High-Quality LLM Pre-Training Texts from Dictionary Data

Marek Medveď[1,2], Radoslav Sabol[1], Ondřej Sotolář[1], and Aleš Horák[1]

[1] NLP Centre, Faculty of Informatics, Masaryk University
Brno, Czechia,
[2] Lexical Computing
{xmedved1, xsabol, xsotolar, hales}@fi.muni.cz

Abstract. The quality of the pre-training texts is an important aspect in the development of a Large Language Model (LLM). High-quality data, such as collections of textbooks, academic papers, and educational forums, has been shown to improve model performance, generalization, and reduce biases. However, obtaining such data at scale can be challenging, especially for non-mainstream languages like Czech.
In this paper, we introduce a method for generating high-quality Czech pre-training data from structured dictionary resources. By employing retrieval-augmented prompting and open-source LLMs, we transform XML-encoded lexicographic dictionary entries into fluent, semantically rich text. The resulting dataset demonstrates that dictionary-grounded generation can effectively enhance data quality.
We present the results of experiments with several LLMs and the process of creating a new Czech pre-training dataset, SlamaHQTrain. This dataset was obtained by processing eight Czech dictionaries containing more than 500,000 entries and 18 million words.

Keywords: large language models; LLMs; pre-training; high-quality data; dictionaries; dictionary entries; Slama models; Czech

1 Introduction

Pre-training Large Language Models (LLMs) requires large quantities of textual data. Various proposed scaling laws have tried to estimate the necessary quantity as related to the number of parameters of the models [15,12]. Braking the scaling laws, that is, training a better model with fewer data, was shown to be possible by improving the quality of the training data [9]. The current research suggests combining both scale and data quality: begin pre-training with large quantity of less-curated data and finish the pre-training with high-quality data.

Simultaneously, further advancement of LLMs is affected by the rising data scarcity, as the easily and publicly available text data have already been included in pre-training corpora. This issue is more pronounced in less-resourced languages, such as languages of the Slavonic group, where the currently available corpora are not extensive enough for state-of-the-art LLM pre-training [32].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 69–83, 2025.

The current research therefore focuses on finding more pre-training data, by retrieval or generation. Training with LLM-generated data can lead to forgetting the true distribution of language, leading to loops and duplicities in generation [29]. Since retrieval is limited by existing resources and generation can degrade the text quality, the combination of the two methods has emerged as a possible solution. When randomness or grounding is introduced to the generation prompt, the distribution of the generated data is positively influenced [9].

In this work, we propose a new pre-training dataset generated by prompting an LLM to produce fluent text grounded in dictionary records. Dictionaries comprise structured records organized by keywords and include information on semantics, part-of-speech, synonyms, and other language features. Pre-training on such structured data is not desirable because of the impact on the output distribution. Therefore, we propose a retrieval-augmented prompting approach to convert the structured records into fluent text. We control the level of distillation of the teacher LLM and correct for hallucination by dynamically composing the prompt for each dictionary record. In the prompt, we include examples based on their similarity, both semantical and on statistical features, to the target record.

The proposed dataset encompasses eight Czech dictionaries and is estimated to amount to 104 M tokens. We generated a sample of the final dataset using four top-scoring LLMs, such as DeepSeek, Llama, and GPT-OSS. To perform a model selection, we measured the quality of the sample for each model using a range of statistical metrics and a human preference score. Finally, we detail the needed infrastructure on the LUMI cluster and calculate the required compute time.

2 Related Work

LLMs such as Alpaca [31], Orca [24], and Phi [21,9,1] demonstrated that high-quality synthetic data produced by more capable teacher models can effectively substitute or complement natural data in pre-training or instruction tuning. The TinyStories project [4] illustrates that even small LLMs can acquire rich linguistic competence when trained on compact, high-quality synthetic datasets. LexMatcher [35] uses bilingual dictionaries to generate a bilingual dataset for machine translation. The Phi models demonstrated, that using lower quantities of high-quality data in pre-training leads to improved performance contrary to the standard scaling-laws [12]. In Phi 3 [1], the authors pre-trained the model in multiple stages with increasing data-quality, although ablations were not performed.

However, the reliance on model-generated data has also drawn significant criticism. Studies like [8] argue that synthetic corpora often replicate the biases, factual errors, and stylistic artifacts of their teacher models, leading to stagnation rather than genuine improvement. Similarly, [29] highlights the risk of performance degradation when generations are recursively used for training, warning of an inevitable quality collapse in self-referential data loops. Furthermore,

Table 1: Dictionary sizes by entry and token count (by tokenizer for *gpt-oss-120b*).

Dictionary	entries	tokens	tokens wo. xml
PSJČ	201,327	42,548,533	22,082,861
SSJČ	123,992	19,762,249	11,225,573
Diderot	90,982	10,302,547	8,480,832
ASCS	45,661	3,202,618	2,618,056
SSC	26,529	4,419,191	1,904,434
SČS	23,276	966,776	565,062
SCFIS	13,552	1,929,538	1,390,761
SCFIN	3,992	776,983	504,114
Total	529,311	83,908,435	48,771,693

some studies, such as the Phi models, do not disclose their data-generation methods; therefore, they are not replicable.

As Czech is less-resourced than languages such as English, with total resources amounting to significantly less than 1 T tokens, there is implicit motivation to extend the training data. This lack of availability of high-quality pre-training is being actively solved with initiatives such as MultiSynt [25], aiming to support the EuroLLM and OpenEuroLLM models.

The evaluation of the quality of data generated for pre-training LLMs remains an unsolved question. Existing studies rely on a mix of proxies such as downstream performance (for Czech the latest being [5]), intrinsic statistics (e.g., perplexity, diversity), LM-as-judge scores, or human annotation. Recent analyses show that automatic metrics often correlate poorly with human judgments and may reward over-specialized synthetic data [16,22]. To address this, several works propose multi-axis evaluation combining controlled pre-training experiments, intrinsic diagnostics, with human or model-based assessments [14,30]. Nevertheless, open challenges persist in balancing competing metrics (e.g., diversity vs. correctness) and developing standardized, scalable benchmarks for synthetic data evaluation.

3 Czech Dictionary Data

The development of a high-quality LLM pre-training data is based on Czech dictionary resources of high lexicographic quality. These datasets were developed by professional lexicographers and represent the most comprehensive Czech dictionaries available. We selected the largest datasets collected within the DEBdict project [13]. In total, over 500,000 dictionary entries comprising more than 18 million tokens have been collected (Table 1). The dictionary texts are then processed by specifically tailored prompting of a selected large language model that transforms their structured descriptions into coherent, natural-language explanations capturing the presented meanings of each entry.

SSJČ

Anton

-a m. dom., zčásti hanl. *Antonín*

```
<root>
  <h>Anton, </h><norm>-a </norm>
  <small>m. dom., zčásti hanl. </small>
  <ital>Antonín</ital>
</root>
```

PSJČ

autor, -a m. *spisovatel literárního díla*. Sebevědomého autora nedostatek čtenářů nikdy nemrzí. V. Mrš. **D** *vůbec původce, tvůrce něčeho*. Autor kresby, projektu, vtipu *a p*.

```
<h>
  <Cil>autor</Cil><Heslo>autor, </Heslo>
  <Tvar>-a </Tvar><Gram>m. </Gram>
  <Vyzn>spisovatel literárního díla.</Vyzn>
  <Dokl>Sebevědomého autora nedostatek čtenářů
  ↪ nikdy nemrzí.</Dokl>
  <Pram>V. Mrš.</Pram>
  <Sep>D</Sep>
  <Vyzn>vůbec původce, tvůrce něčeho.</Vyzn>
  <Dokl>Autor kresby, projektu, vtipu</Dokl>
  <Vyzn>a p.</Vyzn>
</h>
```

ASCS

acylpyrin

-u m ⟨l + ř⟩ farm. lék snižující teplotu a utišující bolesti (kyselina acetylsalicylová)

```
<root>
  <h>acylpyrin</h>
  <t>-u  m <l + ř> farm. lék snižující teplotu
  ↪ a utišující bolesti (kyselina
  ↪ acetylsalicylová)</t>
</root>
```

SČS

adresát

příjemce; #odesílatel;

```
<ENTRY>
  <HEAD>adresát</HEAD>
  <LINE>příjemce; #odesílatel;</LINE>
</ENTRY>
```

SSČ

chemikálie

-e ž *(laboratorní a průmyslová) sloučenina*

```
<root>
  <h>chemikálie</h>
  <gram>-e ž</gram>
  <exp>(laboratorní a průmyslová)
  ↪ sloučenina</exp>
</root>
```

SČFIN

plodný autor

(kniž; pochv; někdy iron) **0** vok *velmi produktivní umělec, zejm. spisovatel hojně publikující; (v iron. smyslu) umělec, v jehož tvorbě převažuje kvantita nad kvalitou* Patřil mezi nejplodnější české a-y. — Tyto plodné a-y v redakci neměli rádi. ◆ Též v přechýl. podobně *plodná autorka*. **Cf** muž práce; muži činu • **A** prolific writer/author **N** fruchtbarer Autor **F** auteur fécond

```
<root>
  <h>plodný autor</h>
  <n>(kniž; pochv; někdy iron) </n>
  <b>0</b><n>vok</n>
  <i>velmi produktivní umělec, zejm.
  ↪ spisovatel hojně publikující; (v iron.
  ↪ smyslu) umělec, v jehož tvorbě převažuje
  ↪ kvantita nad kvalitou </i><n>Patřil mezi
  ↪ nejplodnější české a-y. - Tyto plodné
  ↪ a-y v redakci neměli rádi. - Též v
  ↪ přechýl. podobně </n>
  <i>plodná autorka.</i><b>Cf</b>
  <n>muž práce; muži činu</n>
  <b> - A </b><n>prolific writer/author
  ↪ </n><b>N </b><n>fruchtbarer Autor
  ↪ </n><b>F </b><n>auteur fécond </n>
</root>
```

SČFIS

napadat na nohu

(čl. při chůzi:) viditelně kulhat

```
<root>
  <h>napadat na nohu </h>
  <i>(čl. při chůzi:) viditelně kulhat</i>
</root>
```

Diderot

A

značka pro ampér.

```
<ENTRY>
  <KEY>A</KEY>
  <MEANING>značka pro ampér.</MEANING>
</ENTRY>
```

Fig. 1: Dictionary examples

3.1 SSJČ

The *Dictionary of Standard Czech Language* (Slovník spisovného jazyka českého, SSJČ) is an explanatory dictionary of the Czech language compiled by a team from the Institute of the Czech Language (ÚJČ) under the supervision of Bohuslav Havránek. The last edition was published in 1989 [11]. The dictionary includes several types of entries reflecting its normative and codifying purpose. The core of the dictionary consists of fully developed lexical entries with systematically presented semantic, grammatical, and stylistic information. In addition, it includes cross-references to related or variant forms, phraseological and multi-word units, as well as derivational and morphological variants illustrating word-family relations. The dictionary also covers technical and foreign terms, indicating their origin and field, along with expressions marked for stylistic or temporal usage (archaisms, poetic, colloquial, and literary words). Altogether, it offers a comprehensive overview of standard Czech in its functional and stylistic diversity.

3.2 PSJČ

The *Concise Dictionary of the Czech Language* (Příruční slovník jazyka českého, PSJČ) [10] was the most extensive explanatory dictionary of Czech. It was published in nine volumes between 1935 and 1957 by four different publishers. The main editors were Oldřich Hujer, Emil Smetánka, Miloš Weingart, and Bohuslav Havránek.

The dictionary comprises several types of entries. Its core consists of fully defined lexical units. In addition, the dictionary includes cross-references, fixed expressions and idioms, word-formation variants and derivatives. It also records archaic, regional, poetic and literary words, as well as technical terms from various fields.

3.3 ASCS

The *Academic Dictionary of Foreign Words* (Akademický slovník cizích slov, ASCS) [19] is the most comprehensive Czech dictionary of foreign terms, compiled by the Lexicographic Department of the Institute of the Czech Language of the Czech Academy of Sciences. It covers a broad range of vocabulary—from specialized terminology to everyday expressions—encompassing common, rare, literary, colloquial, and slang words. Each entry provides standard pronunciation, basic etymological data, grammatical and stylistic notes, and examples of use in context.

3.4 SČS

The *Dictionary of Czech Synonyms* (Slovník českých synonym, SČS) [27] systematically presents Czech words together with their synonymous counterparts. For each headword, it lists alternative expressions with identical or closely related meanings, illustrating their use in appropriate linguistic and stylistic contexts.

3.5 SSČ

The *Dictionary of Standard Czech for Schools and the Public* (Slovník spisovné češtiny pro školu a veřejnost, SSČ) [6] is a key normative explanatory dictionary codifying contemporary Czech, compiled by the Institute of the Czech Language (Czech Academy of Sciences). Beyond definitions, each entry provides comprehensive data on orthography, pronunciation, morphological paradigms, stylistic register, and typical collocations. The work includes supplementary appendices on word-formation, frequent anthroponyms, and toponyms.

3.6 SČFI

The *Dictionary of Czech Phraseology and Idiomatics* (Slovník české frazeologie a idiomatiky, SČFI) is the most comprehensive, descriptive dictionary in its field, based on extensive corpus data. This fundamental work is systematically divided into four volumes according to the formal structure of the idioms: I. Comparisons, II. Non-verbal expressions, III. Verbal expressions, and IV. Clausal expressions. Each entry provides a complex description, including variants, definitions, stylistic classification, and usage examples. In our work, we used only volumes II [17] and III. [18].

3.7 Diderot

The *Great Educational Dictionary: the Diderot Encyclopaedia* (Velký slovník naučný: encyklopedie Diderot) [34] is a two-volume encyclopedic work offering an extensive overview of concepts and terminology across diverse fields of science, culture, and technology. It covers disciplines such as the natural sciences, humanities, history, and the arts. Each volume is organized alphabetically and provides clear definitions, explanations, and contextual notes for individual terms.

4 Generating High-Quality Texts from Dictionaries

The basis of our method is seeding a large, state-of-the-art LLM with records from Czech dictionaries and prompting it to generate one or more paragraphs in Czech. The aim is to reflect the lexical and semantic variety of Czech while maximally leveraging the source dictionary records. As the generation is computationally intensive, at the time of submission we have only generated a sample of the final dataset; therefore, the statistics in Table 5 are based on an estimate from this sample.

4.1 Preliminary Experiments

Before processing the complete set of selected dictionaries, we conducted a series of preliminary experiments to determine the most suitable computational setup for converting dictionary entries into natural-language sentences.

Servers

- server **A**:
  - 192 CPU cores
  - 503 GB RAM
  - NVIDIA L40S — 48 GB VRAM, 18,176 CUDA cores
- server **B**:
  - 80 CPU cores
  - 377 GB RAM
  - 5 × NVIDIA A40 — 48 GB VRAM, 10,752 CUDA cores
- server **C**:
  - 256 CPU cores
  - 2 TB RAM
  - no GPU

To establish a controlled testing environment, we used a simple and semantically neutral prompt — "Why is the sky blue?" — and evaluated the Ollama runtime with the **Qwen3** 235b model under three distinct configurations 2.

In the first configuration, the entire model fit within the memory of the graphics card, allowing for full GPU inference without offloading. The second configuration also utilized GPUs but exceeded their available memory, causing the system to partially offload model components into standard system RAM. The third configuration completely excluded GPU acceleration and relied solely on CPU computation and main memory. These experiments were executed on two high-performance computing servers **A** and **B** with different hardware specifications.

A subsequent preliminary test was conducted on the **DeepSeek** v3.1 671b Large Language Model (LLM). Due to the unavailability of GPU resources within our cluster capable of accommodating this model's memory (VRAM) requirements, we utilized server **C**, a high-memory CPU-based machine. The evaluation involved three full prompts, structured according to the template defined in Section 4.3. To analyze performance under different constraints, each prompt was executed twice. The first setup utilized the entire machine's resources, while the second setup was constrained to 100 CPU cores, allowing two instances to be tested simultaneously. The resulting performance statistics are presented in Table 3.

These tests provided valuable insights into the trade-offs between computational efficiency, memory management, and inference speed, guiding the selection of the optimal setup for large-scale dictionary processing in subsequent experiments.

4.2 Prompt Development

Prior to establishing the final version of the prompt, we conducted a series of exploratory experiments aimed at determining the most effective prompting strategy. Two principal approaches were investigated. The first was a direct,

Table 2: Speed testing of Qwen3 235b model in different server settings.

Server	GPU no.	GPU VRAM (MB)	GPU max %	CPU max %	RAM (MB)	Dry run time	Run time
A	5	135471	24	88	333	1m52s	0m59s
A	4	135473	28	93	333	1m38s	1m03s
A	3	119616	15	1855	16192	2m37s	2m42s
B	1	39523	9	4114	96286	5m16s	7m52s
B	0	0	0	7428	135809	2m57s	2m06s

Table 3: Speed testing of deepseek v3.1 671b model on server C.

Server	Run	CPU cores available per run	Throughput tok./sec.
C	single run	256	3.8
C	2 parallel runs	100	3.3

single-step approach, in which the model was instructed to transform a lexical entry provided in XML format directly into the target explanatory output. This output was required to both define and elucidate the entry while remaining as concise as possible.

The second approach adopted a two-stage processing pipeline. In the first stage, the model converted the XML-encoded entry into a plain-text representation. In the subsequent stage, the model generated a natural-language explanation based on this intermediate text. This method allowed for a clearer separation between structural conversion and semantic interpretation, potentially improving output consistency and interpretability.

For the single-step approach, eight prompt variants were evaluated, whereas the two-step method involved ten prompts for the conversion phase and twenty-five prompts for the explanatory phase. The results from these experiments informed the refinement of the final prompt design used in the subsequent stages of model evaluation.

- **Prompts for the single-step method:**
  - **CZ** Transformujte dané XML slovníkové heslo na souvislý text tak, aby podrobně objasňoval každý význam slova a ilustroval jeho použití pomocí příkladů.
  - **EN** Transform the given XML dictionary entry into a coherent text that thoroughly explains each meaning of the word and illustrates its usage with examples.
- **Prompts for the two-step method:**
  - **CZ:** Přepiš následující heslo jako běžný text.
  - **EN:** Rewrite the following entry as plain text.
  - **CZ:** Transformuj tento strukturovaný text na souvislý text složený z vět.
  - **EN:** Transform this structured text into a coherent passage composed of full sentences.

4.3 Final Prompting Setup

The final prompt, developed using guidelines from the OpenAI Cookbook [26], which produced the evaluation sample, consists of eight parts:

1. Role and task: describes the basic task,
2. Input format: a list of dictionary XML tags and their semantics,
3. Output style: output formatting instructions,
4. Abbreviations: abbreviation explanations,
5. Sources: non-abbreviated sources,
6. Examples: pairs of inputs and human-corrected outputs,
7. Task: final reinforcement of task definition,
8. Input.

The and Role (1) and Output Style (3) sections of the prompt are fixed. Section Input format (2) describes the XML syntax for each particular dictionary. Abbreviations (4) and Sources (5) that occur in the input record are retrieved from the each dictionaries lists thereof, or a shared pool if unavailable. Examples (6) are retrieved from a set of hand-corrected examples for each dictionary. They are matched based on sematic similarity using Sentence-BERT-based embeddings. Appendix A contains an example of an input (Figure 2), complete prompt (Figure 4), and outputs from the evaluated models in Figure 3 .

4.4 Model Selection

The selection of the teacher-LLM was guided by the overall model score in the recent benchmark BenCzechMark [5]. The top three scoring models were: *DeepSeek-V3-0324*, *Llama-3.1-405B-Instruct*, and *DeepSeek-R1-0528* as per the leaderboard [3]. Additionally, we included a more recently released *gpt-oss-120B* which scored comparably to the DeepSeek models on the AIME benchmark [2] while having fewer parameters.

4.5 Model Evaluation

To evaluate which LLM produces the best quality outputs, we experimented with the pre-selected models using a set of statistical features complemented by human annotation. The evaluation sample comprises 10 randomly selected examples from each dataset, totaling to 80 examples.

The base score for the statistical features is derived from the length of the text. It is then penalized for instances of: structured text (e.g., lists), markup (Markdown and simple HTML), and abbreviations. The final score, presented in Table 4, is calculated as:

$$S_{stat} = max(0, base - 150 \cdot n_{structure} - 15 \cdot n_{markup} - 100 \cdot n_{abbreviations})$$

We evaluated four LLMs using a comparative preference-based procedure. For each test example, all models were reviewed side-by-side, and the best

Table 4: Evaluation of large language models on BenCzechMark, statistical measures, and human evaluation.

Model	θ B	BenCzechMark	S_{stat}	Points in Human Eval
DeepSeek-V3-0324	685	86.8	522.6	*249
Llama-3.1-405B-Instruct	406	85.7	593.5	166
DeepSeek-R1-0528	685	85.3	573.1	95
gpt-oss-120b	120	-	578.8	*240

response (single or multiple) was selected and awarded points. The results are summarized in Table 4.

Using the nonparametric Kruskal–Wallis H-test [28], we tested whether the differences in S_{stat} across models were statistically significant. The test yielded an H statistic of 2.77 with a corresponding p-value of 0.43, which exceeds the conventional significance threshold of 0.05. This indicates that the observed variations in performance between models are not statistically significant.

The human evaluation involved collecting judgments from human annotators for the examples across the four models. For every item, evaluators awarded a maximum of 5 points to the best model output, or, in the case of a tie, to more. The results in Table 4 show, that DeepSeek-V3 and gpt-oss-120b perform significantly better than the other models, but not significantly better than each other. Further, the Llama-3.1-405B-Instruct model performed significantly better than the DeepSeek-R1 model.

4.6 The Generated Dataset – SlamaHQTrain

At the time of publishing, the full dataset did not finish generation; however, in Table 5, we project the final dataset statistics based on the sample generated so far.

The generation will be performed on LUMI-G supercomputer using `vLLM` [20] for inference and serving and `ray` [23] for distributed scaling to multiple nodes. Each node is equipped with 4x AMD MI250x GPUs with performance and compute capabilities comparable to NVIDIA Ampere A100. Each GPU is split into two software GPUs, each with 64GB of available VRAM. This implies the total GPU capacity of 512GB for each node, equal to the size of the CPU RAM. So far we have managed to deploy `gpt-oss-120b` in a single-node setup due to it's impressive performance-to-parameters ratio and the presence of native MXFP4 quantization [33].

The deployment of larger models like DeepSeek V3 that require multiple nodes to properly fit into GPU VRAM remains challenging due to the experimental nature of LUMI-G hardware. The vLLM server is typically succesfully deployed, but we experience frequent of crashes during the actual inference which is attributed to ray mishandling communication between nodes, which is currently being investigated. Due to the impressive size of DeepSeek V3, at least two nodes are required in `fp8` setting. However, AMD MI250x does not have the

Table 5: Estimate of new tokens generated from the dictionaries. The tokenization was performed using the tokenizer for *gpt-oss-120b*. For the estimate, the ratio between the mean entry tokens and generated tokens was multiplied by the total tokens in each dictionary. The XML tags were kept in the entries before calculating the estimate.

Dictionary	$\bar{t}_{input\_xml}$	$\bar{t}_{input\_clean}$	$\bar{t}_{output}$	gen. tokens (est.)
PSJČ	230	119	231	42,784,297
SSJČ	125	62	168	26,524,891
Diderot	87	66	114	13,571,167
ASCS	79	66	174	7,028,755
SSC	92	41	136	6,516,028
SČS	46	28	177	3,695,512
SČFIS	158	114	237	2,903,796
SČFIN	147	87	235	1,245,715
Total	-	-	-	104,270,161

Table 6: Measured throughput for GPT OSS on LUMI using vLLM

Deployment Mode	Input Throughput (t/s)	Output Throughput (t/s)
Offline Batched	306.49	171.42
vLLM Server	247.16	153.89

hardware support to handle such data types natively, which means scaling to `bfloat16`, possibly doubling the required amount of nodes. In theory, quantization to MXFP4 should be possible, reducing the demand on number of nodes to one; however, such action requires full deployment of the original model and subsequent tuning on a calibration dataset. The usage GPT-Generated Unified Format (GGUF) [7] is also possible to achieve a single-node setup, but is highly experimental and unstable with vLLM.

We measured the expected throughput of `gpt-oss` using two different modes of execution - online serving and offline batched mode. While offline batched mode does not scale beyond one node, it eliminates the overhead introduced by networking. The results can be seen in Table 6.

5 Conclusions and Future Directions

In this paper, we have detailed the process of generating SlamaHQTrain, a new high-quality Czech pre-training dataset based on structured dictionary resources. We have presented the experiments with memory requirements of particular open-source LLMs, the generation prompt development process, and the evaluation of the generated texts using statistical metrics and human annotation. As a result, the SlamaHQTrain dataset is estimated to contain over 104 million tokens of high-quality educational Czech texts that will be used in further pre-training of Slama LLMs.

Acknowledgements. The described study comes from the project "On our own: Opportunities and Risks in the Individualization of Society (PRINS) CZ.02.01.01/00/23_025/0008710", which is co-financed by the European Union.

Computational resources were also provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Abdin, M., et al.: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone (2024), https://arxiv.org/abs/2404.14219
2. Aime 2025 benchmark leaderboard (2025), https://artificialanalysis.ai/evaluations/aime-2025
3. BenCzechMark Leaderboard (2025), https://huggingface.co/spaces/CZLC/BenCzechMark
4. Eldan, R., Li, Y.: TinyStories: How Small Can Language Models Be and Still Speak Coherent English? (2023), https://arxiv.org/abs/2305.07759
5. Fajcik, M., Docekal, M., Dolezal, J., Ondrej, K., Beneš, K., Kapsa, J., Smrz, P., Polok, A., Hradis, M., Neverilova, Z., et al.: BenCzechMark: A Czech-centric multitask and multimetric benchmark for large language models with duel scoring mechanism. Transactions of the Association for Computational Linguistics **13**, 1068–1095 (2025)
6. Filipec, J., et al.: Slovník spisovné češtiny). Academia, Praha (2005), 4. edition
7. Gerganov, G.: Gguf: General-purpose gpu-friendly unified format. https://github.com/ggerganov/llama.cpp (2023), accessed: 2025-11-12
8. Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., Song, D.: The false promise of imitating proprietary llms (2023), https://arxiv.org/abs/2305.15717
9. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Giorno, A.D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H.S., Wang, X., Bubeck, S., Eldan, R., Kalai, A.T., Lee, Y.T., Li, Y.: Textbooks are all you need (2023), https://arxiv.org/abs/2306.11644
10. Havránek, B., et al. (eds.): Příruční slovník jazyka českého (Reference Dictionary of Czech Language, PSJČ). Státní pedagogické nakladatelství/SPN, Praha (1935-1957), electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno
11. Havránek, B., et al.: Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJČ). Academia, Praha, 1st edn. (1960-1971), electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno
12. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
13. Horák, A., Vossen, P., Rambousek, A.: A distributed database system for developing ontological and lexical resources in harmony. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
14. Hu, T., Zhou, X.H.: Unveiling llm evaluation focused on metrics: Challenges and solutions. arXiv preprint arXiv:2404.09135 (2024)

15. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
16. Kim, S., Suk, J., Yue, X., Viswanathan, V., Lee, S., Wang, Y., Gashteovski, K., Lawrence, C., Welleck, S., Neubig, G.: Evaluating language models as synthetic data generators. arXiv preprint arXiv:2412.03679 (2024)
17. kolektiv autorů Ústavu pro jazyk český AV ČR: Slovník české frazeologie a idiomatiky. 2. díl: Výrazy neslovesné. LEDA, Praha (2009)
18. kolektiv autorů Ústavu pro jazyk český AV ČR: Slovník české frazeologie a idiomatiky. 3. díl: Výrazy slovesné. LEDA, Praha (2009)
19. Kraus, J., Petráčková, V., et al.: Akademický slovník cizích slov (Academic Dictionary of Foreign Words, SCS). Academia, Praha (1999), electronic version, LEDA, Praha
20. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
21. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are all you need ii: phi-1.5 technical report (2023), https://arxiv.org/abs/2309.05463
22. Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., Wang, H.: On llms-driven synthetic data generation, curation, and evaluation: A survey. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 11065–11082 (2024)
23. Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M.I., Stoica, I.: Ray: a distributed framework for emerging ai applications. In: Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation. p. 561–577. OSDI'18, USENIX Association, USA (2018)
24. Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A.: Orca: Progressive learning from complex explanation traces of gpt-4 (2023), https://arxiv.org/abs/2306.02707
25. Multisynt: an open multilingual synthetic dataset for llm pre-training (2025), https://openeurollm.eu/ai-factory-leonardo-multisynt
26. Openai cookbook (2025), https://cookbook.openai.com/
27. Pala, K., Všianský, J.: Slovník českých synonym (Dictionary of Czech Synonyms, SČS). Lidove Noviny Publishers, Praha, 3 edn. (2000)
28. SciPy Documentation: Kruskal-Wallis H-test (2025), https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html
29. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R.: The curse of recursion: Training on generated data makes models forget (2024), https://arxiv.org/abs/2305.17493
30. Tam, T.Y.C., Sivarajkumar, S., Kapoor, S., Stolyar, A.V., Polanska, K., McCarthy, K.R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., et al.: A framework for human evaluation of large language models in healthcare derived from literature review. NPJ digital medicine **7**(1), 258 (2024)
31. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)
32. Tommaso, G.: Českỳ korpus pro velkỳ jazykovỳ model. Master's thesis, České vysoké učení technické v Praze. Vypočetní a informační centrum. (2024)
33. Tseng, A., Yu, T., Park, Y.: Training LLMs with MXFP4. In: The 28th International Conference on Artificial Intelligence and Statistics (2025), https://openreview.net/forum?id=a8z5Q0WSPL

34. Velký slovník naučný: encyklopedie Diderot (Great Educational Dictionary: the Diderot Encyclopedia). Diderot, Praha (1999)
35. Yin, Y., Zeng, J., Li, Y., Meng, F., Zhang, Y.: Lexmatcher: Dictionary-centric data collection for llm-based machine translation (2024), https://arxiv.org/abs/2406.01441

Appendix A

Input, Prompt, and Output Example from the ASCS Dictionary

```
<root>
<h>alexandr</h>
<t>-a  m ⟨VJ⟩ zool.  papoušek rodu Psittacula, zelené barvy, pův. z tropické Asie a
 ↪  Afriky</t>
</root>
```

Fig. 2: Example of an XML dictionary entry from the ASCS dictionary.

```
[deepseek-v3:671b]
Alexandr je druh papouška patřící do rodu *Psittacula*, který se vyznačuje svým
výrazným zeleným opeřením. Původně pochází z tropických oblastí Asie a Afriky,
kde obývá zejména lesnaté a křovinaté biotopy. Tento pták je známý pro své živé
zbarvení a společenskou povahu, díky čemuž se často stává oblíbencem chovatelů
exotických druhů papoušků. Alexandr patří mezi středně velké druhy a v přírodě
hraje významnou roli při šíření semen, což přispívá k udržování ekologické
rovnováhy v jeho původním prostředí.

[llama3.1:405b]
Alexandr je označení pro papouška patřícího do rodu Psittacula, který se
vyznačuje svým převážně zeleným zbarvením. Tento druh ptáka pochází původně z
tropických oblastí Asie a Afriky. Vyznačuje se svými charakteristickými znaky a
chováním typickým pro papoušky. Alexandr patří mezi oblíbené exotické ptáky,
kteří jsou někdy chováni v zajetí pro svou pestrou barvu a inteligenci.

[deepseek-r1:671b]
Alexandr je v zoologii název pro papouška patřícího do rodu Psittacula, který
se vyznačuje převážně zeleným zbarvením peří. Tento pták, jehož genitivní tvar
zní "alexandra", pochází původně z tropických oblastí Asie a Afriky, kde je
přirozeně rozšířen. Termín se užívá jako odborné označení pro tento druh
papoušků, reflektující jejich taxonomické zařazení a geografický původ.

[gpt-oss-120b]
Alexandr je v zoologii podstatné jméno mužského rodu, označující papouška rodu
Psittacula. Jedná se o ptáka zelené barvy, jehož přirozený výskyt je původně v
tropické Asii a Africe. Termín je veden jako vlastní jméno (zkratka VJ) a v
odborných textech se používá k přesnému vymezení tohoto druhu papouška.
```

Fig. 3: The output produced by selected models for the input listed in Figure 2.

```
[Role a zadání]
Jsi asistent, který pomáhá generovat dataset vysoce kvalitních textů v českém jazyce, které budou sloužit
pro pre-training LLM. Dostaneš slovníkové heslo pocházející ze Slovníku Cizích Slov  ve formátu XML.
Převeď toto heslo do přirozeného, souvislého odstavce plynulého textu. Vyhni se jakémukoli formátování,
nadpisům nebo výčtům. Zajisti, aby byly v nějaké podobě zachyceny všechny významy, slovní spojení a
idiomatické výrazy uvedené ve slovníku, ať už převyprávěné, nebo zapracované do příkladů. Zachovej bohatství
významových odstínů a kulturních souvislostí. Nevynechávej vzácně používané nebo výrazově silné obraty,
ani když se zdají okrajové. Věnuj zvláštní pozornost zachování nízkofrekvenčních idiomů a kulturně
zabarvených výrazů, i kdyby to vyžadovalo metaforické odbočky. Zachovej a přirozeně zapracuj i morfologické
informace uvedené ve slovníkovém hesle (např. tvary, koncovky, rod, vid, pramen, slovní druh). Pokud
je v hesle uveden odkaz na jiné slovo nebo výraz (např. "viz ..."), uveď primárně informaci o této
významové podobnosti a neuváděj podrobnosti, které ve slovníkovém hesle nejsou. Rozsah výstupu přizpůsob
rozsahu slovníkového hesla - krátké záznamy s minimem informací zpracuj úsporně, bez zbytečného rozvíjení
nebo domýšlení.

[Formát vstupních dat]
Každý XML tag převeď do odpovídající části výstupního textu. Například:
<root>: Kořenový element celého slovníkového záznamu.
<h>: Heslové slovo, tedy hlavní slovo, které je vysvětlováno.
<t>: Text samotného příkladu použití, obvykle uvnitř <exm> nebo <other>.

[Styl výstupu]
Generovaný text má za úkol heslo a jeho významy popsat, informovat o něm, nikoliv vyprávět příběh.
Styl výsledného textu přizpůsob charakteru slovníkového hesla: Pokud výraz pochází z odborné oblasti
(např. medicína, právo, věda, technika), piš věcně, popisně a bez emocionální nebo metaforické
nadsázky. Zaměř se na přesné vysvětlení významu slova a jeho použití v běžné praxi. Používej neutrální
jazyk jako v encyklopedii nebo popularizačním článku. Nevymýšlej nové funkce, role nebo okolnosti,
které nejsou ve slovníku uvedeny. U výrazů knižních, archaických, lidových nebo poeticky zabarvených
můžeš volit styl volnější, literárnější nebo esejistický, ale i zde zachovej věcnou oporu ve významu.
Nepřidávej nové významy a drž se pouze těch uvedených ve slovníkovém záznamu. Vždy se drž významů
a příkladů uvedených ve vstupním hesle. Nevymýšlej vlastní použití, příklady nebo přirovnání. Text nemá
být fikcí, příběhem ani reklamou, ale popisným převyprávěním slovníkových informací. Nepoužívej znak '*'
pro formátování. Pozor na odlišnost českého jazyka, ve kterém bude odpověď generována, od jazyků
příbuzných. Tyto znaky se v češtině nevyskytují: Ĺ, Í, Ŕ, ŕ, Ä, ä, Ô, ô, Ą, ą, Ć, ć, Ę, ę, Ł, ł, Ń, ń,
Ó, ó, Ś, ś, Ž, ź, Ż, ż. Je důležité, abys nevymýšlel neexistující významy slov nebo jejich využití.

[Vysvětlivky zkratek]
pův. původně
v. viz
zool. zoologie

[Příklady]
Vstup: <root>
<h>turboalternátor</h>
<t>-u  m ⟨l⟩ eltech.  rychloběžný, zprav. dvoupólový alternátor s hladkým rotorem, poháněný turbínou</t>
</root>

Výstup: Turboalternátor je v elektrotechnice označení pro rychloběžný, obvykle dvoupólový alternátor
s hladkým rotorem, který je poháněn turbínou. Toto zařízení se využívá tam, kde je potřeba vysokých otáček
a efektivní přeměny mechanické energie na elektrickou, například v elektrárnách nebo průmyslových
aplikacích. Hladký rotor turboalternátoru umožňuje dosahovat vyšších rychlostí a snižuje riziko vibrací,
což z něj činí spolehlivý zdroj energie při vysokém výkonu.

Vstup: <root>
<h>empora</h>
<t>-y  ž ⟨ř⟩ archit.  ochoz nad postranními loďmi chrámu otevřený arkádou do hlavní lodi, galerie;
prostor v patře chrámu situovaný v záp. části lodi proti oltáři, kruchta, tribuna;  emporový  příd.</t>
</root>

Výstup: Empora je architektonický termín označující ochoz umístěný nad postranními loděmi chrámu,
který je otevřený arkádou do hlavní lodi. Tento prostor slouží jako galerie nebo tribuna a nachází
se obvykle v patře chrámu na západní straně lodi, naproti oltáři. Empora může být také označována
jako kruchta. Přídavné jméno „emporový" pak odkazuje na vlastnosti nebo prvky spojené s emporou. Tento
termín se užívá především v kontextu sakrální architektury a je spjatý s historickými stavbami, kde
sloužil jak pro liturgické účely, tak jako prostor pro shromáždění věřících nebo hudebníků.

[Úkol]
Nyní na základě uvedených instrukcí a příkladů generuj kvalitní text z následujícího slovníkového hesla.

Vstup: (see Fig. 2)
Výstup:
```

Fig. 4: The single-stage prompt composed for the XML dictionary entry from listed in Fig. 2.