LLM-Assisted Dialect Lexicography: Challenges and Opportunities in Processing Historical Bavarian Dialects

Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner, Katharina Korecky-Kröll

Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Bäckerstraße 13, 1010 Vienna, Austria

E-mail: philipp.stoeckle@oeaw.ac.at, daniel.elsner@oeaw.ac.at, wolfgang.koppensteiner@oeaw.ac.at, katharina.korecky-kroell@oeaw.ac.at

Abstract

This paper investigates the potential of LLMs in supporting lexicographic work on non-standard linguistic varieties using data from the *Dictionary of Bavarian Dialects in Austria (WBÖ)*. Based on approx. 2.4 million digitized and TEI-encoded dialect paper slips published via the Lexical Information System Austria (LIÖ), we construct a domain-specific corpus and evaluate LLMs in semantic classification and dictionary entry generation. Key preparatory steps include metadata enrichment, glossary and ontology development, and prompt engineering combined with Retrieval-Augmented Generation (RAG) techniques. Preliminary results suggest that LLMs can assist in organizing dialectal material into coherent semantic groupings. However, challenges persist regarding data preprocessing, structural conformity, and selection of representative examples. We discuss methodological implications and outline future directions, including the integration of agent-based systems and fine-tuning approaches tailored to dialect resources. This study contributes to the broader discourse on AI-assisted lexicography, highlighting both the potential and limitations of current LLM technologies in handling underrepresented language varieties.

Keywords: computational lexicography; historical dialect lexicography; large language models; metadata enrichment; semantic classification

1. Introduction

Since the public release of ChatGPT in late 2022, Large Language Models (LLMs) have become central to discussions around the future of language processing and linguistic analysis. Their impressive capabilities in handling standard language have sparked interest across multiple disciplines, including lexicography. However, the application of LLMs to dialect lexicography presents distinct challenges. Many dialects and (other) non-standard varieties are underrepresented – or entirely absent – from the training data of these models, limiting their ability to process such material effectively.

This paper explores the potential and limitations of LLMs in dialect lexicography using a case study from the *Dictionary of Bavarian Dialects in Austria (WBÖ)*, a long-term research project hosted at the Austrian Academy of Sciences. The WBÖ documents the lexis of historical Austrian dialects based on approx. 3 million handwritten paper slips, a large part of which have been digitized, encoded in XML/TEI, and published via the Lexical Information System Austria (LIÖ) since 2022. Our research focuses on how LLMs can assist in semantic classification and dictionary entry creation for dialect data, provided that the data sovereignty remains with the Austrian Academy of Sciences.

We first outline the research context (section 2), including the background and structure of the WBÖ, followed by a discussion of relevant state-of-the-art research on LLMs in lexicography, highlighting the specific challenges regarding the WBÖ data, the role of knowledge-enhanced neuronal networks, Retrieval-Augmented Generation, and prompt engineering techniques.

In section 3, we present our experimental setup, detailing the LLM environment, hardware, and software used, as well as our methodological approach to semantic classification based on XML/TEI-encoded dialect paper slips. We then describe the process of grouping lexical evidence into core meanings and present preliminary results from our experiments.

Finally, section 4 provides a discussion of our findings and their implications for future research, including agent-based pipelines and domain-specific fine-tuning approaches that aim to improve LLM support for dialect lexicography.

2. Research Context

2.1 The WBÖ

The WBÖ (Stöckle, 2021) represents one of the most comprehensive dialectological documentation projects in the German-speaking area. As a long-term lexicographic endeavor initiated in 1911, the WBÖ aims to systematically document the lexical variation across the dialects of historical Austria and South Tyrol. Given that most of the data originates from the first half of the 20th century, the material can be considered historically significant, capturing linguistic varieties that have since undergone substantial change or even disappeared.

The so-called "main catalog" – an archive of approx. 3 million paper slips – serves as the empirical foundation for all WBÖ research activities. The paper slips contain systematically categorizable information of lexicographic value, including the lemma (headword), geographic origin, semantic information, pronunciation details, and frequently illustrative sample sentences that demonstrate authentic usage contexts. Fig. 1 provides an example of a paper slip documenting the lemma *gäh* ('sudden, abrupt'), illustrating the structure and content of these primary sources.

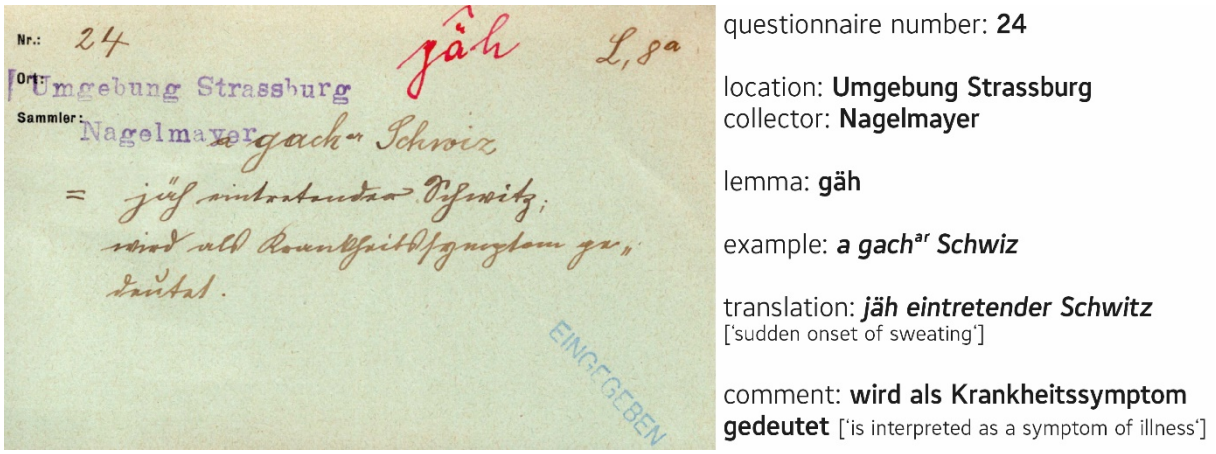

Fig. 1: Paper slip on the lemma *gäh* ('sudden, abrupt')

The paper slip contains questionnaire number, geographic attribution, collector identification, and a contextualizing example sentence supplemented by explanatory commentary. Notably, while explicit sense descriptions are sometimes absent, the semantic content can often be inferred from the provided usage examples, demonstrating the contextual richness of the documentation approach.

Recognizing the need to enhance accessibility and facilitate modern lexicographic workflows, digitization of the WBÖ collection began in the early 1990s, focusing initially on lemmas beginning with the letters *D* onwards, since the lemmas starting with *A*, *B/P* and *C* had already been integrated into the first three (printed) WBÖ volumes. The resulting digital database encompasses approx. 2.4 million entries, representing the vast majority of the original paper slip collection.

The digitization process initially employed TUSTEP (Tuebingen System of Text Processing Tools) for manual transcription of paper slip contents. However, responding to rapid technological developments and evolving digital humanities standards, the TUSTEP database was subsequently converted to XML/TEI format, ensuring long-term sustainability and interoperability with contemporary digital research infrastructures (Bowers & Stöckle, 2018). The XML/TEI conversion process involved not only technical transformation but also substantial content enrichment. Fig. 2 demonstrates the TEI representation of the paper slip shown in Fig. 1, illustrating how the digital format preserves all original information while adding valuable metadata enhancements such as part-of-speech classification, detailed geographic coding and specifications regarding the questionnaire number.

```
<entry xml:id="g297_qdb-d1e67011" n="371188" source="#orig-g297_qdb-d1e67011">
    <form type="hauptlemma" xml:id="tu-58545.15">
      <orth>gâh</orth>
    </form>
    <gramGrp>
      <pos>Adj</pos>
    </gramGrp>
    <note type="anmerkung" resp="O" corresp="this:BD">wird als Krankheitssymptom gedeutet</note>
    <cit type="kontext" n="1" xml:id="tu-58545.20">
      <quote xml:lang="bar">a <pRef>gachᵃʳ</pRef>
        <seg type="gram">[P,fl]</seg> Schwiz</quote>
      <def xml:lang="de">jäh eintretender "Schwitz"</def>
      <ref type="fragebogenNummer" xml:id="tu-58545.19">24L8a: Schwitz "Schweiß" m. Adj. (gesunder, kalter,*)</ref>
    </cit>
    <ref type="quelle" xml:id="tu-58545.16">Umg. Straßburg, Nagelmayer</ref>
    <ref type="quelleBearbeitet" xml:id="tu-58545.17">{2.3b05} obGurkt.:nöMKä. </ref>
    <ref type="bibl" corresp="this:QDB">
      <bibl>FbB.NAGELMAYER· (19xx) [SFb./WFb.2/Mtlg.]</bibl>
    </ref>
</entry>
```

Fig. 2: XML/TEI version (slightly shortened) of the paper slip on the lemma *gäh*

The enhanced digital format significantly improves research usability through implementation of Elasticsearch, enabling sophisticated querying capabilities and presenting data in accessible tabular formats (Breuer & Stöckle, 2023). Since 2022, the entire database is publicly accessible through LIÖ.

2.2 State of research

The public release of OpenAI's ChatGPT back in late 2022 had significant impact on lexicography as well, even though this scientific field has entered the phase of "post-editing lexicography" even before, as a considerable number of lexicographic tasks could already be processed semi-automatically (Rundell, 2023: 12–13). Soon after the seminal arrival of ChatGPT, evaluations started on how this tool was applicable in lexicographic work (De Schryver, 2023 for a comprehensive overview on such endeavors; see also Chen et al., 2024: 90; Rundell, 2023: 13).

This includes, e.g., a test of how ChatGPT performs in writing dictionary entries for a given set of (English) headwords: As Rundell (2023: 13–14) reports, single-sense words were more adequately defined than polysemous words in general, the latter especially lacking certain meanings, or including "pure inventions, unsupported by corpus data, and not recorded in mainstream dictionaries". Furthermore, example sentences for the respective dictionary entries given by ChatGPT were "consistently bad" (in the sense of 'inauthentic'), and the sources used for selecting them remained unclear (Rundell, 2023: 14; Jakubíček & Rundell, 2023). Rundell (2023: 16), thus, concludes: "ChatGPT can produce plausible-looking dictionary text, at least for headwords at the simpler end of

the spectrum. But closer examination almost always reveals problems".

Another study assessed "the quality of AI-generated sense definitions and example sentences in comparison to those crafted by human lexicographers for the same headwords" (Lew, 2023: 3), contrasting the entries generated by ChatGPT with the ones from Collins COBUILD Advanced Online. Four (human) lexicographic experts then evaluated a total of 15 entries qualitatively without knowing the respective entries' origin (Lew, 2023: 4). As a general result, the study concludes "that ChatGPT is capable of producing dictionary definitions emulating COBUILD style [i.e., non-condensed definitions; own addition] that are practically indistinguishable in quality from those written by highly trained human lexicographers" (Lew, 2023: 8). Regarding the quality of example sentences generated, results differ, though: They were evaluated as "significantly worse than those crafted by professional human lexicographers" (Lew, 2023: 9). This issue might be addressed by task-specific prompt engineering, as a follow-up session with improved instructions to the LLM and a re-evaluation of the – then refined – example sentences suggest (Lew, 2023: 9).

Due to pragmatic reasons, it is out of scope of this paper to highlight every study evaluating the usability of LLMs in lexicography. The two selected studies might act as exemplary showcases on how to tackle this topic and we will tie in here with our own approach (section 3). What follows is an aggregated summary on findings and implications based on individual studies that need to be taken into consideration when dealing with LLMs in (and for) lexicographic tasks.

First and foremost, the importance of well written prompts can hardly be overestimated: Proper prompt engineering is crucial for the quality of the output and (optional) fine-tuning of initial results (De Schryver, 2023: 377; Jakubíček & Rundell, 2023: 524; Lew, 2023: 9; Phoodai & Rikk, 2023: 348; Rundell, 2023: 15). Clearly, LLMs – chatbot ChatGPT based on OpenAI's GPT language models obviously was in focus of a great deal of the studies reviewed for this paper – seem to perform certain lexicographic tasks remarkably well such as, e.g., writing (dictionary entry) definitions (De Schryver, 2023: 371; Jakubíček & Rundell, 2023: 529; Lew, 2023: 8). The easy usability due to its natural language communicative style (Jakubíček & Rundell, 2023: 528), the cost-efficiency compared to human work load in general (Lew, 2023: 9), and the incredibly rapid enhancements of models with their vast increase of data used for training (Jakubíček & Rundell, 2023: 528; Rundell, 2023: 15) are further aspects worth noticing.

At the same time, the data used for model training is typically based to a considerable amount on English, which results in better performance in this language than on (rather) low resource ones (Li et al. 2024). Assuming that the datasets used for training predominantly include conceptually written texts in standard language varieties, the question of how and which non-standard varieties are actually represented in the datasets arises in this regard, too. In addition, the model input has repeatedly been

characterized as a "black box" (Chen et al. 2024: 97; De Schryver, 2023: 377; Rundell, 2023: 16), addressing the fact that the actual content of the dataset remains opaque as "the model [currently] does not keep references to training sources" making verbatim citations (of corpus content) practically impossible (Jakubíček & Rundell, 2023: 522). Another issue is the non-deterministic character of LLMs, resulting in different output at stable input, i.e., using the same prompt does not lead to the identical answer (De Schryver, 2023: 358; Chen et al. 2024: 97; Jakubíček & Rundell, 2023: 521–522; Phoodai & Rikk, 2023: 349; Rundell, 2023: 16). This, at the same time, "makes it hard to evaluate even a single prompt" (Jakubíček & Rundell, 2023: 524). In addition, LLMs might (tacitly) create facts that are completely invented, so-called "hallucinations", which is a fundamental problem in any task – not just lexicography – relying on factual correctness (Chen et al. 2024: 97; De Schryver, 2023: 358; McKean & Fitzgerald, 2023: 18; Rundell, 2023: 14). While some of these inventions might be easily detectable at once, others could require thorough verification. Apart from these (more) systemic issues, there might also be technical challenges due to the fact that the maximum token-length both for LLM input and output is restricted to a certain length for hardware reasons and that (readily) trained models are considered static insofar as the inclusion of further training data would lead to a complete retraining of the whole model as Jakubíček & Rundell (2023: 522) point out. Amongst other things, such changes might trigger the necessity to revise the by then (effectively) used prompting strategies (Jakubíček & Rundell, 2023: 524).

There are a number of specific challenges to unfold the potential of LLMs in supporting lexicographic work based on WBÖ data. This includes the challenge that WBÖ data relies to a great deal on German non-standard varieties (i.e., Bavarian dialect varieties), on which LLMs are currently not trained on specifically. Thus, the development of a rationale is crucial to ensure such training, both from the perspective of language (i.e., German) and variety (i.e., dialect). In addition, the dictionary entries of the WBÖ rely on accurate information provided by the lexicographic database, i.e., the paper slip catalogue, only. For reasons of lexicographic exactness, it, thus, is important that only those semantic definitions that are actually found in the data are included in the dictionary entry, or – put differently – that factually wrong or invented definitions are excluded from these LLM-generated dictionary entries. While we are currently focusing on semantic classification, a further step will include the linking of definitions with the corresponding paper slips used and to provide example sentences to illustrate the different definitions.

2.3 Knowledge enhanced Neuronal Networks

"Attention is All You Need" (Vaswani et al., 2017) – the paper published in 2017 knowingly revolutionized the research around neuronal networks with their groundbreaking "Transformer Architecture" that is used in state-of-the-art Language Models. Since then, a lot of research was published about how to train models with

factual knowledge besides the ability to write meaningful and well formatted text. Large-scale pretrained language models (PLM), also named Large Language Models (LLM) (Pan et al., 2023; Wei et al., 2021), are trained on several knowledge sources, namely, encyclopaedia knowledge, common-sense knowledge and linguistic knowledge on different levels of granularity. Text, entity, relations or subgraph are relevant in this regard for applications like text generation, entity-related tasks and questions answering for example (e.g., Pan et al., 2023).

Although LLMs have significantly improved over the last few years, they often lack the ability to bind the generated text to factual knowledge and are black boxes using opaque internal working mechanisms that are difficult to interpret (Pan et al., 2023; Wei et al., 2021; Baldazzi et al., 2023a, s. also 2.2). To address the problem of hallucinations and interpretability, researchers suggest synergizing factual knowledge with pretrained models. Two approaches are relevant for our research endeavour: a) knowledge enhanced LLM Pre-Training, b) knowledge enhanced LLM Inference (e.g., Wei et al., 2021; Baldazzi et al., 2023a). At the current stage of our research, our focus is on option b) while reserving option a) for potentially improving our results. Finetuning or training of existing models often leads to better results since during training a model can align new factual knowledge with linguistic context (Baldazzi et al., 2023a). However, it takes more resources to achieve and requires careful considerations in data preparation. On the other hand, inference with Retrieval-Augmented Generation (RAG) offers a more flexible approach that can be tested on several LLMs for comparisons.

Nevertheless, the tree-like DOM structure of XML in combination with TEI offers factual data nested in tags. The TEI tags hold ambiguous metadata and information about their relationship to ancestors, parent and children tags. With additional preprocessing like retrieving metadata from additional sources, all required components (factual data, metadata, relations and structure) are available for creating a text corpus by applying metadata enrichment, text linearization and verbalization methods. This corpus can be used for either fine-tuning or RAG.

2.4 Retrieval-Augmented Generation (RAG)

A workflow or pipeline for RAG requires a Retriever Algorithm to fetch data from a larger corpus. As the name already suggests, a generator mechanism is the next stage to complete the process. With an LLM as generator the retrieved inputs are then fused during inference and ensure the generated output is aligned with the given context (Gupta et al., 2024).

The result of this approach provides some answers to the problems of hallucinations, interpretability and data bias. However, it does not solve them. In many cases, models do not correctly incorporate the retrieved inputs in the generated texts as we will show in our preliminary results. A reduced data bias can only be achieved if the corpus attached to the retriever is carefully curated and addresses data bias. Otherwise, it may

amplify the existing model bias (Gupta et al., 2024). LLMs providing meaningful texts grounded in factual data, where data provenance can be evaluated are a promising prospect and could foster new linguistic research fields.

2.5 Prompt Engineering

The mastery of communicating with computers are prompts and the field of "Prompt Engineering" (s. also 2.2) has become a new discipline to communicate with LLMs. Especially the invention of chat bots like ChatGPT has created a lot of attention for the field. To improve the generated text created by LLMs a careful drafted prompt is often key to more accurate results. It is imperative to align the prompt to a correct terminology relevant for the topic and carefully informing the model about its role, task and desired output (e.g., Pan et al., 2023; Vatsal & Dubey, 2024).

In our research, Chain of Thought (CoT) prompt engineering has become a key role of improving results. The idea behind CoT is breaking down a complex task into smaller sub-tasks. Models finetuned with instruction-based data and reasoning capabilities are generating additional text during reasoning for each sub-task. The additional text will then be used as CoT template to complete the main task (Wei et al., 2022; Vatsal & Dubey, 2024).

3. Experiments

3.1 LLM, Hardware and Software Setup

Next to experimenting with LLMs in general, the aim of this research project was to use and test open-source models and compare them with proprietary solutions. Open-source models have the advantage that data sovereignty may remain within our research institution. However, working with open-source models comes with additional challenges since they must be self-hosted while proprietary solutions offer a ready-to-use model attached to a paywall. Solving these challenges will create much needed expertise in the field. This paywall, on the other hand, takes care of all the hardware and software requirements to run a model. Luckily, the strong open-source community for programmers in general but also within the machine learning and AI community publishes many very useful tools. One very prominent tool is Ollama[1]. For us relevant advantages of Ollama are:

- It can run model inference on CPU with smaller models

[1] Ollama is an installable client for Linux, Windows and Mac, running a server that is exposing an API endpoint for running model inference among other things. https://ollama.com/

- and run hybrid model inference on GPU and CPU if the model size exceeds the available GPU VRAM memory.
- It offers chat completion templates for various models.
- Most importantly, it can run the GGUF model file format.[2]

The GPT-Generated Unified Format (GGUF) is key for running large models that can have hundreds of gigabytes. Without diving deep into the specifics, it offers model quantization[3]. The community (ggml[4]) behind GGUF has published the llama.cpp library[5] used in Ollama.

Model	**Size (in billion parameters)**	**Availability**
LLaMa 3.1 Instruct	70B	open via Ollama or Huggingface[6]
LLaMa 3.2 Instruct	3B	open via Ollama or Huggingface
LLaMa 3.3 Instruct	70B	open via Ollama or Huggingface
Deepseek R1 – LLaMa based	70B	open via Ollama or Huggingface
Mistral	7B	open via Ollama or Huggingface
Gemma 3	27B	open via Ollama or Huggingface
Claude 3.7	unknown	proprietary via Anthropic
Claude 4	unknown	proprietary via Anthropic

Table 1: All tested LLMs

The first open-source LLMs we decided to use is LLaMa published by Meta. By comparison LLaMa provided the most promising results (s. 3.3) and therefore we discontinued using other open-sources models seen in table 1. According to their own published research it is the most capable open-source model able to compete with proprietary solutions like GPT3.5/4/4o, Mistral and Anthropic Claude 3.5. They offer

[2] File format description and availablility on Github:
https://github.com/ggml-org/ggml/blob/master/docs/gguf.md (3 June 2025)

[3] Huggingface article about quantization.
https://huggingface.co/docs/optimum/concept_guides/quantization (4 June 2025)

[4] Open-Source software community on Github. https://github.com/ggml-org

[5] Open-source C/C++ framework for running model quantization, finetuning and inference.
https://github.com/ggml-org/llama.cpp

[6] Online platform for hosting open-source machine learning models including LLMs.
https://huggingface.co/

405B (billion parameters), 70B, and 8B versions. The largest one provides the best results (Llama Team, AI @ Meta 2024). Without quantization this model would require more than 3 terabytes of VRAM only to load the model while the 4-bit quantized version only uses about 240 gigabytes. Even so, our limited hardware resources would only allow us to use the 70B version which comes in 43 gigabytes with 4bit quantization. For our first experiments, the hardware available to us is a Tesla V100 GPU with 32 GB.[7] The hybrid GPU, CPU model inference technique from llama.cpp via Ollama already comes in handy. It significantly reduces the inference speed but still handles generating output. However, the context window for RAG is limited to about 2048 input tokens, which is also the default setting in Ollama. The input tokens required for our project range from around 5k (thousand) up to the total limit of 128k Llama3 can handle. With this limitation, the results LLaMa3 provided still looked promising to continue our aim to use open-source models moving to the hardware offered by the CLIP[8] data center. With the publication of LLaMa3.3 in 70B[9], an instruction finetuned version of LLaMa3.1 offering similar evaluation statistics to the 3.1 405B version, we continued using the smaller version.

3.2 Methodological Approach

To introduce our domain specific factual data (s. 2.1) and enhance LLMs with this knowledge, a RAG Pipeline as seen in Fig. 3 is in development. The pipeline has three parts: a) the creation of a LLM text corpus, b) a Retriever and Algorithm to combine the context data with predefined prompts, and c) the LLM receiving all inputs and generating core meanings.

In the first part a), the data from sources like TEI/XML, Collection Cat, Glossary, and Ontology (s. 3.3) are processed. With text linearization and verbalization methods, structured data and metadata are reduced into text tokenizable for LLMs. The text corpus is stored in a Vector store database that holds token embeddings, the original text, and metadata about the text. In the second part b), the Vector store database is used as a Retriever. Currently each item in the database holds a complete collection of one lemma, e.g., *Gefrette* ('arduous work'; 'hardship'; 'misery'). The lemma *Gefrette* is added as metadata, and the Retriever retrieves the correct collection with a provided keyword. The retrieved inputs are combined with the predefined prompts (s. 3.3). The last part c) has two options. In option 1), the total inputs are sent to an API provided by a proprietary solution like Claude 4 for example, or option 2), to a self-hosted open source model like LLaMa3.3.

[7] GPU: Graphics Processing Unit, CPU: Central Processing Unit, VRAM: Video Random Access Memory, RAM: Random Access Memory

[8] Cloud Infrastructure Platform for High Performance Computing at the Austrian Academy of Sciences. https://www.clip.science/

[9] Benchmarks by Meta for LLaMa 3.3. https://www.llama.com/models/llama-3/#benchmarks

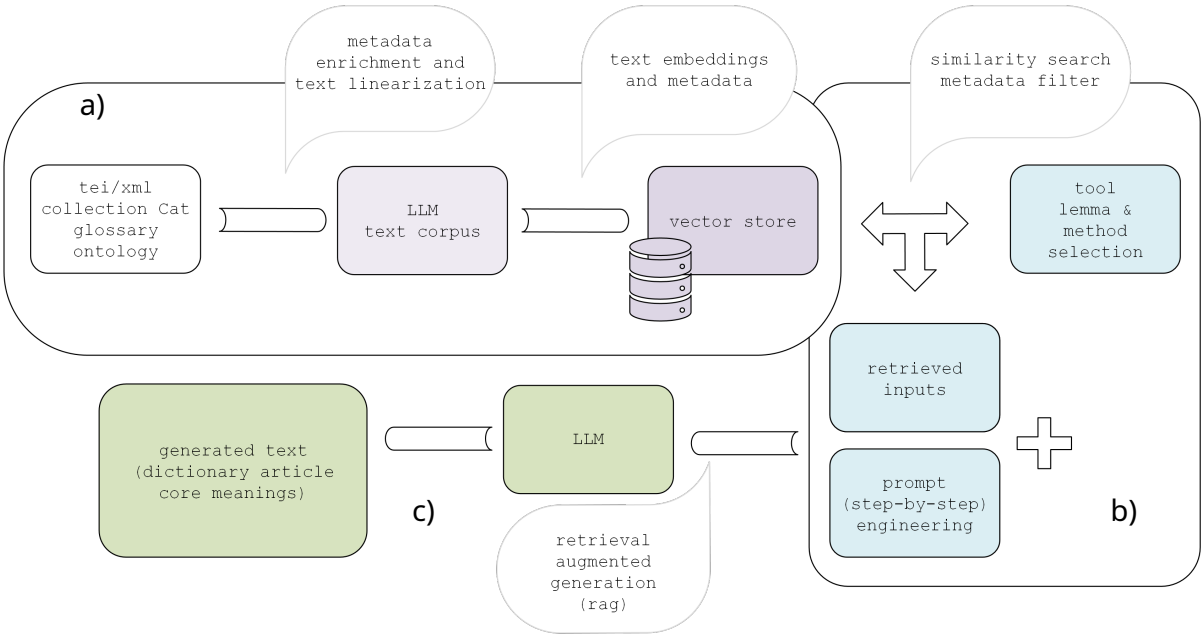

Fig. 3: LLM text corpus generation and RAG pipeline.

In part a) of Fig. 3, the creation of an LLM text corpus aims to introduce an ontology by creating KGs from the existing dataset and potentially finetuning LLaMa3.3 via Ontological Reasoning and KG enhanced pretraining (Pan et al. 2024; Baldazzi et al. 2023a, b). Currently the data is created only by combining TEI/XML source data, Collection Cat metadata and a glossary (s. 3.3).

3.3 Semantic Classification

To conduct the first round of testing, we selected a sample of 100 lemmas. For all of these, dictionary entries had already been authored by trained human lexicographers, allowing direct comparison between human- and machine-generated results. To examine potential effects of LLM pretraining on published material, we included 50 lemmas whose articles were already publicly available via LIÖ and 50 unpublished lemmas.

As outlined in 2.1, the basis for all lexicographic work in the context of the WBÖ is a structured XML/TEI database derived from approx. 2.4 million digitized historical paper slips. For practical lexicographic work, these TEI entries are accessed and annotated using a custom-built web-based tool called Collection Cat. This interface allows queries to the underlying BaseX database via Elasticsearch. While the layout is similar to the LIÖ online interface, Collection Cat is tailored to lexicographic workflows, offering features such as custom tagging of individual entries for semantic classification.

The TEI format preserves a wealth of metalinguistic information – e.g., lemma form, pronunciation, grammatical labels, sources – which is visualized in tabular form within the Collection Cat interface. However, as LLMs are not optimized to process this kind of structured, tagged data, their performance improves significantly when working with plain text. We therefore constructed a new corpus suitable for LLM input, transforming the structured XML data into readable text while preserving relevant content. To ensure consistent interpretation of category labels, we created a glossary defining all categories relevant to semantic classification, as well as an ontology specifying hierarchical and associative relationships between these categories. A separate text file was prepared for each lemma. For the experiment, we proceeded lemma by lemma, applying a standardized workflow. To reduce complexity in this initial stage, we limited the dataset to simplex forms, excluding compounds which would require tailored prompts and additional parsing.

The LLMs were provided with a text file for each lemma and instructed using a sequence of carefully designed prompts[10]. This stepwise prompting strategy follows the CoT approach, which has been shown to outperform direct prompting in tasks requiring reasoning and structured outputs (cf. Wei et al., 2022; Vatsal & Dubey, 2024). The stepwise method also mirrors the workflow of human lexicographers and was therefore considered particularly appropriate for our purposes. All prompts were formulated in German, in line with the language of both input and expected output.

The **first prompt** instructed the model to analyze the input data with respect to its lexicographic utility. The LLM was asked to proceed step by step, relying solely on the information in the text file, and to deliver the output in German. If the model did not support file uploads, the input text was pasted directly into the interface.

Despite the instruction to generate German output, two of the five LLMs tested – Mistral 7b and Gemma 3:27b – responded in English. Additionally, we observed frequent instances of hallucination, such as invented geographic attributions not present in the input. Among the open-source models tested, LLaMa 3.3 (70b) produced the most plausible results, but was still outperformed by Anthropic's commercial model Claude 4 Sonnet, which delivered more structured, detailed, and comprehensive analyses. Output by Claude typically included content overviews, structural summaries, grammatical annotations, word senses, and notes on geographic distribution.

The **second prompt** instructed the model to assume the role of a lexicographer and extract word senses, grouping them into coherent semantic categories. The model was asked to consider information from the following fields in order of priority: (1) word sense, (2) sense of sample sentence, (3) original (dialectal) sample sentence and (4) questionnaire number (s. Fig. 1). Other relevant information in additional fields could also be used. For each identified sense, the model was asked to specify regional usage

[10] For details on prompting and results see https://github.com/acdh-oeaw/llm-assisted-dialect-lexicography/.

and, where available, to include illustrative sample sentences along with the associated place of origin.

At this stage, only LlaMa 3.3 and Claude 4 Sonnet delivered usable results. Mistral 7b continued to output English texts, while Gemma 3 and DeepSeek r1 failed to return any output, even after extended waiting periods (sometimes exceeding several minutes).

Only the two successful models – LLaMa 3.3 and Claude 4 Sonnet – were carried forward to the **third prompt**, the final step. In this phase, the models were asked to compile full dictionary entries using a formal article scheme. The prompt provided the scheme in text form, describing the structural layout of the articles, including metalinguistic labels and expected content. All output should be derived solely from the data provided.

What follows is a detailed comparison of the entries generated by Llama 3.3 and Claude 4 Sonnet, focusing on their accuracy, structural coherence, and lexicographic adequacy. These outputs are evaluated alongside corresponding entries produced by a human lexicographer to assess their relative strengths and limitations.

3.4 Preliminary Results

As the current phase of the project is still focused on data preparation, the construction of an LLM-specific, and prompt engineering, preliminary results will be presented using the article *Gefrette* ('arduous work'; 'hardship'; 'misery') as an illustrative example. Fig. 4 displays the article as published in the WBÖ on the LIÖ platform. All WBÖ articles follow a standardized structure which, in addition to the lemma and its meanings, includes information on grammar and geographical distribution.

To evaluate the quality and assess the extent to which LLMs are capable of generating dialect dictionary articles in WBÖ-style, we examine the following aspects:

1) Structure: Are the expected components of the article present and correctly structured according to the WBÖ scheme (e.g., heading, grammatical information, senses with regional distribution, and example sentences with geographic provenance)?

2) Semantic classification: Is the categorization of senses coherent, comprehensible, and comparable to that of a human lexicographer[11]?

3) Data fidelity: Does the information provided originate from the underlying WBÖ data, or does the LLM introduce hallucinated content?

[11] Before being published, each WBÖ article passes three corrective steps to ensure scientific quality control.

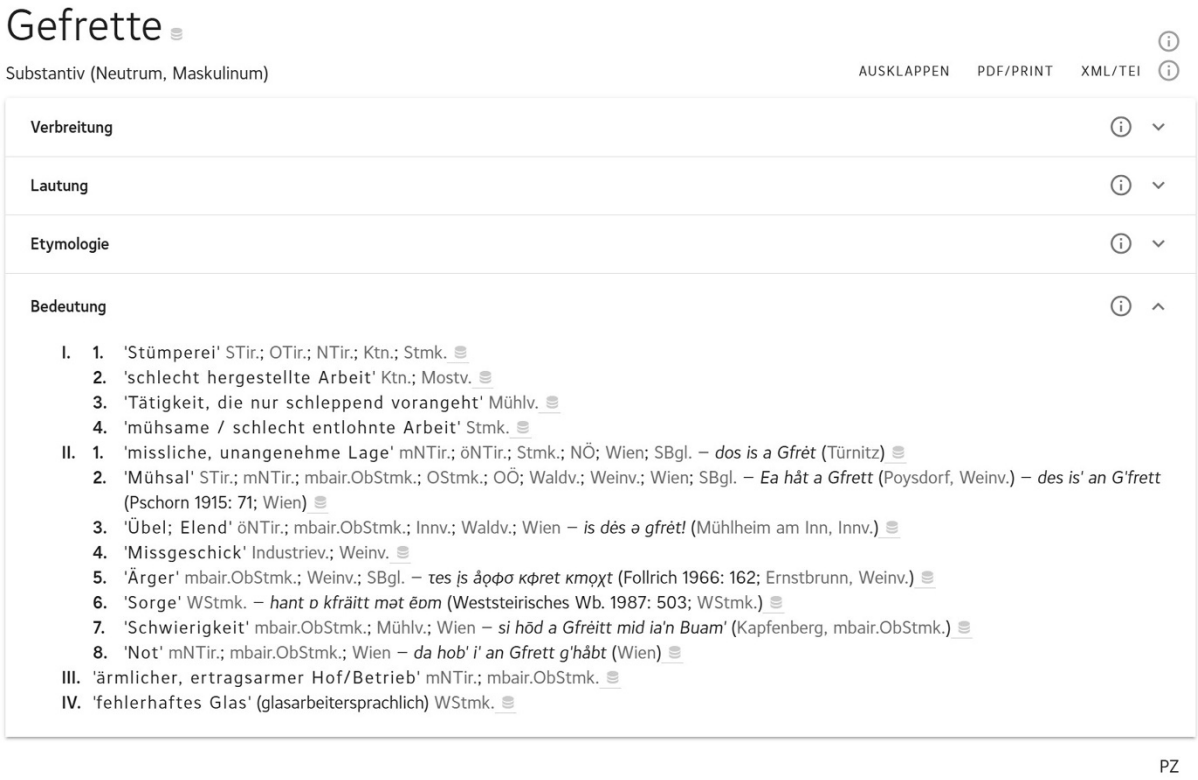

Fig. 4: WBÖ article *Gefrette* ('arduous work'; 'hardship'; 'misery') on LIÖ

There are several observations regarding structure of the dictionary entries generated for lemma *Gefrette*: In general, both models, Claude 4 Sonnet (henceforth Claude) and LLaMa 3.3 70B (henceforth LLaMA), structure their results corresponding to the input given, i.e., the scheme prescribed via prompt. This includes a header (in this case a denomination of the lemma *Gefrette*) with basic grammatical information. While both models indicate *Gefrette* as noun ('Substantiv'), only Claude provides additional information on the lemma's gender (here: neuter). According to information in the WBÖ article, *Gefrette* can also be interpreted as a masculine noun. However, this observation relies on a single record only, which might be the reason that neither model provides this option.

In addition, the AI-compiled semantic categories are presented in a structured way, which also includes information on the areal distribution of the respective categories. As is the case in WBÖ articles, the areal distribution orients towards defined Austrian major regions ('Großregionen'), which, thus, were also included in the glossary during data preparation (s. 3.3). To provide a regional overview per semantic category (as WBÖ articles do), Claude lists every major region found in the records (i.e., data entries based on the paper slips) assigned to the respective category. If there are example sentences given in the records, Claude uses a selection of them to illustrate the respective category's meaning, too. In addition, the model provides information on the regional origin of the selected example sentences, like in WBÖ articles, too. At the end of each category, Claude additionally lists all database IDs ('Beleg-ID') included in the

respective category. Like Claude, LLaMa also lists all major regions per semantic category. However, LLaMa additionally provides the (internal) numbers assigned to each major region. If the prompt was followed entirely correctly, the model should have excluded this information in its dictionary entry and relied on the names of the major regions only (as this is the way it is handled in WBÖ articles, too). LLaMa also adds example sentences for illustrative purposes to the respective categories, but unlike Claude, it does not provide regional information on the sentences' origin. LLaMa complements the example sentences by listing the respective database IDs, but – contrary to Claude – it does not list all database IDs that obviously were used to generate the respective semantic category.

Both Claude and LLaMa differentiate their semantic categories to a scheme of main- and subcategories similarly to the one used in WBÖ entries (i.e., Roman numerals, subclassified by Arabic ones, followed by Latin ones). Thus, both models follow the scheme outlined in the initial prompt. In total, Claude lists 15 (sub-)categories, LLaMa provides six. In comparison, the WBÖ article of *Gefrette* makes use of 14 (sub-)categories.

As already pointed out, both models provide example sentences for illustrative purposes upon availability in the respective records (which not always is the case). Up to two example sentences per semantic category are given, which corresponds to the way example sentences are provided in WBÖ articles. Both models reproduce the respective example sentences in their original form, i.e., in dialect, making use of the phonetic spelling in Teuthonista (if available). In LLaMa, however, the example sentences are quoted twice. Leaving this issue of double quotations aside, both models, in general, correspond to the instructions given in the respective prompt and are in line with how example sentences are handled in WBÖ articles such as, e.g., *Gefrette*. It is also worth mentioning that if there was additional grammatical information in the data (e.g., number and gender), this information also was included in the example sentences presented in the model's result.

We now turn to the second focus of evaluative remarks, dealing with (rather) qualitative aspects of the AI-generated semantic classifications. As already pointed out, the WBÖ article on *Gefrette* (s. Fig. 4) acts as an initial point of comparison. Based on human lexicographic work, four main semantic categories were deducted from the data (i.e., the respective paper slips collection). Two categories target the (broadly paraphrased) semantic fields "(arduous) work" ('[schlechte/stümperhafte/etc.] Arbeit'; cat. I.) and "hardship/misery/trouble" ('Mühsal/Notlage/Ärger/etc.'; cat. II.). These two main categories are subdivided into a number of subcategories to cope with semantic specifics inherent to the data (Fig. 4). The further two semantic categories are "poor, unprofitable farm/firm" ('armer, ertragsarmer Hof/Betrieb'; cat. III.) and "defective glass" ('fehlerhaftes Glas'; cat. IV). Both lack any subcategories and, thus, are reduced to a single semantic field each.

Results by Claude indicate diversified semantic categories, too. In total, there are seven main categories. Similarly to the WBÖ article, two categories deal with "hardship/trouble" (cat. I.) and "(arduous) work" (cat. II). Both categories are subdivided as well, corresponding with the categories' structure found in the human generated WBÖ article. The third category comprises "misfortune" ('Missgeschick, Pech'; cat. III.1) and "(constant) trouble" ('Ärger, dauernder Ärger'; cat. III.2). Both subcategories could also be assigned to cat. I., as there are no clear semantic boundaries justifying a separate, third category. Cat. IV includes the subcategories "issues in interpersonal relationships" ('Schwierigkeiten in zwischenmenschlichen Beziehungen'; cat. IV.1) and "calamity, difficult situation with/in economic problems" ('Kalamität, schwierige Situation mit wirtschaftlichen Problemen'; cat. IV.2). The semantic coherence of these two subcategories remains somehow opaque. A closer look at the respective records of these categories reveals that both categorical denominations are deducted from data of a single record each. Other records included in these two categories fit their semantic classification to a lesser extent. Claude's cat. V. semantically corresponds to cat. IV of the WBÖ article ("defective glass"). In addition, Claude generates a sixth category (cat. VI) with two subcategories (both based on a single record each). However, cat. VI.1, "running in an awkward position" ('ins Gefrette kommen: in missliche Umstände geraten'), could also be included in cat. I.2, "awkward, desperate, annoying position" ('unangenehme, verzweifelte, ärgerliche Lage'), and cat. VI.2, "existential experience of life" ('existentielle Lebenserfahrung'), would semantically fit into cat. I.1, "trouble, difficult situation" ('Mühe, Plage, schwierige Situation'), as well – at least, if we take the example sentence given in cat. VI.2 (literally "I do have my trouble in the world"; *I hô schă maî˜ gfréit ăv-dă wälld*) into account. The categorical denomination of cat. VI.2. seems suboptimal in this regard. So, both cat. IV and VI are exemplary instances of semantically defragmenting data stronger than necessary, or, put differently, creating more categories than necessary. In addition, results indicate that categories relying on single records might semantically fit into other categories as well (and possibly even better). Finally, Claude generates a seventh category as well, which is labeled as "ironic: abundance" (ironisch: Überfluss; cat. VII). It is based, again, on a single record. While the human lexicographer skipped this record due to its ambiguous content, Claude obviously deducted a distinct semantic sense.

As Claude includes more than one record in its semantic categories, let us take an impressionistic look at how well all these records align to their respective categories (i.e. its denomination / paraphrasing of their meaning). In general, if there is more than one record per category, there seems to be sufficient semantic (in-group) coherence. Sporadically, there are single records, though, that could have been assigned to another semantic category as well. This includes, for example, two records in cat. I.1 ("trouble, difficult situation"; database-IDs: f283_qdb-d1e1059, f283_qdb-d1e4051) that would fit cat. II ("[arduous] work") as well (and probably better).

LLaMa structures its results into main- and subcategories as well. Cat. I. consists of "trouble, difficulty" ('Mühe, Plage, Schwierigkeit'; cat. I.1) and "calamity, evil"

(‘Unheil, Übel’; cat. I.2), which corresponds to a certain extent to cat. I of the WBÖ article (“hardship/trouble”). As already pointed out, LLaMa does not list all database IDs used to generate the respective category in its result (i.e., the dictionary article). In most of the times, it only offers database IDs of the example sentences used, which makes it harder to evaluate how well the records fit the category (denomination) semantically. Only in categories, where no example sentences are found in the records, LLaMa provides up to two database IDs (this is the case for cat. III.1.b and III.2, s. below). One of the two example sentences of cat. I.1, *is dės ə gfrėt* [n]; das is a Gfret n, literally translated as “this is (an) evil”, seems to rather fit into cat. I.2. as well. LLaMa paraphrases cat. II., which is not subdivided into further subcategories, as “incompetence [or: amateur work], paltry/poor efforts” (‘Stümperei, armseliges Bemühen’). This corresponds in parts to cat. II of the WBÖ article, dealing with “(arduous) work)”. Cat. III in LLaMa is subdivided into three subcategories, two of which are subdivided into further, second level. Cat. III.1.a is labeled as “difficulty, annoyance” (‘Schwierigkeit, Schererei’). As “difficulty” is also used to paraphrase the semantic content in cat. I.1, this raises questions of precise inter-categorical semantic delimitation. Cat. III.1.b is paraphrased as “misfortune, unfortune situation” (‘Mißgeschick [sic!], unglückliche Situation‘). There are no example sentences available in this category. Nevertheless, LLaMa provides two database IDs: one refers semantically to “evil” (‘Übel’; database ID f283_qdb-d1e821), the other to “misery” (‘Not’; database ID f283_qdb-d1e2711). These two records do not really fit well into this category’s denomination. Finally, cat. III.2, “defective work, especially in the context of glass” (‘fehlerhafte Arbeit, besonders im Kontext von Glasgefäßen’) complements LLaMa’s third category. Based on the database IDs given in cat. III., an overarching semantic field is harder to detect than is the case for the previous two categories.

This leads to the question of data fidelity: Does the information provided originate from the underlying WBÖ data, or does the LLM introduce hallucinated content? A fundamental aspect of lexicographic work in the context of WBÖ is linking (illustrative) examples (typically in the form of example sentences) to the respective semantic categories generated[12], and to provide (major) regional information tied to this data. Based on the state of research (s. 2.2), there are obviously challenges of “citing” the actual/correct data used when working with LLMs. Regarding this paper and its focal points of evaluation, this includes the questions a) whether the LLM provides adequate example sentences per category and b) in how far the provided (major) regional information provided with the respective categories and example sentences is factually correct.

Regarding Claude, there is rather high congruence of the record used (presented in the results by indicating the respective database-ID of the record), example sentence found in the record (as far as there is one, of course) and the indication of the major region of

[12] All records used for writing the respective WBÖ article are clickable at the LIÖ platform, irrespective of their use as example sentence.

both, record and example sentence (upon availability). As previously stated above under structure, Claude does not only provide the major region per example sentence, but it also additionally lists all major regions found in the records of the respective semantic category directly after the category's denomination – as a kind of (major) regional summary and in line with the way, data is presented in WBÖ articles. There is a small number of instances where the provided major regional information of this kind is incorrect, though. In two subcategories generated by Claude, there are major regions listed which are not found in the corresponding records (based on the database IDs given in the results) of this category: This is the case for cat. I.2, where "Western Styria" ('Weststeiermark') and "Waldviertel" ('Waldviertel') were incorrectly listed (while the further regions "Vienna" / 'Wien' and "Weinviertel" / 'Weinviertel' were correct), and in cat. II.1, where "Central-Bavarian Upper Styria" ('mittelbairische Obersteiermark') was not found in the records (while "Mostviertel" / 'Mostviertel' and "Mühlviertel" / 'Mühlviertel were correctly listed). Conversely, in cat. II.2 "Vienna" ('Wien') should have been listed alongside the (correctly listed) "Central Northern Tyrol" ('mittleres Nordtirol') and "Waldviertel" ('Waldviertel') - but it was not, irrespective of the fact that the example sentence given to illustrate the meaning of cat. II.2 indicated "Vienna".

As already mentioned above, for LLaMa, an evaluation of the congruence of the records used per category is more difficult as the model provides less database IDs of the records used to generate the semantic categories in its results, i.e., the dictionary entry. As a consequence, the generated list of major regions per category cannot be evaluated in terms of correctness, as there are not enough database IDs to do so. Put differently: The number of major regions per category exceeds the given database IDs in the majority of categories. However, we can compare whether the major regions found in the available database IDs are also present in the list of major regions per category. This is not always the case: In cat. I.2, the database ID of one of the example sentences refers to "Mostviertel" ('Mostviertel'; database ID f283_qdb-d1e2864), but this region is not found in the list of major regions of this category. Same goes for cat. II, which incorrectly excludes "Middle Carinthia" ('Mittelkärnten'; database ID f283_qdb-d1e2687) from the results. In cat. III.1.b, "Middle Northern Tyrol" ('mittleres Nordtirol'; database ID f283_qdb-d1e2711) is missing as well. The regional information of the other five database IDs given in the dictionary article is correctly reflected in the list of major regions of the respective categories, though (in one case, there is no major region available in the data).

Apart from these issues, blatant invented or hallucinated content could not be found in both models' results on *Gefrette*. This result was not necessarily expected, as earlier tests (using, amongst other, less diversified prompt engineering) did include clearly hallucinated (such as, e.g., major regions in Germany for which there is no WBÖ data). In addition, it is worth mentioning that both models seemed to refrain from adding data on occasions where there was no data available in the records. If, for example, there was no regional information found, or no example sentence given in the respective record,

the models did not add such data to the results presented – at least what can be said based on the database IDs provided in the results.

As a final, yet remarkable result, Claude largely uses the same example sentences as found in the WBÖ article: Six out of a total of eight given example sentences (all categories combined) can be found in the WBÖ article as well. In LLaMa, this value amounts to one (out of six). The WBÖ article was published online in October 2020. It remains unclear whether it was included in any training data of the models used, though. We will be able to tackle this question in the forthcoming evaluations, though, which will include both published and unpublished lemmas, as outlined in 3.3.

4. Discussion and Outlook

4.1 Resources, Sustainability and Fine-Tuning Strategies as alternative to RAG

Whether or not fine-tuning an LLM is necessary for this project depends on the evaluation of the results based on the RAG based text generation. The discussed LLMs already went through unsupervised pre-training and fine-tuning steps to achieve a specific usage such as chat completions (system – assistant – user) by showing the model a selection of curated chat completion examples. Or summarization capabilities with curated examples of text and summarized text combinations using SFT data (e.g., Touvron et al. 2023; Meta-AI 2024). Next to training LLMs on specifics use-cases and data the size of the context e.g. tokens for RAG during inference is an issue. As mentioned in 3.1 the LLaMa context token window is 128k and according to Anthropic their newest models can process up to 200k during inference. Most of our data collections will fit inside these limitations while some collections are exceeding. Fine-tuning could solve this problem.

Tests on CLIP have shown that LLaMa 3.3 (70B) with 128k token context window uses up to 83 gigabytes of GPU VRAM memory; about 43 gigabytes for loading the model and 40 gigabytes for the context. This requires three of the available Tesla A100 GPUs each providing 40 gigabytes of VRAM memory or two GPUs by moving the remaining required memory to the regular CPU RAM. The number of resources required for fine-tuning has yet to be determined but discussed libraries like llama.cpp offer fine-tuning strategies for quantized models and a similar hardware setup should be feasible.

What speaks against fine-tuning is research showing that even fine-tuned systems still struggle with data provenance and hallucinations let alone interpretability. Their research suggests using few- or one-shot RAG to improve the results on fine-tuned systems. On the other hand, the research offers insights in how to potentially solve interpretability of LLM generated texts by training a model on “ontological reasoning”

thus allowing questions like “why this conclusion?” (Baldazzi et al., 2023a, 2023b). The training dataset they suggest is based on network-based data modelling with Enterprise Knowledge Graphs (EKGs). This dataset consists of a general KG with factual data enhanced with ontological information of a specific domain or enterprise. In our project this would mean creating a KG enhanced with ontological information about semantics and lexicography.

Finally, are LLMs sustainable? This project has not yet collected data about power consumption and the answer is connected to the evaluation results and whether LLMs can efficiently assist in dialect lexicography. The elapsed time during the latest inference round with LLaMa3.3 was approx. 32 minutes with a context token size of around 10k. The above-mentioned hardware requirements already outline the rather resource intensive nature of using AI tools like LLMs. Research estimates that by 2030 AI technologies might contribute to about 5% of the global CO2 emissions raising concerns and questions to be addressed when conducting research on e.g., LLMs (Singh et al. 2025).

4.2 Lessons learned and next steps

There are several conclusions we can draw from the current state of our research and the model inference tests we conducted. We can confirm that RAG in combination with CoT prompting significantly improves the model generated text. Furthermore, the latest tests as discussed in 3.4 did not show obvious hallucinated facts. It makes a great difference if a model is prompted with one prompt that holds all the instructions – even if it is carefully formulated and broken down into several steps – or whether the model is prompted several times, with every new prompt increasing the context data by appending the model's previous output to the next prompt. Additionally, newer models offer “extended thinking”[13]. These models simulate thinking out loud during inference before generating the final output. Models like Claude 4 already include this feature and we believe that it is one of the reasons why it creates much better results than LLaMa 3.3 which does not use “extended thinking”. In our future tests we will therefore try to adapt new model versions like LLaMa 4 to level the playing field. There is one question that arises, however: Why not go the easier way and just use proprietary solutions? And the answer is simple: With self-hosted open-source models we keep data sovereignty. Additionally, finetuning models on a larger non-standard language dataset is only possible with an open-weight model like LLaMa 3.3. Finally, the open-source community might be key to developing new solutions to make AI more accessible by creating solutions like quantization as discussed in 3.1 and potentially reducing the carbon footprint.

[13] Extended thinking is a step-by-step reasoning feature enabled by additional superivised finetuning data (SFT) see https://docs.anthropic.com/en/docs/build-with-claude/extended-thinking.

What remains open is the question of interpretability. As discussed in 3.4, while hallucinated facts do not seem to be an issue, it is still unclear how the models conclude the core meanings. Hence, raising the question "why this answer?". As outlined in 4.2 EKGs used for finetuning LLMs on reasoning could help improve interpretability. Therefore, we aim to create the described EKG of our data and use it for finetuning LLaMa 3.3 or one of its future versions. Whether our strategy leads to efficiency gains by assisting in writing dictionary entries remains to be seen. Once we have concluded the data preparations, prompt engineering strategy and finetuning strategy, we will be able to clarify this question. As discussed in 4.1, the lemma *Gefrette* requires about 32 minutes of inference time. A lexicographer might have used several working days to carefully investigate the data and come to a conclusion. If we can further improve the results and interpretability of the model outputs and remain within similar timeframes, we believe LLMs could potentially make research projects such as the WBÖ more efficient.

5. References

Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A., Sallinger, E. & Atzeni, P. (2023a). Explaining Enterprise Knowledge Graphs with Large Language Models and Ontological Reasoning. Available at: https://doi.org/10.4230/OASIcs.Tannen.1 (18 July 2025).

Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A., Sallinger, E. (2023b). Fine-tuning Large Enterprise Language Models via Ontological Reasoning. Available at: https://doi.org/10.48550/arXiv.2306.10723 (18 July 2025)

Bowers, J. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A.U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.) *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2), 25–26 January 2018 Vienna, Austria*. Wien: Gerastree Proceedings, pp. 45–54. Available at: https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf (18 July 2025)

Breuer, L.M. & Stöckle, P. (2023). Das WBÖ-online im ‚Lexikalischen Informationssystem Österreich' – Zugriff und Vernetzungsmöglichkeiten, Version 2. In T. Krefeld, S. Lücke & C. Mutter (eds.) *Berichte aus der digitalen Geolinguistik (II): Vernetzung und Nachhaltigkeit (Korpus im Text 9), Version 30*. Accessed at: https://www.kit.gwi.uni-muenchen.de/?p=54448&v=2. (18 July 2025)

Chen, L., Dao, H.-L. & Do-Hurinville, D.-T. (2024). AI empowerment: Where are we in the automation of lexiocography? A metaphraseographic study. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *ASIALEX 2024 Proceedings*, Sep 2024, Tokyo, Japan, pp. 90–98.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The current State of Art Using ChatGPT. *International Journal of Lexicography* 36(4), pp. 355–387.

Elasticsearch. Accessed at: https://www.elastic.co/elasticsearch (18 July 2025)

Gupta, S., Ranjan, R., Narayan Singh, S. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. Available at: https://doi.org/10.48550/arXiv.2410.12837. (18 July 2025)

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (eds.) *eLex 2023. Electronic lexicography in the 21st century*, Lexical Computing CZ: Brno, pp. 518–532.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities & Social Sciences Communications* 10(704), pp. 1–10.

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., Du, M. (2024). Language Ranker: A Metric for Quantifying LLM Performance Across High and Low-Resource Languages. Available at: https://doi.org/10.48550/arXiv.2404.11553. (18 July 2025)

LIÖ: *Lexikalisches Informationssystem Österreich* ('Lexical Information System Austria'). Accessed at: https://lioe.dioe.at/. (18 July 2025)

McKean, E. & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In*: AsiaLex 2023. Lexicography, Artificial Intelligence and Dictionary Users*, pp. 18–27.

Meta Llama Model 3 = Meta AI: Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed at: https://ai.meta.com/blog/meta-llama-3/. (18 July 2025)

Llama Team @ Meta: The Llama 3 Herd of Models. Available at: https://doi.org/10.48550/arXiv.2407.21783. (18 July 2025)

Pan S., Luo L., Wang Y., Chen C., Wang J., Wu X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. Available at: https://doi.org/10.48550/arXiv.2306.08302. (18 July 2025)

Phoodai, C. & Rikk, R. (2023). Exploring Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (eds.) *eLex 2023. Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, Lexical Computing CZ: Brno, pp. 345–375.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *AsiaLex 2023. Lexicography, Artificial Intelligence and Dictionary Users*, pp. 9–17.

Stöckle, P. (2021). Wörterbuch der Bairischen Mundarten in Österreich (WBÖ). In A. N. Lenz & P. Stöckle (eds.) *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Stuttgart: Steiner, pp. 11–46. Available at: https://doi.org/10.25162/9783515129206. (18 July 2025)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., 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.

Available at: https://doi.org/10.48550/arXiv.2302.13971. (18 July 2025)
TUSTEP: *Tuebingen System of Text Processing tools*. Accessed at: https://www.tustep.uni-tuebingen.de/tustep_eng.html. (18 July 2025)
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N.G., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. Available at: https://doi.org/10.48550/arXiv.1706.03762. (18 July 2025)
Vatsal, S. & Dubey, H. (2024). A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. Available at: https://doi.org/10.48550/arXiv.2407.12994. (18 July 2025)
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Available at: https://doi.org/10.48550/arXiv.2201.11903. (18 July 2025)
Wei, X., Wang, S., Zhang, D., Bhatia, P. & Arnold, A. (2021). Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey. Available at: https://doi.org/10.48550/arXiv.2110.08455. (18 July 2025)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/