaluations, *google/gemma-4-31B-it* consistently outperformed all other models; accordingly, FC and NC conditions, as well as fine-tuning experiments, are reported for this model only. The remaining models are benchmarked under IC and CC but not analysed further.

We report macro-average F1, which treats all sentiment classes equally regardless of their frequency in the dataset, and weighted-averaged F1 metric, which accounts for class imbalance by weighting each class's contribution proportionally to its support.

4 Results and discussions

4.1 Overall model performance

Across all 9 open-source LLMs tested, performance varied significantly with model scale. The best-performing model, *google/gemma-4-31B-it*, achieved a macro F1 (CC) of 56.66% (weighted F1: 80.57%). The second-best model, *google/gemma-3-27B-it*, reached a macro F1 (FC) of 54.49% (weighted F1: 78.61%). Critically, all remaining 7 models failed to exceed 50% macro F1,

¹ <https://repository.clarin.hr>

suggesting that the size of the model plays a direct and substantial role in the performance of the zero-shot sentiment classification. The dramatic performance gap between the 31B and 27B Gemma variants and the widespread failure of sub-50% macro F1 among all other models indicate that scale is far more predictive of sentiment classification ability than model family or architecture. This corroborates recent findings in few-shot learning that parameter count strongly correlates with in-co-text reasoning capacity on complex multi-class tasks (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022).

Table 1: Zero-shot and fine-tuned performance of google/gemma-4-31B-it under four co-text conditions. NC = no co-text; IC = initial 5 sentences; CC = surrounding window (k=5); FC = full document.

Co-text condition	Macro-F1	Weighted-F1
NC	56.06	80.11
IC	56.36	80.53
CC	56.66	80.57
FC	56.16	80.86
Fine-tuned (FC)	57.51	81.48

4.2 Negligible effect of co-text

Contrary to our initial hypothesis, providing co-text — whether a full document, surrounding co-text windows (co-text with $k = 5$), or initial framing sentences — did not materially improve performance across the model population. This suggests that the manually annotated Croatian dataset may not inherently require discourse co-textualisation for consistent annotation or that the dataset’s characteristics (e.g., explicit sentiment markers and limited implicit polarity) diminish the utility of co-text at zero-shot inference time.

4.3 Error analysis: best-performing model

Error analysis of the best-performing model (*google/gemma-4-31B-it*) reveals a systematic pattern: the model struggles most with the `mixed` and `other` sentiment classes. In contrast, `positive` and `neutral` sentiment are classified with substantially higher accuracy and F1 scores. This imbalance likely reflects both the inherent difficulty of mixed-sentiment utterances (which require nuanced reasoning about competing sentiments) and potential class imbalance in the training/annotation distribution. The `other` category, typically a catch-all class in annotation schemes, may benefit from clearer definition or richer co-textual prompting to disambiguate from core sentiment labels.

4.4 The Mixed Class

Of 2,797 true-mixed sentences, only 1,435 (51.3%) are correctly predicted. The following example illustrates the pattern:

“Radnja je pomalo kompleksna za jedan dječji film , zbog mnogo putovanja kroz vrijeme i mnogo retrospekcije, ali svejedno ostajem pri mišljenju da je ovo čisto dječji film ali je i upitno da će se i njima dopasti”. (En: The plot is somewhat complex for a children’s movie, due to extensive time travel and flashbacks, but I still maintain that it is strictly a children’s film – though it is questionable whether even they will like it.) TRUE label: `mixed`, PRED label: `negative`. Negative criticism (“kompleksna za dječji film”) dominates; the positive qualifier after *ali* (but) is underweighted.

Croatian contrast markers — *ali* (but), *međutim* (however), *unatoč* (despite), *iako* (although) — are the critical discriminators. The model underweights these function words relative to the content words they connect.

4.5 The Other Class

The other class is not a single category — it covers five qualitatively different phenomena (rhetorical questions, sarcasm/irony, idioms/proverbs, meta-commentary, structural transitions) sharing only the absence of lacking evaluative stance. With 344 training examples split across five sub-types (≈ 69 per type), no classifier reliably learns any consistent decision boundary.

5 Conclusion

We have presented a systematic evaluation of 9 open-source LLMs on co-text-aware sentiment annotation for Croatian movie reviews, together with a QLoRA fine-tuning study of the best model. Our key findings are:

1. *google/gemma-4-31B-it* substantially outperforms all other models (56.66% macro F1 and weighted F1: 80.57%), with performance dropping sharply for smaller models;
2. co-text provision does not meaningfully improve zero-shot classification across any model, suggesting either dataset characteristics or fundamental limitations of zero-shot prompting in leveraging discourse;
3. fine-tuning the best model yields a further gain (macro F1: 57.51%), with concise prompts outperforming definition-heavy variants;
4. error analysis reveals systematic difficulties with the *mixed* and *other* classes while simpler classes (*positive*, *neutral*) are more reliably classified.

Future work will examine whether fine-tuning recovers the utility of co-text at scale, explore dynamic and asymmetric co-text windows, and investigate whether the *mixed* and *other* classes benefit from refined annotation guidelines or hierarchical classification strategies.

References

- Agić, Ž., Ljubešić, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in croatian. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 1164–1167.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. ArXiv, abs/2001.08361. <https://api.semanticscholar.org/CorpusID:210861095>
- Lima, J. V. R. J., Pinheiro, V., & Caminha, C. (2026). Portuguese sentiment analysis with open-source LLMs: Models, prompts, and efficient deployment. *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*, 212–221.
- Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.
- Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 6818–6825.
- Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research (JAIR)*, 70, 1373–1411.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86.
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.

- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 527–536.
- Šmíd, J., & Dušek, O. (2026). Leveraging fine-tuning of large language models for aspect-based sentiment analysis in resource-scarce environments. *Knowledge-Based Systems*, 312, 115277. <https://doi.org/10.1016/j.knosys.2026.115277>
- Thakkar, G., Mikielić Preradović, N., & Tadić, M. (2023, May). Croatian film review dataset (cro- FiReDa): A sentiment annotated dataset of film reviews. In J. Piskorski, M. Marcińczuk, P. Nakov, M. Ogrodniczuk, S. Pollak, P. Přibáň, P. Rybak, J. Steinberger, & R. Yangarber (Eds.), *Proceedings of the 9th Workshop on Slavic natural language processing 2023 (SLAVICNLP 2023)* (pp. 25–31). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bsnlp-1.4>
- Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 417–424.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (08/2022)

Tamás Váradi

ELTE Research Centre for Linguistics, Budapest, Hungary, varadi.tamas@nytud.hu

WoLaLa Wiki: A Research Wiki for Language and Language Model

Keywords: LLM Wiki, human supervised agentic workflow, scholarly knowledge curation

1 Introduction

We present WoLaLa Wiki, an experiment in developing the LLM Wiki pattern into a curated scholarly knowledge resource for research on language and language models. The project currently integrates 250 sources through differentiated ingestion and cumulative synthesis within a human-supervised agentic workflow. We describe methodological lessons arising from its development, illustrate how new literature modifies accumulated synthesis, and report a preliminary comparison with an expert synthesis of overlapping literature.

2 Motivation and rationale

WoLaLa Wiki started as an experiment with a new way of using large language models for research. The immediate inspiration was Karpathy’s proposal for an “LLM Wiki” (Karpathy, 2026): a collection of documents that an LLM gradually turns into a structured body of linked notes. Instead of treating the results of repeated work with documents as transient, useful results are preserved and can become the basis for subsequent work.

We explored this idea through WoLaLa, the Workshop on Language and Language Models, whose subject matter spans linguistics, NLP, cognitive science, philosophy of language, and research on large language models. Our question was what additional methods are required to develop Karpathy’s deliberately simple personal knowledge pattern into a controlled scholarly resource: how should knowledge derived from a growing research literature be selected, represented, maintained, synthesized and published?

Karpathy’s proposal rapidly gave rise to several experiments that develop the LLM Wiki pattern for different purposes. Zhang (2026) evaluates the Wiki as a knowledge base for operational decision support, while Huerta (2026) treats it as a compressed knowledge representation for LLM inference. Ming et al. (2026) develop it as an agent-native retrieval structure for multi-hop reasoning. Moreira and Sweet (2026) use it to preserve the history of collaborative knowledge work — including findings, decisions, failed approaches and revised claims — across researchers and agents. WoLaLa Wiki shares mechanisms with these systems but differs in its primary object: rather than primarily supporting a downstream task or preserving the history of a research project, the Wiki itself is intended as a persistent scholarly representation of an external research literature. Its central methodological problems therefore concern the controlled selection of sources, differentiated representation, cumulative synthesis and publication of scholarly knowledge.

3. From source collection to scholarly Wiki

WoLaLa Wiki began with a collection of papers and books assembled around the workshop themes and subsequently expanded through deliberate additions and cascading discovery from relevant sources. As the collection grew, distinctions emerged that are largely unnecessary for a small personal Wiki. In particular, we distinguish between possessing a source, admitting it to the research collection, representing it in the Wiki, and publishing its Wiki representation. Not every available document warrants ingestion, and not everything useful in the internal research environment is appropriate for public release. A quantitative breakdown of the resulting collection and its public subset is given in Table 1. The public WoLaLa Wiki is available online.¹

¹ <https://drop-6520b376-555.rwsnapshot-2608.workers.dev/>

A representative selection of the scholarly sources represented in the Wiki can be inspected separately.¹

3.1 Differentiated and cumulative ingestion

Admitted sources are processed at different levels of depth. Indexed sources are recorded but not given substantive Wiki treatment; brief ingestion captures their main contribution and relevance; deep ingestion reconstructs arguments, qualifications, relations to other work and points of disagreement in considerably greater detail. The choice is a scholarly judgment rather than a mechanical function of document length or prominence.

Table 1: Current size and composition of WoLaLa Wiki

	Current figure
Files in the source archive	773
Document files	620
Sources selected for the curated inventory	292
Ingested source pages	250
Deep source pages	92
Brief source pages	158
Indexed sources not yet ingested	42
Higher level Wiki pages	22
among them overview pages	11
concept pages	8
analysis pages	3
Public Wiki	
source pages	238
higher level Wiki pages	22
index page	1
total	261

This distinction emerged from experience. Early development tended to be depth-first, producing detailed treatments of a relatively small number of sources. At larger scale this risked creating islands of rich analysis surrounded by poorly represented areas. The workflow therefore shifted toward breadth-first coverage followed by selective deepening.

Ingestion is also cumulative. New sources are not treated simply as independent summaries. They are related to material already represented and may extend, qualify or challenge existing concept, analysis or overview pages. Earlier synthesis thus becomes part of the context in which later sources are interpreted. This is the form of compounding that is central to WoLaLa Wiki.

3.2 Research Wiki and public Wiki

The internal research repository and public Wiki serve different purposes. The former contains the working environment in which WoLaLa Wiki is developed; the latter is a separately generated reader-facing resource. Candidate pages are checked for links, metadata, provenance and accidental exposure of local or private information. Only explicitly approved pages enter the sanitized public subset. Automated validation and linting support this process, but consequential editorial and publication decisions remain under human control.

These procedures suggest that the significant transition from a personal LLM Wiki to a scholarly resource is not simply an increase in size. It introduces a governance problem: decisions are required about corpus membership, representational depth, provenance, synthesis, quality control and publication. These decisions are interdependent. Admitting a source creates an

¹ <https://drop-6520b376-555.rwsnapshot-2608.workers.dev/overviews/Sources-in-the-WoLaLa-Wiki>

editorial commitment different from merely possessing it; the choice of brief or deep ingestion affects how strongly it can contribute to later synthesis; integration may modify interpretations already represented elsewhere in the Wiki; and publication requires a further judgment about whether the resulting representation is sufficiently stable, attributable and appropriate for public use. Governance therefore operates throughout the knowledge-construction process rather than merely as a final publication filter.

4. Human-supervised knowledge construction

WoLaLa Wiki was built through a human-supervised agentic workflow involving a human researcher and two AI systems with different roles. ChatGPT served mainly as a discussion and supervisory partner, helping to formulate strategy, assess results and turn methodological decisions into operational instructions. Codex acted as the repository-level agent: it inspected the source collection and Wiki, carried out multi-step tasks, created and revised pages, maintained project records and ran validation procedures. The human researcher supplied domain expertise and scholarly judgment, set priorities and retained control over consequential decisions.

The systems did not communicate autonomously. The human mediated the process: plans were discussed with ChatGPT, instructions passed to Codex, and Codex's reports returned for assessment and further planning. This arrangement is not proposed as a novel agent architecture; rather, it documents the division of labour through which the scholarly resource was constructed.

Human supervision does not imply manual review of every operation or generated page. Many routine operations—including source processing, Wiki maintenance, integration and structural checking—are delegated to the agent, while human intervention is concentrated on decisions requiring scholarly judgment, such as ingestion depth, major changes to synthesis and public release. This selective allocation of human attention is important to the scalability of the workflow.

Automated validation currently focuses on structural integrity, metadata, provenance and publication safety. It does not systematically verify the factual fidelity of generated Wiki content against the underlying sources, and systematic evaluation of source fidelity therefore remains an important area for future work. Operational instructions, model information and project methodology will be documented with the research resource to support reproducibility.

5. Compounding in practice

A concrete example illustrates what cumulative synthesis means in WoLaLa Wiki. During selective ingestion of *A Companion to Chomsky*, four chapters concerning meaning and reference, semantics, pragmatics and intentionality were added to a Wiki that already contained concept, analysis and overview pages addressing meaning in language models.

Integration did not simply add four source summaries. It reorganized existing synthesis around these distinct evidential targets, clarified that evidence for semantic structure does not by itself establish reference or intentionality, and propagated these distinctions across existing pages. Thus new literature modified the accumulated representation on which subsequent work builds rather than merely enlarging the document collection. This was a human-supervised editorial process, not autonomous self-modification.

The same cumulative structure could eventually operate in the opposite direction. Research questions may expose missing connections, inadequately represented positions or gaps in the Wiki. We are developing Ask WoLaLa to investigate whether such findings, after human review, can lead to revisions of the persistent knowledge resource. This query-inspired enhancement remains future work.

6. Preliminary evaluation and conclusions

As a preliminary evaluation, we compared WoLaLa Wiki with Piantadosi’s expert synthesis (Piantadosi, 2023) of a substantially overlapping literature. From a researcher-assembled collection of 115 valid works cited by Piantadosi, WoLaLa Wiki had independently ingested 91 (79%). We compared their thematic organization rather than requiring agreement at the level of individual claims or editorial prominence. Both gave substantial attention to syntactic generalization, learnability, meaning and reference, language–thought dissociation, functional explanation and evidential standards. WoLaLa Wiki more explicitly distinguished behavioural, representational, mechanistic, developmental and neural evidence. Prominence did not align: works discussed extensively by Piantadosi were no more likely to receive deep treatment in WoLaLa Wiki.

We therefore interpret the result narrowly: it provides preliminary evidence that an independently constructed Wiki achieves broad coverage of a recognizable research landscape while organizing it according to partly different editorial priorities. It does not evaluate the factual fidelity of individual Wiki pages or establish expert-equivalent synthesis quality. Systematic evaluation of these dimensions is an important next step.

In conclusion, WoLaLa Wiki demonstrates that Karpathy’s personal LLM Wiki pattern can be developed into a governed scholarly knowledge resource spanning hundreds of sources. The main lessons concern not scale alone but the methods required for controlled corpus growth, differentiated representation, cumulative synthesis and publication. The project also suggests that scalability need not require eliminating human scholarly judgment: routine knowledge-processing and maintenance operations can be delegated to agents while human attention is concentrated on decisions where scholarly judgment is consequential. Community maintenance, portability to other domains, systematic content evaluation and query-inspired enhancement remain to be tested.

References

- Huerta, J. M. (2026) *WiCER: Wiki-memory Compile, Evaluate, Refine: Iterative Knowledge Compilation for LLM Wiki Systems*. arXiv:2605.07068.
- Karpathy, A. (2026) *LLM Wiki*. GitHub Gist, <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>
- Ming, H., Li, F., Wu, X. & Que, W. (2026) *Retrieval as Reasoning: Self-Evolving Agent-Native Retrieval via LLM-Wiki*. arXiv:2605.25480.
- Moreira, P. S. & Sweet, C. R. (2026) *Beyond Memory: A Templated Substrate for Heterogeneous Collaborative Knowledge Work with LLM Agents*. arXiv:2607.24759.
- Piantadosi, S. T. (2023) *Modern Language Models Refute Chomsky’s Approach to Language*. <https://zenodo.org/records/12665933>.
- Zhang, B. (2026) *Does the Wiki Grow Smarter? A Pilot Evaluation of LLM-Maintained Knowledge Bases in an Exception-Dense Decision-Support Task*. <https://www.sciexiv.com/article/sciexiv202605.0291.v1>.

Boglárka Vermeki

University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia,
vermekiboglarka@yahoo.com

Corpus-Grounded Retrieval-Augmented Generation for Beginner Hungarian: Comparing Pedagogical and Authentic Child-Language Grounding in Communicative Task Generation

Keywords: retrieval-augmented generation (RAG), beginner Hungarian language learning, child-language corpora, educational natural language processing (NLP), corpus-grounded text generation

I Introduction and background

Large language models (LLMs) are increasingly used to generate educational materials, but their predominantly adult-oriented training data may limit their suitability for beginner and child learners. Educational language technologies require not only grammatical fluency but also lexical simplicity, contextual appropriateness, communicative naturalness, and learner-oriented discourse patterns (Nayeem & Rafiei, 2024; Frank, 2011). This challenge is particularly relevant for morphologically rich and comparatively underrepresented languages such as Hungarian, for which educational NLP resources remain limited.

Retrieval-augmented generation (RAG) conditions LLM outputs on external material to improve contextual relevance and controllability (Gao et al., 2024). Because the quality and character of retrieved material can substantially influence generation (Chen et al., 2024), retrieval corpora may function as forms of linguistic grounding and not merely as sources of factual information. However, the pedagogical application of RAG to learner-oriented discourse remains underexplored.

Pedagogically simplified textbook corpora and authentic child-language corpora provide different forms of linguistic mediation. Textbook corpora prioritise lexical control, grammatical transparency, curricular progression, and beginner accessibility, whereas child-language corpora preserve conversational dynamics, discourse markers, emotional stance, and socially situated interaction. Naturalistic child-language data can capture communicative and context-sensitive properties often absent from standardised materials (van Dijk et al., 2023). This perspective also aligns with calls for smaller, interpretable, and socially grounded alternatives to weakly documented web-scale datasets (Bender et al., 2021).

The present study investigates corpus-grounded RAG for beginner Hungarian by comparing LLM-generated communicative learner texts under three conditions: (1) unguided base generation, (2) generation grounded in the pedagogically simplified *Magyar felfedező 1.* textbook corpus, and (3) generation grounded in the authentic KorSzak Child Language Corpus. A lightweight, prompt-based RAG framework retrieves relevant corpus excerpts as linguistic grounding during generation.

The experimental units are communicative learner tasks and short discourse-level texts. Drawing on task-based language-teaching research (Bygate, 2016; Ellis, 2003; Ellis et al., 2019; Long, 2015; Willis, 1996), the tasks include *introducing oneself or another person, presenting family members or favourite animals, creating a room tour, describing one's home, discussing weekend plans, and organising a shared programme.*

The study hypothesises that the two corpora provide complementary forms of grounding. Pedagogical grounding is expected to support grammatical transparency, beginner suitability, and instructional usefulness, while child-language grounding may contribute conversational naturalness, interactional authenticity, and child-oriented discourse patterns. Unguided generation provides the baseline against which these potential effects are evaluated.

2 Research Questions

The study investigates the following research questions:

1. How does corpus grounding affect the linguistic and pedagogical properties of LLM-generated beginner Hungarian learner texts?
2. Do pedagogically simplified and authentic child-language corpora produce different communicative and pedagogical effects in generated discourse?
3. Which grounding condition produces outputs perceived as more beginner-friendly, natural, pedagogically useful, and child-oriented?

3 Methodology

The study adopts a corpus-grounded retrieval-augmented generation framework using prompt-based contextual conditioning. Two Hungarian corpora representing distinct communicative and pedagogical registers are used as grounding sources: (1) Magyar felfedező 1. Corpus, a pedagogically simplified Hungarian textbook corpus designed for beginner language learners, and (2) an authentic Hungarian child-language corpus, the KorSzak Child Language Corpus, containing spontaneous child-child interactions and conversational speech.

The experiment focuses on eight communicative tasks for beginner-level Hungarian learners, selected from recurrent activity types in the textbook corpus. These include *self-introductions, introductions of a friend or family members, room tours, favourite-animal presentations, weekend plans, programme organisation, and short vlog-style narratives about everyday activities*. The tasks were manually selected based on pedagogical relevance, communicative centrality, and comparability across the two corpora.

Experiments were conducted using GPT-5.5 in a controlled prompting environment. For each of the eight communicative tasks, the model generated five outputs under three conditions: (1) unguided base generation, (2) pedagogical-corpus grounding, and (3) authentic child-language grounding, yielding a full generation set of 120 learner texts. In the grounded conditions, corpus material is segmented into meaningful discourse-level units, represented with source and task-relevant metadata, and retrieved by semantic similarity to the task prompt; the five highest-ranked relevant excerpts are inserted into the generation context. The model is instructed to use these excerpts as linguistic grounding for style, complexity, and communicative register while producing a novel beginner-friendly text.

Because full expert annotation of all 120 texts is still in progress, the present work reports a balanced pilot evaluation. For each task, two outputs were randomly selected from each condition (8 tasks × 3 conditions × 2 texts), producing a 48-text pilot sample. This design preserves representation of every task and grounding condition while enabling expert evaluation. Three annotators with backgrounds in Hungarian language pedagogy, applied linguistics, and corpus linguistics independently evaluated the 48 pilot texts. The annotators assessed the texts without access to their generation conditions. Texts were rated on a 1–5 scale for grammaticality, beginner-level suitability, pedagogical usefulness, naturalness, discourse coherence, and child-orientedness. The principal quantitative comparison reported here uses the five dimensions for which complete ratings from all three evaluators were available. Qualitative comments additionally identified recurrent linguistic, pedagogical, and discourse-level problems. This pilot analysis reports descriptive results across the available ratings and examines correspondence among the three evaluators. Annotation of the remaining texts in the full 120-text dataset is ongoing. The present analysis compares the three generation conditions within the completed 48-text pilot.

Pilot Results

The balanced pilot comprised 48 texts, with 16 outputs from each generation condition and equal representation of all eight communicative tasks. The generated texts were generally evaluated positively across the five dimensions for which complete ratings from all three

evaluators were available. Pedagogical-corpus grounding achieved the highest overall mean across the five analysed dimensions ($M = 4.60$), followed by authentic child-language grounding ($M = 4.52$) and unguided base generation ($M = 4.44$). The conditions showed partly complementary profiles. Pedagogically grounded outputs received the highest mean ratings for grammaticality ($M = 4.88$), beginner-level suitability ($M = 4.25$), pedagogical usefulness ($M = 4.71$), and naturalness ($M = 4.46$). Child-language-grounded outputs received the highest discourse-coherence rating ($M = 4.81$) and were rated as more natural than Base outputs ($M = 4.40$ versus 4.21), although they received the lowest beginner-suitability rating ($M = 4.08$). Base generation received the lowest overall mean and showed the greatest variation between individual texts. Pedagogical grounding produced the most consistent ratings across the pilot sample.

Inter-annotator reliability was examined using a two-way random-effects, absolute-agreement intraclass correlation coefficient. Single-rater reliability was moderate across the five dimensions ($ICC = .58-.66$), while the reliability of the mean ratings of the three evaluators was good ($ICC = .81-.85$). Exploratory task-blocked comparisons did not reach statistical significance (all $p \geq .197$). Given the small pilot sample and the concentration of ratings at the upper end of the scale, the observed condition differences should therefore be interpreted as descriptive tendencies. Communicative-task-fulfilment scores were not available in the exported evaluation summaries and were excluded from the present analysis. Only two evaluators completed child-orientedness, and it showed no variation, with all available ratings receiving the maximum score. Therefore, it was not included in the principal condition comparison.

Qualitative analysis further suggested that the effects of grounding were task-dependent. Child-language grounding performed especially well for *introducing a friend* and produced the highest mean discourse-coherence rating overall. Pedagogical grounding performed particularly well for *favourite-animal presentations*, *descriptions of home and housework*, and *programme-organisation dialogues*. The lowest-rated individual outputs included a child-language-grounded *family* text containing grammatical errors and a Base *programme-organisation dialogue* containing an internal temporal contradiction. Five of the eight lowest-rated outputs belonged to the Base condition.

These pilot findings partially support the proposed complementarity of pedagogical and authentic child-language grounding. Pedagogical grounding produced the highest and most consistent overall ratings, particularly for grammaticality and pedagogical usefulness. Child-language grounding contributed strong discourse coherence and improved naturalness relative to Base generation, but did not consistently improve strict beginner-level suitability. The results therefore suggest task-dependent benefits of corpus grounding. The pilot provides initial empirical indications that corpus selection can influence the pedagogical and communicative properties of generated beginner Hungarian discourse, while annotation of the remaining texts in the full 120-text dataset continues.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. <https://dl.acm.org/doi/10.1145/3442188.3445922>
- Bygate, M. (2016). Sources, developments and directions of task-based language teaching. *The Language Learning Journal*, 44(4), 381–400. <https://doi.org/10.1080/09571736.2015.1039566>
- Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence. arXiv. <https://arxiv.org/abs/2309.01431>
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.
- Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice*. Cambridge University Press.
- Frank, M. C. (2011). Computational models of early language acquisition. *WIREs Cognitive Science*, 2(5), 548–561. <https://osf.io/download/a95t2>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. <https://arxiv.org/abs/2312.10997>

- Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley-Blackwell.
- Nayeem, M. T., & Rafiei, D. (2024). KidLM: Advancing language models for children – Early insights and future directions. arXiv. <https://aclanthology.org/2024.emnlp-main.277/>
- van Dijk, B., van Duijn, M., Verberne, S., & Spruit, M. (2023). ChiSCor: A corpus of freely told fantasy stories by Dutch children for computational linguistics and cognitive science. Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), 352–363. <https://aclanthology.org/2023.conll-1.23/>
- Willis, J. (1996). *A framework for task-based learning*. Longman.

Claus Zinn
Erhard Hinrichs

University of Tübingen, Tübingen, Germany, {claus.zinn, erhard.hinrichs}@uni-tuebingen.de

Evaluating the Usage of LLMs for an Example-based Word Sense Disambiguation Task

Keywords: word sense disambiguation, interplay between large language models and computational lexicography, logistic regression for data analysis

I Introduction

Large language models (LLMs) can play a supportive role in lexicon development. One such application is the generation of example sentences to enrich lexical entries. Examples illustrate usage, they are particularly helpful to illustrate the paraphrase or definition of a word sense. In a past experiment, three LLMs generated example sentences for lexical entries in GermaNet, the largest lexical-semantic wordnet for German.

A wordnet is a network of word senses that are connected to each other by a set of conceptual relations (e.g., hypernymy) and lexical relations (e.g., antonymy). The first wordnet was developed at Princeton University by the research group of the cognitive psychologist George Miller (Fellbaum, 1998). In the wordnet literature, word senses are called lexical units. Lexical units that are synonymous or are near synonyms are grouped into synonymy sets, or synsets for short. In the Princeton wordnet each word sense is illustrated by a manually constructed example and accompanied by a definition of its meaning.

GermaNet (Hamp & Feldweg, 1997), the wordnet for German, has been under development at the University of Tübingen since 1993. It has since become the largest lexical-semantic wordnet for German (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010). Each of the 225,000+ lexical units (version 19), grouped into 174,579 synsets, has been manually constructed and curated by highly trained lexicographers. While their work is supported by specialised tools such as GernEdit (Henrich & Hinrichs, 2010), access to various German language corpora, newspaper collections, treebanks, and frequency databases, lexicographical work remains a labour-intensive manual work.

To support the GermaNet team, three large language models were instructed to generate example sentences that illustrate the meaning of GermaNet word senses in context. In total, 3600 examples sentences were generated; 900 sentences for 100 monosemous nouns, and 2700 sentences for polysemous nouns (300 word senses). All sentences were rated by two annotators in terms of acceptability following the criteria of (Atkins & Rundell, 2008). In our study, we use the corpus of rated example sentences for polysemous nouns for a task that turns the sentence generation task upside down.

2 Experimental design

We define a word sense disambiguation (WSD) task for LLMs on an LLM-generated corpus of example sentences: for a given lemma and its word senses, can they identify, for a given example sentence, the word sense used in the sentence? In this context, we would like to answer the following research questions: How do LLMs perform in the word sense disambiguation task, and which LLM is best? How does the degree of the lemma's polysemy influence the outcome of LLM performance in the WSD task? Is there a performance difference between acceptable and non-acceptable sentences?

2.1 LLM Selection

The corpus of LLM-generated example sentences was built using three LLMs: gpt-4o (OpenAI, 2024), deepseek (v3-0324) and claude-3-5-haiku-20241022. The latter two

LLMs are no longer available so that we used their successor systems `deepseek (v3.2)` (DeepSeek-AI, 2025) and `claude-haiku-4-5-20251001` (Anthropic, 2025). The LLM `gpt-4o` has been deprecated, but it still available for use.

2.2 Context and Prompt Design for the WSD Task

The role of the LLM is to correctly identify the word sense that is used in a given example sentence, which has been created to document this word sense in the best possible manner. In the actual prompt, we list all the word senses for a given lemma, and then ask the LLM to identify the word sense that is described by an example sentence. The LLMs have been assigned the following context on their role slot (all LLM interaction was conducted in German):

In der Vergangenheit wurden Beispielsätze erzeugt, um die typische Verwendung einer gegebenen Wortbedeutung bzw. Lesart möglichst gut zu dokumentieren. Im nachfolgenden Prompt gebe ich dir eine Liste von Wortbedeutungen. Deine Aufgabe ist es, einem gegebenen Beispielsatz die jeweils korrekte Lesart zuzuordnen. Deine Antwort muss ausschließlich die passende Wortbedeutung enthalten!¹

With the context, the following prompt was given (here, instantiated with 'Anzeige' and its three senses):

Das Wort 'Anzeige' hat drei Bedeutungen, die durch die folgenden drei Oberbegriffe umschrieben werden: 'Anklage', 'Ausschreibung' und 'Bauteil'. Nenne die Wortbedeutung, die im folgenden Beispielsatz verwendet wird: 'Die Anzeige am Thermometer fiel plötzlich unter den Gefrierpunkt'.²

The three LLMs mostly followed the prompt's directive. Occasionally, an LLM's answer was embedded in markup or preceded or followed-up by an explanation. In these cases, task success was rated manually.

2.3 Corpus of LLM-Generated Example Sentences

In a previous experiment, three LLMs generated three example for each of 100 monosemous lemmas and 300 polysemous word senses. The resulting 3600 example sentences were rated by two human annotators as either acceptable or non-acceptable. A sentence was judged non-acceptable if it had one of the following eight error types: wrong lemma (FL), wrong word sense (FA), grammar error (FG), orthographic error (FO), not natural (NN), not informative (NI), not typical (NT), and data hallucination (HA). In the sequel, we focus on the 2700 example sentences for polysemous nouns.

Each of the 2700 sentences were given to the three LLMs for the WSD task resulting to 8100 LLM responses on the WSD task. However, it would be unfair to judge LLM performance on sentences that received at least one FA rating from our human annotators. Those 371 sentences were found to elicit another word sense than the requested one, and hence were taken out from the study. For the subsequent main evaluation we are left with 2329 example sentences, and hence, 6987 LLM judgements on the WSD task from the three LLMs we used.

¹ I Role in English: In the past, example sentences were generated in order to document the typical usage of a given word meaning or sense as accurately as possible. In the following prompt, I will provide you with a list of word meanings. Your task is to assign the correct sense to a given example sentence. Your answer must contain exclusively the appropriate word meaning.

² Prompt in English: The word "Anzeige" has three meanings, which are described by the following three general terms: "criminal charge", "advertisement", and "component". Name the word meaning that is used in the following example sentence: 'The display on the thermometer below freezing.'

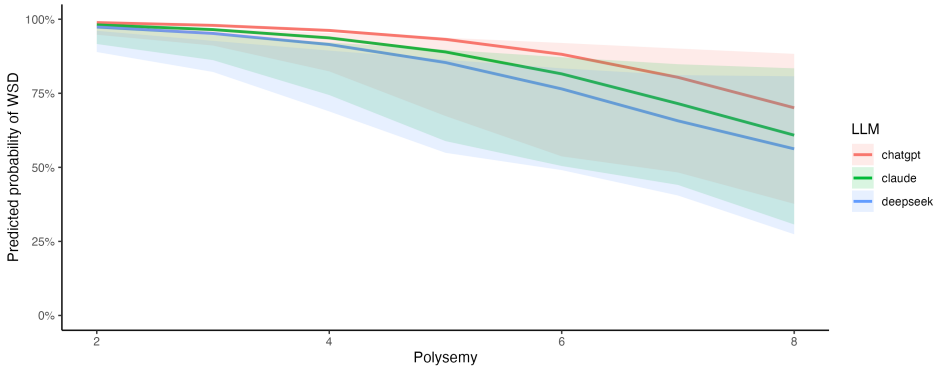


Figure 1: Both polysemy and LLM as main effects in the WSD task.

3 Evaluation

We start with an analysis of the 1567 example sentences that both annotators judged as acceptable. In this data, 122 different lemmas with varying degrees of polysemy were used:

number of senses	2	3	4	5	6	7	8	
number of lemmas	93	14	9	3	1	1	1	$\sum$ 122
number of wsd tasks	3051	732	465	171	54	108	120	$\sum$ 4701

A generalized linear mixed-effects model predicting WSD with polysemy and LLM as fixed effects and lemma as a random intercept

$$wsd \sim polysemy + llm + (1 | lemma)$$

showed a significant negative effect of polysemy ($\beta = -0.62$, $SE = 0.20$, $z = -3.07$, $p = .002$), corresponding to an odds ratio of 0.54. Thus, each one-unit increase in polysemy was associated with approximately 46% lower odds of WSD. LLM also significantly affected WSD: compared with the reference LLM `gpt`, `claude` showed lower odds of WSD ($\beta = -0.54$, $p = .007$, $OR = 0.58$), while `deepseek` showed still lower odds ($\beta = -0.87$, $p < .001$, $OR = 0.42$). Fig. 1 depicts the results.

It is remarkable that the probability of success in the WSD task surpasses 75% for highly polysemous lemmas across all three LLMs tested. Also, it is worth to note that the origin of the 2700 example sentences (i.e., the LLMs that generated them) was not statistically significant overall.

The previous results are for example sentences that both our annotators judged acceptable. To investigate whether the quality of the sentences has an effect on LLM's WSD performance, we divided the dataset of 2329 sentences into three acceptability conditions: (i) `acc_acc` (both annotators agreed that the example sentence is acceptable – 1567 sentences); (ii) `acc_nonacc` (one annotator found the example sentence acceptable while the other disagreed – 342 sentences); and (iii) `nonacc_nonacc` (both annotators agreed that the example sentence is not acceptable – 420 sentences).

A binomial logistic regression

$$wsd \sim acceptable$$

showed that acceptability significantly affected the likelihood of WSD. The estimated probability of WSD was .953 (95% CI [.946, .958]) in the `acc_acc` condition, compared with .916 (95% CI [.898, .932]) in the `acc_nonacc` condition and .933 (95% CI [.917, .945]) in the `nonacc_nonacc`

condition. Pairwise comparisons with Tukey adjustment showed that the odds of WSD were significantly higher in `acc_acc` than in `acc_nonacc` ($OR = 1.84$, $z = 4.61$, $p < .001$) and `nonacc_nonacc` ($OR = 1.45$, $z = 2.84$, $p = .013$). The difference between `acc_nonacc` and `nonacc_nonacc` was not significant ($OR = 0.79$, $z = -1.48$, $p = .302$). Fig. 3 depicts the results. It shows that LLMs perform remarkably well in the WSD task even for sentences our annotators found non-acceptable and that WSD performance even increases slightly in the `nonacc_nonacc` condition. To better understand this phenomenon, we looked into the error types of the `nonacc_nonacc` sentences. Here, both annotators agreed that the example sentence is non-acceptable, but they may have assigned different error codes to the sentences. To increase annotator agreement in this subset of sentences, we conflated NI, NN, and NT labeled sentences into the NX category.¹ Due to data sparsity, we also conflated the FO and FG error types into the Fx category. With this conflation, in the set of 1260 `nonacc_nonacc` sentences, annotators agreed on 1239 sentences and disagreed on 21 sentences.

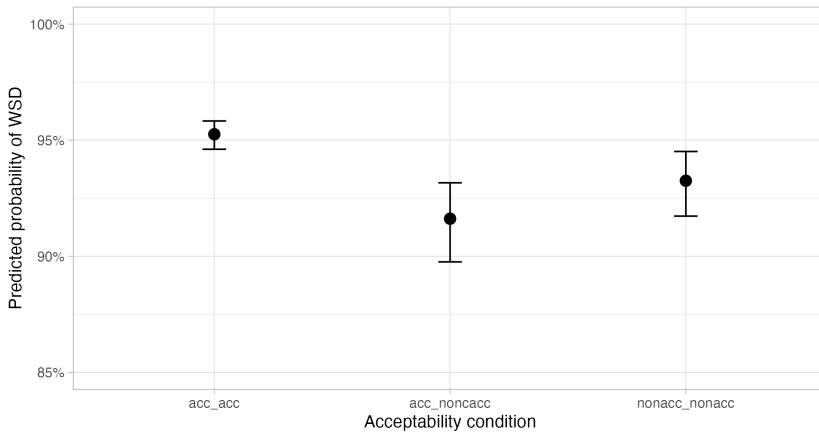


Figure 2: Sentence acceptability as factor in the WSD task.

A binomial logistic regression examining the conflated annotation categories showed that NX had significantly higher odds of WSD than FL ($\beta = 0.99$, $SE = 0.31$, $p = .001$, $OR = 2.70$). The difference between FX and FL was not statistically significant ($\beta = 1.09$, $SE = 0.73$, $p = .134$, $OR = 2.99$).²

4 Discussion

LLMs did very well in the WSD task within the corpus of acceptable sentences with success rates well above 90%. A binomial logistic regression revealed significant main effects of polysemy and the LLM used for the WSD task. Higher polysemy was associated with lower odds of a successful response. Our data shows that LLMs perform equally well across polysemy levels, that is, polysemy affected all LLMs similarly. We also showed that WSD performance does degrade in the corpus of less- or unacceptable example sentences, but not too much. It shows that WSD performance seems robust against NX flaws in the data — the lack of typicality, naturalness, informativeness did not deter LLMs to select the correct word sense used in the sentence. A significant performance drop can only be identified for example sentences that exhibit the FL (wrong lemma) flaw.

¹ In the main study, we found that annotators often assigned these labels but often disagreed on whether the sentence is not natural, not typical, or not informative.

² There were only 3 sentences marked HA. The estimate for HA was hence highly unstable, as indicated by its extremely large standard error, and was therefore not interpreted.

References

Anthropic. (2025). Claude 4.5 Haiku [Accessed: April-May 2026].
Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
DeepSeek-AI. (2025). DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models [Accessed: April-May 2026]. <https://arxiv.org/abs/2512.02556>
Fellbaum, C. (1998, May). *WordNet: An Electronic Lexical Database*. The MIT Press. <https://doi.org/10.7551/mitpress/7287.001.0001>
Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German [Madrid, Spain]. *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.
Henrich, V., & Hinrichs, E. W. (2010). GernEdiT - The GermaNet Editing Tool. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the International Conference on Language Resources and Evaluation*. ELRA.
OpenAI. (2024). GPT-4o system card [Accessed: April-May 2026]. arXiv preprint arXiv:2410.21276.

Appendix: Additional Analysis

Fig. 3 visualises the results of a generalized linear mixed-effects model predicting WSD with only polysemy as fixed effect (with lemma as a random intercept).

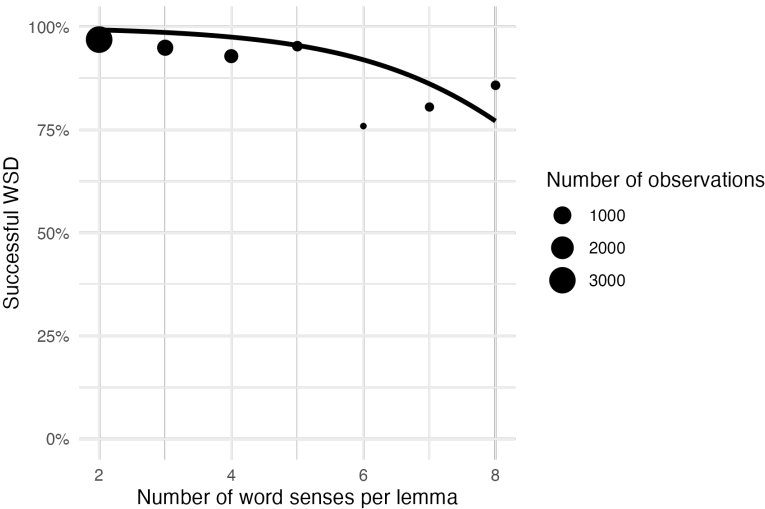


Figure 3: Polysemy as main effect in the WSD task.



The WoLaLa2026 Book of Abstracts is licensed under Creative Commons 4.0 CC-BY-ND license.

Conference Programme and Book of Abstracts
2nd International Workshop on Language and Language Models
WoLaLa2026

Dubrovnik, Croatia, 12 – 13 October 2026



ISBN 978-953-55375-6-4



9 789535 537564