Improving Slovene Language Models for Lexicographic Question Answering through Continued Pretraining and Instruction Fine-Tuning

Timotej Knez, Slavko Žitnik
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
{timotej.knez, slavko.zitnik}@fri.uni-lj.si

Abstract

This paper presents a two-stage training approach to improve the performance of Slovene large language models on lexicographic question-answering tasks. We developed a comprehensive lexical pretraining corpus containing 356,294 Slovene word entries. We constructed the corpus by converting structured data from multiple lexicographic sources into markdown format. Additionally, we created a question-answering dataset with 16,508 QA pairs from diverse sources, including automatically generated questions, a linguistic advisory portal, and community forums. Using the Slovenian GaMS model (based on Gemma 2 9B) and GaMS 3 model (based on Gemma 3 12B), we performed continued pretraining on the lexical corpus followed by instruction fine-tuning with our QA dataset combined with translated general-domain questions. We compared results to different model configurations. Our results demonstrate significant improvements (text similarity increasing from 0.226 to 0.542, BERTScore F1 of 0.915) in answering Slovene lexicographic questions, validating the effectiveness of domain-specific continued pretraining for low-resource languages.

Keywords: Slovene, language models, lexicography, continued pretraining, instruction fine-tuning, question answering

1. Introduction

Lexicographic knowledge plays a fundamental role in natural language processing, providing essential information about word meanings, morphological forms, usage patterns, and semantic relationships (Horák and Rambousek, 2017). While large language models (LLMs) have demonstrated remarkable capabilities across various tasks, they often struggle with specialized lexicographic questions, particularly for low-resource languages where pretraining data may be limited and less diverse (Merx et al., 2024). For Slovene, a South Slavic language with complex morphology and relatively scarce digital resources, this challenge is especially pronounced.

The gap between general pretraining corpora and domain-specific lexicographic knowledge presents a significant obstacle. Standard pretraining approaches expose models to broad linguistic patterns but may not adequately capture the structured, detailed information contained in lexicographic databases. This is particularly problematic for questions that require precise knowledge of word forms, definitions, collocations, synonyms, and morphological paradigms, information that is systematically organized in lexicographic resources but may be underrepresented or scattered across general text corpora.

This paper addresses a research question: Can structured lexicographic data improve Slovene LLM performance through a combination of continued pretraining and instruction fine-tuning? We hypothesize that converting structured lexicographic resources into natural language text and using them for continued pretraining will inject domain-specific knowledge into the model, while subsequent instruction fine-tuning will align the model with question-answering tasks.

In summary, our experiments show that lexicographic adaptation is highly effective: direct lexicographic instruction fine-tuning (Lex-FT) produced the best overall results, substantially improving Slovene lexicographic QA quality over baseline models, while continued pretraining provided limited additional gains and general benchmark performance remained broadly stable.

Our main contributions are:

- A lexical pretraining corpus containing 356,294 Slovene word entries synthesized from multiple lexicographic sources and structured in markdown format
- A comprehensive QA dataset with 16,508 question-answer pairs from diverse sources, including automatically generated questions, linguistic advisory portals, and community forums
- A two-stage training methodology combining continued pretraining with instruction fine-tuning, applied to the GaMS model (based on Gemma 2 9B architecture)
- Comprehensive evaluation demonstrating significant improvements in answering Slovene

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 114–123

lexicographic questions while maintaining general language capabilities

2. Related Work

2.1. Slovene Lexical Resources

Slovene benefits from a comprehensive ecosystem of digitized lexical resources developed primarily by the Center for Language Resources and Technologies (CJVT) at the University of Ljubljana. These resources collectively provide definitions, usage examples, word forms, and collocations that form the foundation of our lexicographic knowledge base.

The Digital Dictionary Database of Slovene (DDDS) serves as the central lexicographic repository. Linked to DDDS, the Open Slovene WordNet (OSWN) provides a semantic network of Slovene synsets: OSWN 1.0 contains approximately 95,262 synsets covering 164,904 word forms, derived from Princeton WordNet with Slovene translations (Čibej et al., 2023). CJVT also maintains the Thesaurus of Modern Slovene, the largest open Slovene synonyms collection, which is fully digital and crowdsourced (Krek et al., 2023). This thesaurus interlinks with other resources and enables users to compare headwords and synonyms along collocational profiles with corpus examples.

For contextual information, the Collocations Dictionary of Modern Slovene provides 78,046 headwords with 4.4 million collocation pairs and approximately 14.45 million usage examples (Kosem et al., 2023). For morphology, Sloleks 2.0 (Slovene Morphological Lexicon) covers 100,802 lemmas and about 2.79 million inflected word forms with automatic stress annotation (Dobrovoljc et al., 2019). Together, these resources provide nearly all facets of a dictionary entry: sense definitions, usage examples, inflectional paradigms, and collocational patterns, making them good knowledge base for language model enhancement.

2.2. Slovene Large Language Models

Recently, CJVT has developed Slovene-specific LLMs. The GaMS family ("Generative Model for Slovene") includes models continually pretrained on Slovene and related South Slavic data (Vreš et al., 2024). Vreš et al. (2024) created GaMS-1B by continuing pretraining on the English OPT model with Slovene data. Larger models followed: GaMS-9B (9B parameters) and GaMS-27B (27B parameters) are based on Google's Gemma 2 multilingual architecture, continually pretrained on Slovene plus Croatian, Serbian, Bosnian, and English (CJVT, 2024). More recently, CJVT released GaMS3-12B-Instruct, the next-generation GaMS based on Google's Gemma 3 architecture (CJVT, 2026). These models serve as powerful Slovene backbones for further fine-tuning. In this work, we use GaMS models as our base to specialize in lexicographic data.

2.3. LLMs for Lexicography and Lexical Semantics

There is growing interest in using LLMs for lexicographic tasks. Pham et al. (2025) found that ChatGPT's word definitions are highly accurate, comparable to traditional dictionaries. Similarly, Lew (2023) showed that ChatGPT could generate COBUILD-style dictionary entries for English, with AI-generated sense definitions rated on par with human lexicographers' definitions, though example sentences scored lower. These results demonstrate LLMs' raw capacity for lexicographic output but also their limitations, particularly in producing illustrative examples. Importantly, both studies suggest that fine-tuning could improve results.

On the other hand, several studies have shown that injecting structured lexical knowledge into LLMs via fine-tuning can boost performance on lexical-semantic tasks. Moskvoretskii et al. (2024) created TaxoLLaMA by compiling an instruction-tuning dataset from English WordNet hypernym relations, then fine-tuning LLaMA-2 on this data, achieving state-of-the-art results on taxonomy enrichment and hypernym discovery tasks. This illustrates that feeding WordNet-style definitions and relations into an LLM can make its implicit word knowledge more explicit and usable.

For Slovene specifically, Škvorc and Robnik-Šikonja (2025) used GPT-3.5 to extend Slovene dictionary entries: they took short examples from the Dictionary of Standard Slovenian Language (SSKJ) and asked GPT-3.5 to generate full-sentence contexts, improving data for sense disambiguation. Their pipeline showed that LLMs could augment lexicographic examples while preserving sense.

More broadly, Lew (2024) argues that traditional lexicography remains vital for languages like Slovene, especially given current LLM shortcomings in low-resource settings. Yet he also notes that AI tools offer new opportunities to automate routine tasks under a lexicographer's guidance. Our work builds on these efforts by instruction-tuning Slovene LLMs on curated dictionary resources. By combining multiple Slovene lexicographic datasets during fine-tuning, we aim to teach the model accurate, up-to-date lexical knowledge that reflects structured resources.

3. Data Collection and Preparation

We collected data for continued pre-training and fine-tuning to adapt a Slovene LLM for lexicographic tasks. We combine five complementary Slovene

lexicographic resources covering lexical, morphological, semantic, and usage-related information. The Digital Dictionary Database of Slovene (DDDS) serves as the core resource, providing structured lexical data for approximately 300,000 entries. Morphosyntactic annotations are drawn from the SUK corpus (Arhar Holdt et al., 2024). Definitions are supplemented by the Bridge Dictionary (CJVT, 2024) and OSWN (Čibej et al., 2023), while a synonym dictionary (Krek et al., 2023) provides sense-level synonym mappings.

3.1. Lexical Pretraining Corpus

Structured lexicographic entries were converted into natural language text to enable continued pretraining. The corpus is restricted to single-lexeme entries to ensure consistent template-based conversion.

Each entry lists all morphological forms from the DDDS, followed by word senses and definitions from SSKJ, Open Slovene Wordnet (OSWN), and the Bridge Dictionary, or semantic indicators where definitions are unavailable. Usage examples, collocations, and sense-organized synonyms are included where available.

Conversion was performed using rule-based templates that generate structured markdown with clear entry boundaries. The final corpus consists of a single markdown file containing 356,294 entries and approximately 875 MB of text.

An example of the pretraining corpus format is shown below (truncated):

```
# Opis besede vzmetnica
Beseda vzmetnicah je samostalnik, občno ime ž
     enskega spola množine v mestniku.
Beseda vzmetnico je samostalnik, občno ime ženskega
     spola ednine v orodniku.
...

## Definicije besede vzmetnica
Definicija besede vzmetnica je: Vzmetnica je velika
     ravna blazina,
ki jo damo v posteljo, da nam je udobneje.

## Kolokacije besede vzmetnica
- nova vzmetnica
- spati na vzmetnici
- posteljna vzmetnica
- zamenjati vzmetnico
...

## Sopomenke besede vzmetnica
- žimnica
- vzmet
```

3.2. QA Fine-Tuning Dataset

For conversational lexicographic QA, we constructed a dataset of 16,508 question–answer pairs from five sources (Table 1). The dataset combines automatically generated, expert-curated, and authentic user questions.

The largest subset contains 12,173 DDDS-derived QA pairs generated using templates, covering morphology, definitions, and common collocations. Additional data includes 112 questions generated based on patterns identified during a promptathon event in which we organized manual efforts to collect samples of lexicographic and linguistic questions that might be answered by such model. The dataset also includes 1,350 QA pairs derived from SUK annotations, 896 expert-answered questions from Jezikovna svetovalnica, a Slovene linguistic forum, and 342 questions from other Slovene online forums.

All data are stored in a ShareGPT JSON format with additional *data_type* and *source* fields.

An example QA item from the fine-tuning corpus is shown below:

```
{
    "conversations": [
        {
            "from": "human",
            "value": "Katere sopomenke ima beseda
     eksperimentalen?"
        },
        {
            "from": "gpt",
            "value": "Sopomenke za eksperimentalen
     so:\n- poskusen\n"
                     "- vzorčen\n- testen"
        }
    ],
    "source": "ddd",
    "data_type": "automatic"
}
```

Source	Description	Count
DDDS	Template-generated QA pairs	12,173
Promptathon	Community prompt patterns	112
SUK	POS-annotated examples	1,350
Svetovalnica	Linguistic advisory portal	896
Forums	Forum discussions	342
Total		**16,508**

Table 1: QA dataset composition by source

4. Evaluation Methodology

We conducted a comprehensive evaluation of our trained models by combining automated text-similarity metrics with LLM-based qualitative assessment. The evaluation compares the base GaMS model with our continued pretrained and fine-tuned variants across both domain-specific lexicographic questions and general-domain tasks.

4.1. Evaluation Datasets

Our primary evaluation was performed on a held-out test set from the lexicographic QA dataset, containing question-answer pairs with reference

(ground-truth) answers. The test split contains 10% of the full QA dataset and preserves the same question-source distribution as the training split. Each sample included:

- A prompt (question)
- A reference answer (correct answer)
- Generated responses from multiple model variants
- Source metadata indicating the origin of each question

Additionally, we evaluated model performance on general-domain tasks using the Slovenian-LLM-eval benchmark dataset, which consists of improved Slovenian translations of widely used English evaluation benchmarks, including ARC (Easy and Challenge), BoolQ, GSM8K, HellaSwag, NQ Open, OpenBookQA, PIQA, TriviaQA, TruthfulQA, and Winogrande, to assess whether domain specialization negatively impacted general language understanding capabilities.

4.2. Text Similarity Metrics

For lexicographic QA evaluation, we computed three complementary automated similarity metrics between each model's generated answer and the reference answer:

1. **Levenshtein Similarity**: Normalized edit distance measuring character-level differences, calculated as $1 - \frac{\text{distance}}{\max(\text{len}_1, \text{len}_2)}$
2. **Sequence Matcher Similarity**: Python's SequenceMatcher ratio measuring sequence alignment quality
3. **Word Overlap Similarity**: Jaccard similarity coefficient of word sets, computed as $\frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$

The final text similarity score was computed as the arithmetic mean of these three metrics, which were all normalized to provide a value between 0 and 1. This provided a robust measure of lexical and structural similarity to the reference answer:

$$\text{TextSim} = \frac{\text{Leven.} + \text{SeqMatcher} + \text{WOverlap}}{3}$$

Additionally, we computed BERTScore (Zhang et al., 2020), a learned metric that evaluates semantic similarity using contextual embeddings. BERTScore computes precision, recall, and F1 scores by matching tokens between candidate and reference texts using cosine similarity of BERT embeddings, providing a more nuanced semantic comparison than lexical overlap metrics.

These metrics served as simple reference metrics for quick comparison between the models as they are computationally efficient; however, for the main evaluation of the model performance, we used an LLM-as-a-judge approach, which judged the models closer to how a human would evaluate the response.

4.2.1. LLM-as-a-Judge Evaluation

To assess qualitative aspects beyond simple text matching, we employed GPT-5.2 as an LLM judge to perform comparisons of model answers. The use of an LLM judge enabled us to evaluate the models on a dataset with around 1000 examples and provided a solution for the lack of any standardized benchmarks for Slovene lexical quesiton answering. For each question, we evaluated model pairs across three dimensions:

1. **Semantic Similarity**: Which answer conveys meaning most similar to the reference answer, considering semantic equivalence rather than lexical matching
2. **Formatting Quality**: Which answer demonstrates superior structure, readability, and appropriate use of formatting elements (whitespace, bullet points, paragraphs, etc.)
3. **Grammatical Correctness**: Which answer exhibits the highest linguistic quality in Slovene, including proper spelling, syntax, and morphological case agreement

To mitigate position bias, where for example the model would prefere to select the first response, we randomized the order of model answers in each comparison. The GPT-5.2 judge was prompted to select the best answer and to respond only with the answer identifier. We set the temperature to 0.3 to balance consistency with nuanced evaluation capabilities. The model only had access to the reference answer in the semantic similarity setting, while the formatting quality and grammatical correctnes were evaluated independently of the reference answer.

While the use of an LLM judge might introduce a bias into the evaluation, we manually checked a small subset of model answers (30 answers from each model) and got results that coincided with the ones provided by the LLM judge.

4.3. Analysis and Aggregation

Results were aggregated at two levels:

- **Overall**: Average similarity scores and judge win counts across all test samples

- **By Source**: Separate aggregation for each data source (DDDS, SUK, Svetovalnica, Forumi, Promptathon) to identify domain-specific performance patterns

5. Results

We evaluated seven model variants to assess the impact of different training strategies on lexicographic question answering performance. Figure 1 shows how each of them was trained. For clarity, we introduce the following naming convention:

- **Base**: GaMS 9B with standard instruction tuning (cjvt/GaMS-9B-Instruct)
- **Base-Nemotron**: Base model retrained on translated Nemotron dataset (NVIDIA, 2024) (cjvt/GaMS-9B-Instruct-Nemotron)
- **Base-Nemotron-R**: Nemotron variant with reasoning prompting
- **Lex-FT**: Base model fine-tuned with lexicographic QA pairs
- **CPT-Lex-FT**: Model with continued pretraining on lexical corpus, then instruction tuned
- **GaMS3-Lex-FT**: Larger GaMS3 12B model fine-tuned with lexicographic data
- **GaMS3-CPT-Lex-FT**: GaMS3 12B model with continued pretraining on lexical corpus, then instruction tuned

5.1. Overall Performance

Table 2 presents the overall evaluation results across all test samples. The evaluation employs text-similarity scores (averaged from Levenshtein distance, sequence matching, and word overlap), BERTScore metrics (precision, recall, and F1), and LLM-as-a-judge assessments across three dimensions: semantic similarity, formatting quality, and grammatical correctness.

The lexicographic fine-tuning models substantially outperformed the baseline variants. The Lex-FT model achieved the highest text similarity (0.542), BERTScore F1 (0.915), and judge preference scores, representing a significant improvement over the Base model. Notably, the Base-Nemotron model performed poorly (0.091 in text similarity, 0.816 BERTScore F1), while excelling in the answer formatting quality dimension, likely because directly translating general-domain datasets can produce well-formatted responses that lack Slovene-specific knowledge. We also observed that the nemotron-trained model tends to provide very verbose responses that included substantial irrelevant information alongside the actual answer, reducing similarity scores.

The reasoning-prompted variant (Base-Nemotron-R) partially mitigated this verbosity issue, improving text similarity to 0.153, though still underperforming the lexicographic models. Among specialized models, the direct fine-tuning approach (Lex-FT) outperformed continued pretraining followed by fine-tuning (CPT-Lex-FT and GaMS3-CPT-Lex-FT), suggesting that for our dataset size and task, supervised fine-tuning provided more efficient knowledge transfer than the two-stage approach. The GaMS3-CPT-Lex-FT model showed comparable performance to GaMS3-Lex-FT, indicating that the additional continued pretraining stage did not provide substantial benefits for the larger model architecture.

5.2. Performance by Data Source

Figure 2 shows the model performance evaluated by semantic similarity to the reference answer, split by the source of examples. More detailed information from two representative data sources: SUK (sentence analysis questions) and Digital Dictionary Database of Slovene (lexical knowledge questions) is also presented in Table 3 and Table 4.

On SUK questions requiring sentence-level morphosyntactic analysis, lexicographic fine-tuning models achieved near-perfect text similarity scores (>0.95), with Lex-FT reaching 0.985. This represents a 202% improvement over the Base model. The baseline models completely failed on these structured analytical tasks (0 judge wins), while specialized models excelled.

For Digital Dictionary Database questions testing direct lexical knowledge, Lex-FT again led with 0.681 text similarity, though the absolute scores were lower than on the morphosyntactic analysis from SUK, suggesting greater task complexity and answer variability. The Base-Nemotron model catastrophically failed (0.028), while Base-Nemotron-R recovered somewhat (0.269), demonstrating the impact of prompting strategies on response quality.

Performance patterns remained consistent across different data sources. On Jezikovna Svetovalnica (language advisory portal) questions, all models achieved modest text similarity scores (0.08–0.11), with no clear winner, suggesting these authentic user questions present greater complexity with more varied valid answer formulations. The lexicographic models maintained competitive performance without dramatic improvement, indicating that human-generated linguistic advice may be harder to replicate than structured lexical knowledge.

On the promptathon questions, specialized models again dominated, with Lex-FT and CPT-Lex-FT

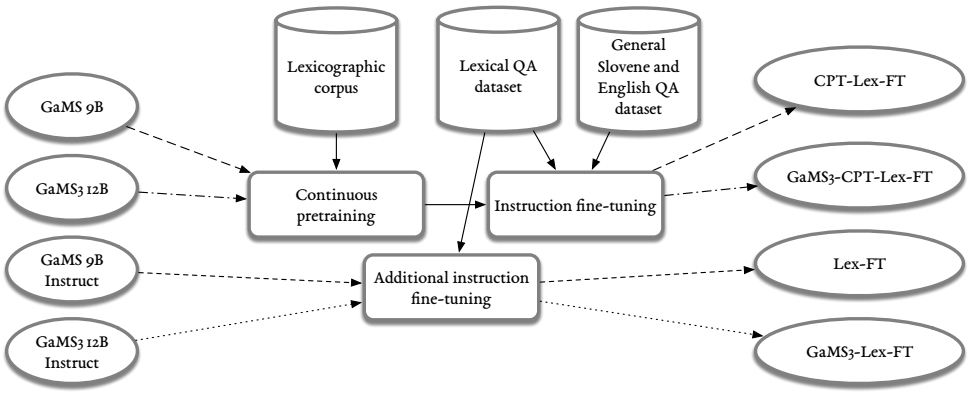

Figure 1: Overview of the training process showing the two-stage approach: continued pretraining on lexical corpus followed by instruction fine-tuning with QA pairs, applied to both GaMS 9B and GaMS3 12B base models.

Model	Text Sim.	BERT-P	BERT-R	BERT-F1	Semantic	Format	Grammar
Base	0.226	0.863	0.851	0.856	31	26	49
Base-Nemotron	0.091	0.792	0.844	0.816	16	**170**	13
Base-Nemotron-R	0.153	0.826	0.842	0.833	22	12	24
Lex-FT	**0.542**	**0.919**	**0.910**	**0.915**	**93**	34	**90**
CPT-Lex-FT	0.516	0.907	0.908	0.907	61	20	45
GaMS3-Lex-FT	0.497	0.905	0.900	0.902	33	10	38
GaMS3-CPT-Lex-FT	0.503	0.908	0.901	0.904	39	23	36

Table 2: Overall evaluation results. Text similarity and BERTScore metrics range 0–1 (higher is better). BERT-P/R/F1 denote BERTScore precision, recall, and F1. Judge metrics show number of wins in pairwise comparisons across all samples.

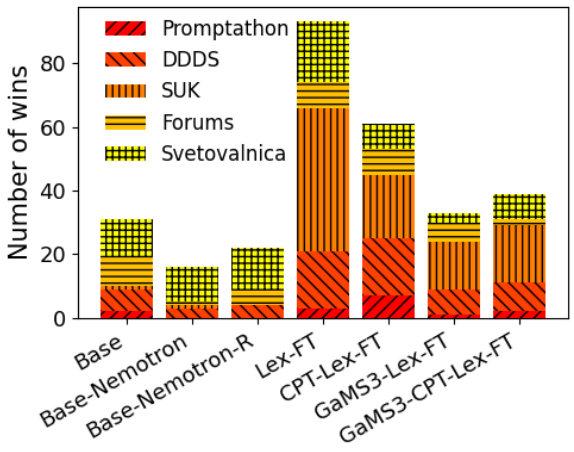

Figure 2: LLM-as-a-judge evaluation results showing the number of wins each model achieved across different data sources in pairwise comparisons for semantic similarity to the reference answer.

achieving 0.86 text similarity compared to 0.46 for Base. The Slovene forums' questions showed less pronounced differentiation, with Lex-FT reaching 0.236 compared with 0.136 for Base, likely reflecting the informal, discussion-oriented nature of forum answers.

Model	Text Sim.	Semantic
Base	0.326	0
Base-Nemotron	0.142	0
Base-Nemotron-R	0.161	0
Lex-FT	**0.985**	**54**
CPT-Lex-FT	0.959	21
GaMS3-Lex-FT	0.953	25

Table 3: SUK sentence analysis results (subset of judge metrics shown for brevity).

Model	Text Sim.	Semantic
Base	0.197	10
Base-Nemotron	0.028	1
Base-Nemotron-R	0.269	8
Lex-FT	**0.681**	**40**
CPT-Lex-FT	0.571	23
GaMS3-Lex-FT	0.635	18

Table 4: Digital Dictionary Database lexical knowledge results.

5.3. Qualitative Error Analysis

Beyond quantitative metrics, a qualitative examination of model outputs reveals characteristic error patterns that shows the impact of training data qual-

ity and domain adaptation. We present representative examples of common failure modes across model variants.

5.3.1. Language Confusion in Translation-Based Models

The most common failure pattern occurs in models trained on automatically translated data. The Base-Nemotron model frequently confuses English and Slovene linguistic concepts, producing responses that discuss English grammar when answering questions about Slovene. For example, when asked whether the Slovene phrase "kolikor" can appear at the beginning of a sentence, the Base-Nemotron model responded (in Slovene): "Yes, the phrase kolikor can stand at the beginning of a sentence, but in formal or literary English this is not common. It is more commonly used in more informal or colloquial language or in specific grammatical structures."

This error demonstrates that automatic translation of instruction-tuning datasets can introduce fundamental confusion about the target language's identity. The model correctly understands the question structure but applies knowledge about English grammar to answer a question explicitly about Slovene. The lexicographic fine-tuning eliminated this problem, maintaining proper language context throughout their responses.

5.3.2. Repetition and Degeneration

Another characteristic failure involves repetitive generation, particularly when models attempt to provide information they lack. When asked to list synonyms for the Slovene verb "priznati" (to admit/acknowledge), which has no clear synonyms, the Base-Nemotron model produced an extensively structured but contentless response that repeatedly listed "priznati" itself as its own synonym across multiple categories.

This pattern of repetitive degeneration appears when the model attempts to conform to an expected response structure (categorized synonym lists with examples) without possessing the necessary lexical knowledge. The response demonstrates superficial understanding of the task format but complete failure to provide useful content.

In contrast, the lexicographic fine-tuned models produced more appropriate responses, by providing semantically related forms like "priznati se" (reflexive form) or "priznavati" (imperfective aspect), which represent morphological variants rather than true synonyms. While these responses are not perfect, they demonstrate substantially better understanding of Slovene lexical structure.

5.3.3. Cross-Linguistic Interference in Specialized Terminology

A third error pattern involves providing information about the wrong language's linguistic system. When asked to list examples of Slovene auxiliary verbs ("pomožni glagoli"), both Base-Nemotron models provided extensive lists of *English* auxiliary verbs in addition to the Slovene verbs.

This failure is particularly problematic because the question asks about Slovene verbs, yet the model responds mostly in terms of English grammar. While the lexicographic models maintained proper language context by providing Slovene terms such as "biti" (to be), "imeti" (to have), "morati" (must), and "smeti" (may), they also produced incorrect answers: in standard Slovene grammar, only "biti" is classified as an auxiliary verb. This reveals a limitation of purely fine-tuning-based approaches, even with domain-specific training, models may generate plausible but factually incorrect information when lacking access to authoritative structured resources.

These error patterns collectively demonstrate that automatic translation of general-domain training data introduces systematic language confusion that persists even in models with strong general capabilities. Domain-specific fine-tuning on native Slovene lexicographic data substantially improves language alignment and eliminates cross-linguistic interference, but does not guarantee factual accuracy on all specialized queries. This limitation points to the need for retrieval-augmented generation systems that can directly query authoritative lexicographic databases to ensure correct, verifiable answers, particularly for questions with precise, definitional answers.

5.4. Key Findings

Our evaluation reveals several key insights. (1) Lexicographic models outperformed baselines, with even larger gains on structured tasks like sentence analysis. (2) The Base-Nemotron model's poor results (0.091 text similarity, 0.816 BERTScore F1) demonstrate that automatically translated general-domain datasets may introduce artifacts problematic for specialized tasks. (3) Contrary to our initial hypothesis, the Lex-FT model (direct fine-tuning) consistently outperformed CPT-Lex-FT and GaMS3-CPT-Lex-FT (continued pretraining + fine-tuning), suggesting that for datasets of our size (15K examples), supervised learning on task-specific data provides more efficient adaptation than unsupervised continued pretraining. (4) The GaMS3-Lex-FT and GaMS3-CPT-Lex-FT models (12B parameters) did not substantially outperform the 9B Lex-FT model, indicating that training data quality and task alignment matter more than

raw parameter count for specialized domains. (5) Models excelled on structured tasks (SUK: 0.985) but showed modest improvements on open-ended linguistic advice (Svetovalnica: 0.11), highlighting the importance of answer structure and definedness in evaluation outcomes.

5.5. General Benchmark Evaluation

To assess whether our domain-specific training negatively impacted general language understanding capabilities, we evaluated our models on the Slovenian LLM Evaluation Dataset, which contains Slovenian translations of widely used English benchmarks. This dataset, developed by Arčon et al. (2024), includes improved translations of popular benchmarks such as ARC (Easy and Challenge), BoolQ, HellaSwag, NQ Open, OpenBookQA, PIQA, TriviaQA, and Winogrande.

Table 5 presents the results for selected models and benchmarks. The full results across all benchmarks are shown in Figure 3.

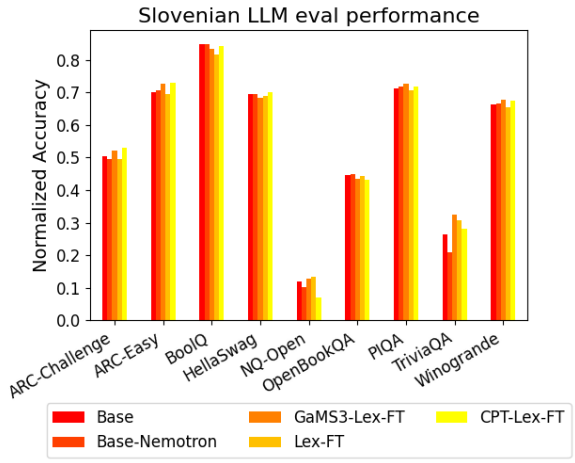

Figure 3: Comprehensive benchmark evaluation results across all models and evaluation tasks, showing accuracy scores on the Slovenian LLM Evaluation Dataset benchmarks.

The results demonstrate that our lexicographic fine-tuning had minimal negative impact on general language understanding. The CPT-Lex-FT models (which underwent continued pretraining followed by instruction fine-tuning) maintained or even slightly improved performance on most benchmarks compared to the Base model (0.756 on ARC-Easy, 0.842 on BoolQ, 0.700 on HellaSwag).

The direct fine-tuning models (Lex-FT) showed a modest decrease in some benchmarks, particularly on ARC Challenge and BoolQ, suggesting a slight trade-off between domain specialization and general capabilities. However, the differences remain relatively small (typically 2-5 percentage points), indicating that the models retained strong general language understanding while gaining substantial improvements in lexicographic tasks.

Notably, the GaMS3-Lex-FT model (based on the larger 12B parameter architecture) maintained competitive performance across all benchmarks, demonstrating that the larger model capacity helped preserve general capabilities despite domain-specific training.

6. Discussion

Our results demonstrate the effectiveness of domain-specific fine-tuning for lexicographic question answering. Gains were especially pronounced on structured tasks (SUK: 0.985), while performance on open-ended questions from Jezikovna Svetovalnica remained modest, suggesting that expert linguistic advice is harder to replicate with limited supervised data.

Contrary to expectations, direct fine-tuning (Lex-FT) outperformed the two-stage approach (CPT-Lex-FT), suggesting that for moderate-sized datasets, supervised learning provides more efficient adaptation than multi-stage training. Similarly, increasing model size from 9B to 12B parameters yielded minimal gains, indicating that data quality and task alignment matter more than parameter count.

The Base-Nemotron model's failure highlights critical risks of automatically translated training data, which conflates English and Slovene linguistic concepts. While domain-specific fine-tuning eliminates cross-linguistic interference, it does not guarantee factual accuracy, motivating retrieval-augmented generation approaches.

Importantly, domain specialization did not harm general language capabilities, with only modest decreases (2-5 percentage points) on general benchmarks. Our findings offer a practical strategy for low-resource languages: existing linguistic resources can be effectively repurposed, direct fine-tuning often suffices, and moderately sized models are sufficient.

7. Conclusion

In this work, we investigated domain-specific adaptation of large language models for Slovene lexicographic question answering. We proposed a two-stage training approach combining continued pretraining on a large-scale lexical corpus with instruction fine-tuning on curated question-answer pairs. To support this process, we constructed a comprehensive lexical pretraining corpus containing 356,294 word entries and a dedicated instruction dataset of 16,508 lexicographic QA pairs drawn from both structured resources and authentic user queries.

Model	ARC-C	ARC-E	BoolQ	HellaSwag	PIQA
Base	0.515	0.732	0.848	0.693	0.702
Base-Nemotron	0.506	0.739	0.847	0.696	0.705
GaMS3-Lex-FT	0.500	0.740	0.832	0.683	0.702
Lex-FT	0.469	0.711	0.805	0.671	0.690
CPT-Lex-FT	0.515	0.756	0.842	0.700	0.707

Table 5: General benchmark evaluation results (accuracy) for selected models on key benchmarks. ARC-C = ARC Challenge (normalized accuracy), ARC-E = ARC Easy (accuracy), BoolQ = Boolean Question accuracy, HellaSwag = commonsense inference (normalized accuracy), PIQA = physical commonsense reasoning (accuracy).

Our experimental results show that domain-specific fine-tuning substantially improves model performance on lexicographic tasks. In particular, direct instruction fine-tuning of the GaMS model led to large gains over the baseline across multiple evaluation settings, demonstrating that targeted supervision effectively transfers structured lexicographic knowledge into a conversational QA format. While continued pretraining did not provide additional benefits in our setting, the overall approach successfully enhanced the model's ability to deliver accurate, well-structured lexicographic answers without degrading general language understanding.

This work demonstrates that high-quality, native-language linguistic resources can be efficiently repurposed to build specialized language models for low-resource languages. Our findings highlight that careful data curation and task-aligned fine-tuning can outweigh increased model size or complex training pipelines, offering a practical pathway for developing domain-specific language technologies in smaller language communities.

8. Acknowledgements

This work was financially supported by the Slovenian Research and Innovation Agency through the research project Large Language Models for Digital Humanities (GC-0002). This work is also co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 AI4DH. We thank Luka Dragar for assistance with data preparation and Domen Vreš for providing access to the latest GaMS models and help with their training.

9. Bibliographical References

Tjaša Arčon, Timotej Petrič, and Domen Vreš. 2024. Slovenian llm evaluation dataset. Hugging Face.

CJVT. 2024. GaMS: Generative model for slovene.

CJVT. 2026. Gams3-12b-instruct. Hugging Face.

Aleš Horák and Adam Rambousek. 2017. Lexicography and natural language processing. In *The Routledge handbook of lexicography*, pages 179–196. Routledge.

Robert Lew. 2023. ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1):1–10.

Robert Lew. 2024. Dictionaries and lexicography in the ai era. *Humanities and Social Sciences Communications*, 11(1):1–8.

Raphaël Merx, Ekaterina Vylomova, and Kemal Kurniawan. 2024. Generating bilingual example sentences with large language models as lexicography assistants. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 64–74.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350.

NVIDIA. 2024. Nemotron post-training dataset v1. Hugging Face.

Bach Pham, JuiHsuan Wong, Samuel Kim, Yunting Yin, and Steven Skiena. 2025. Word definitions from large language models. In *2025 19th International Conference on Semantic Computing (ICSC)*, pages 158–162. IEEE Computer Society.

Tadej Škvorc and Marko Robnik-Šikonja. 2025. Solving word-sense disambiguation and word-sense induction with dictionary examples. *arXiv preprint arXiv:2503.04328*.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. *arXiv preprint arXiv:2410.06898*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

10. Language Resource References

Arhar Holdt, Špela and Krek, Simon and Dobrovoljc, Kaja and Erjavec, Tomaž and Gantar, Polona and Čibej, Jaka and Pori, Eva and Terčon, Luka and Munda, Tina and Žitnik, Slavko and Robida, Nejc and Blagus, Neli and Može, Sara and Ledinek, Nina and Holz, Nanika and Zupan, Katja and Kuzman, Taja and Kavčič, Teja and Škrjanec, Iza and Marko, Dafne and Jezeršek, Lucija and Zajc, Anja. 2024. *Training corpus SUK 1.1*. Slovenian language resource repository CLARIN.SI.

Jaka Čibej, Luka Terčon, Simon Krek, Andraž Repar, Erik Novak, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Kaja Dobrovoljc, Amadea Berginc, Irena Hvala, Damijan Klement, Manja Kolenc, Ana Močnik, Tina Munda, David Pavlas, Anamari Pečan, Aleksandra Poljak, Davorin Sečnik, Jure Šešet, Jan Štumberger, Tina Toličič, and Laura Trpin. 2023. Open slovene WordNet OSWN 1.0. Slovenian language resource repository CLARIN.SI.

CJVT. 2024. *Bridge Dictionary of Slovene*. Centre for Language Resources and Technologies. CJVT Language Resources.

DDDS. List of linked senses from open slovene wordnet and the digital dictionary database of slovene 1.0. https://b2find.eudat.eu/dataset/5a2c5642-240c-5a27-8aad-a7896d7ceb77. Dataset - B2FIND.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik, and Marko Robnik-Šikonja. 2019. Morphological lexicon sloleks 2.0. Slovenian language resource repository CLARIN.SI.

Iztok Kosem, Špela Arhar Holdt, Simon Krek, Polona Gantar, Eva Pori, Jaka Čibej, Bojan Klemenc, Cyprian Laskowski, Kaja Dobrovoljc, Vojko Gorjanc, Nikola Ljubešić, Karolina Zgaga, and Rebeka Roblek. 2023. Collocations dictionary of modern slovene KSSS 2.0. Slovenian language resource repository CLARIN.SI.

Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc, Kaja Dobrovoljc, Eva Pori, Rebeka Roblek, and Karolina Zgaga. 2023. Thesaurus of modern slovene 2.0. Slovenian language resource repository CLARIN.SI.