LLMic: Building a Romanian Foundation Language Model

1st Vlad-Andrei Bădoiu
Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
vlad_andrei.badoiu@upb.ro

2nd Alexandru M. Gherghescu
Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
alexghergh@upb.ro

3rd Alexandru Agache
Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
alexandru.agache@upb.ro

4th Mihai-Valentin Dumitru
Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
mihai.dumitru2201@upb.ro

5th Costin Raiciu
Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
costin.raiciu@cs.pub.ro

***Abstract*—Recent advances in Large Language Models (LLMs) have shown remarkable capabilities across various tasks, with commercial models leading the way. While open models usually operate at smaller scales due to constraints on available corpora and hardware resources, they maintain competitiveness through specialization and fine-tuning. However, a significant challenge persists: the under-representation of low-resource languages in open datasets results in weak model capabilities in many languages.**

In this paper, we document the complete process of pretraining a foundation model for Romanian, a low-resource language, including corpus construction, architecture selection, and hyperparameter optimization. As part of this work, we introduce FuLG, a hundred-fifty-billion-token Romanian corpus extracted from CommonCrawl, alongside a 3-billion-parameter bilingual model, LLMic. Our evaluation shows that it is worthwhile to train language-specific models for specialized tasks, achieving results comparable to other much larger open and closed models. We show that fine-tuning LLMic for language translation after the initial pretraining phase outperforms existing solutions in the English-to-Romanian translation task. We hope through this work to advance the standing of the Romanian language in the world of LLMs.

I. Introduction

In recent years, Large Language Models (LLMs) adoption has increased tenfold. These models have demonstrated impressive capabilities across diverse applications, from code generation and natural language processing to image analysis. This adoption is sustained by an ever-growing ecosystem, with a variety of new open and proprietary models being released to the public at a fast pace. Leveraging hyperscalers' computing infrastructure, models such as GPT-4, Claude, and Gemini have attained the trillion-parameter threshold [1], achieving unprecedented performance for a variety of tasks [2], [3]. At very large scales, models become excellent general-purpose problem solvers, showing sophisticated reasoning capabilities and emergent behaviors such as knowledge transfer between languages, which enables effective multilingual processing.

Open-weights models, despite operating at significantly lower parameter counts, have a vibrant ecosystem [4]. Models such as Llama [5], Deepseek [6], Qwen [7], and Mistral [8] are able to be competitive with their much larger counterparts on a variety of tasks [9]. Open models particularly shine when they are specialized through fine-tuning for a specific task, to align in performance with their much larger counterparts. The advantages of open-weights models are numerous. Among them: data can be kept completely on-premises, operational costs can be reduced through infrastructure-optimized deployments, and results and research can be reproduced and shared.

At the heart of training open models lies the pretraining corpus. The research community has produced numerous data corpora. These are often derived from CommonCrawl [1], a public repository of crawled web pages, which has indexed more than 250 billion pages. Other corpora go further by including curated data such as books, social media discussions, and research papers. Trillion-token corpora such as Dolma [10], FineWeb [11], or RedPajama [12] have enabled the development of billion-parameter models with loss-optimal training [13], but they generally cover only a small fraction of languages.

A significant challenge in training language models at a lower parameter count lies in achieving consistent performance across diverse languages. Open-weights models, often constrained by their reduced parameter counts and English-centric pretraining corpora, typically underperform on low-resource languages compared to commercial alternatives [14]. This can be traced to public corpora such as Dolma [10] and FineWebEdu [11], targeting only widely spoken languages.

We thus observe that languages with limited representation often face challenges in the ecosystem around open-weight models. This limitation becomes apparent in language-specific benchmarks, and is also noticeable in real-world use,

[1] https://commoncrawl.org/

such as during conversations when retrieving language-specific knowledge. For instance, OLMo model family [15] support only six languages, while other initiatives such as LLM360 K2 [16] only support English. Even models developed by tech giants, such as Meta's Llama [17] or Google's Gemma [18], have limited output capabilities in less commonly spoken languages. Only recently have we seen some developments in this area, with Llama 3 supporting 30 languages [5], albeit with lower performance levels compared to English and limited information about their training.

While fine-tuning can enhance a model's capabilities in specific languages [19], [20], the results are underwhelming compared to the performance in languages that dominate the pretraining data. Recent research has shown that it is possible to classify an LLM's neurons into "language-specific" and "language-agnostic" categories [21], with language-specific neurons being only a tiny fraction of the entire model. The authors of [21] go as far as showing that, through manual intervention in the LLM's inference process, activations of language-specific neurons can be modified to steer the model towards a particular language. While this is promising, this is mainly a theoretical approach that has yet to yield results for practical applications.

Given this context, this paper focuses on improving the standing of the Romanian language in the open-weights LLM ecosystem. Romanian is, a language that represents approximately 0.6% of Common Crawl pages [22] and where existing open-weights models demonstrate limited capabilities [19]. To address this issue, we introduce an open Romanian pretraining corpus, resulting from filtering Romanian content from CommonCrawl. We built a pipeline that employs deduplication techniques, common signal filtering, and FastText for language detection, documenting the entire process of creating such a text corpus. The corpora is made available openly as FuLG[2], a 156B-token corpus with our tokenizer (589GB tokenized), or 220B tokens with the Llama 3 tokenizer. Alongside the corpus, we release LLMic[3], a 3B-parameter bilingual Romanian-English foundation model released under the Apache 2.0 License, and document its training process.

This paper presents a comprehensive study on pretraining a foundation model from scratch for the Romanian language. First, we document the construction of a pretraining corpus for the Romanian language—a component notably absent from state-of-the-art open corpora [10]–[12], [23], [24]. Next, we dive into the process of selecting training a bilingual model, covering aspects such as model architecture, optimization of hyperparameters, and development of an efficient training configuration. To showcase the usefulness of language specific models such as LLMic, we fine-tune the model for translation tasks and evaluate its performance against existing solutions, including other fine-tuned models for the Romanian language.

II. Building a State-of-the-Art Romanian Corpus

A crucial factor in model performance is the size, quality and diversity of the pretraining corpus. These datasets typically comprise content from various sources, primarily web pages, due to their sheer number and availability, followed by social media discussions, academic papers, books, code repositories, and encyclopedias, covering a wide range of topics. The main sources of data are either private crawls of the Internet or the publicly available CommonCrawl repository.

While information about closed models' training data is scarce, even open-weights models like Llama or Gemma lack transparency regarding their training datasets, as it is a central part of the competitive advantage. We are thus in a peculiar scenario, where the best open-source models aren't truly open-source when it comes to pretraining data (aptly named open-weights models), which deeply affects any smaller initiative and open innovation.

We sourced the main pretraining corpus for Romanian, named FuLG, from CommonCrawl, an online collection of crawled internet web pages dating back to 2007. CommonCrawl releases snapshots of portions of the Internet several times annually, each snapshot exhibiting low similarity rates between releases, thus enabling us to consistently extract new data from each snapshot. We used snapshots from 2013 to May 2024. Given the petabytes scale of data to process, we ran a data processing pipeline in a distributed environment with multiple nodes for data acquisition. For deduplication and quality filtering, we used a single large memory node.

a) Data Acquisition: To process CommonCrawl snapshots, we leveraged the CCNet pipeline [25]. CCNet facilitates distributed processing of snapshots, including downloading in the WET format, language identification via the FastText algorithm [26], and deduplication of common paragraphs. We faced two challenges with CCNet:

1) Since CCNet development has stalled, we quickly hit roadblocks related to package versions and environment-specific problems with our distributed cluster setup, which needed numerous modifications to the source code.
2) We ran the pipeline on a SLURM-managed cluster, but with several constraints. We faced limitations on maximum supported job sizes and limited batch array sizes. We also faced limitations related to checkpointing. To overcome these issues, we updated CCNet to work under these constraints.

In the first phase, we filtered out documents in Romanian using FastText's language identification. We retained documents with a FastText language score above 0.5.

b) Deduplication: For quality filtering and deduplication, we relied on code from RedPajama [12]. Exact deduplication reduced the dataset size by 37%, while fuzzy deduplication with a 0.8 threshold further reduced it by 50%.

c) Content Filtering: Next, we introduced a content filtering step to our pipeline. Using a regex-based approach, we filtered HTML and Javascript artifacts that are sometimes present in the WET files. To filter potentially controversial

[2] https://huggingface.co/datasets/faur-ai/fulg
[3] https://huggingface.co/faur-ai/LLMic

content, we utilized a dictionary, removing documents containing specific words in either the content or URL. For Personally Identifiable Information (PII), we replaced phone numbers, email addresses, and links with special tokens.

d) Quality Filtering: As with deduplication, we utilized existing code from RedPajama to compute a set of quality signals. We extended the code to filter documents based on the quality signals introduced by Gopher [27], reducing the dataset size by 50%. This resulted in a final disk size of 589GB. In terms of tokens, we obtain 156B tokens using our tokenizer, or 220B tokens using Llama 3 tokenizer which has a much lower compression ratio for Romanian.

The thresholds and rules we employed for filtering are as follows:

- Fraction of characters in the most common n-grams exceeded: 0.2 for 2-grams, 0.18 for 3-grams, 0.16 for 4-grams, 0.15 for 5-grams, 0.14 for 6-grams, 0.13 for 7-grams, 0.12 for 8-grams, 0.11 for 9-grams, 0.10 for 10-grams
- Fewer than 50 words or more than 100,000 words
- Median word length less than 3 or greater than 10 characters
- More than 90% of lines start with bullet points
- More than 30% of lines end with ellipses
- Less than 30% of lines end with punctuation

The data acquisition process spanned several months, while deduplication and filtering were completed in less than 18 hours on a single large-memory node.

III. Pretraining Data

Data scarcity presents a fundamental challenge in training a language model for the Romanian language. Acquiring hundreds of billions of tokens requires extensive filtering and cleaning of CommonCrawl's petabyte-scale dataset as shown in the previous section. Obtaining additional curated content often necessitates applying optical character recognition (OCR) to public domain documents.

The corpus for training LLMic consists of a mixture of filtered web data and curated sources in both Romanian and English. The data is split into 300B tokens for Romanian and 700B tokens for English, as shown in Table I. For the English subset of the corpus, we leveraged FineWebEdu [11] due to its LLM-assisted quality filtering, and extended it with a curated set of documents from Dolma [10]—forum discussions, books, and research papers. For Romanian, we use the entirety of FuLG, extended it with publicly available curated content, discussions, and parallel Romanian-English documents. A detailed overview of the corpus is discussed for the remainder of this section.

Source	**Size**
Romanian (300B)	
Web Sources	621 GB
Discussions, Curated & Parallel	10 GB
English (700B)	
FineWebEdu	1.3 TB
Dolma Subset	109 GB

TABLE I: Dataset Composition by Language and Source

Web Sources. We leverage two filtered CommonCrawl sources for Romanian language data: i) FuLG, presented in the previous section which contains 220B tokens when processed with the Llama tokenizer; ii) mC4 [28], a multilingual cleaned version of CommonCrawl, comprising 42B Romanian tokens. While they share the source, the two corpora differ in their language detection algorithms, filtering and deduplication techniques. To remove duplicates resulted from the merge, we apply both fuzzy and exact de-duplication using the RedPajama [12] pipeline. We further augment our dataset by incorporating filtered content from recent CommonCrawl dumps (2024-22, 2024-26, 2024-30, and 2024-33):

a) ***Curated Data****:* We extend our Romanian web corpus with curated data from multiple sources, including the Romanian Wikipedia and public documents from Romanian institutions. For the English portion, we utilize a curated subset of Dolma [10], specifically incorporating high-quality content from books, Wikipedia, and research papers.

b) ***Discussions****:* Our corpus includes discussions from public forums. For Romanian content, we utilized forum data from the Pushshift [29] dataset. The English forum discussions were sourced from Dolma.

c) ***Parallel Data****:* We incorporated parallel Romanian-English data from multiple sources, including translated official documents from the European Union and the ParaCrawl [30] parallel corpus.

IV. Tokenizer

The tokenizer plays a crucial role in both model inference speed and the overall capabilities of the model [31]. A tokenizer that can express the text in fewer tokens results in faster inference, since fewer tokens are needed during the autoregressive decoding process.

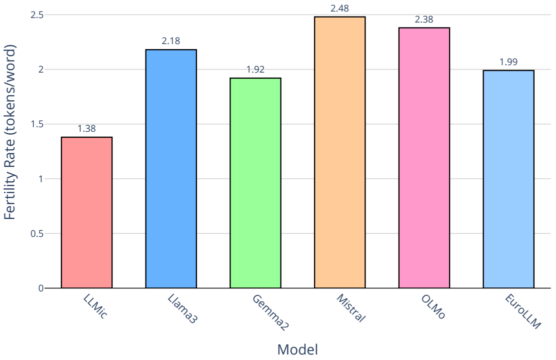

Fig. 1: Tokenizer fertility analysis across different languages and text sources. The graph shows the relationship between input text and resulting tokens.

To address this limitation, we built a BPE tokenizer based on the GPT-NeoX architecture [32]. The tokenizer was trained on 7 billion tokens, sourced with equal distribution between FineWebEdu [11] and our Romanian corpus. A design choice we made early one was to use an uncased vocabulary of 128,000 tokens and intentionally omits the Romanian diacritics: ă, â, î, ș, ț, replacing them with a, i, s, t respectively. This decision is based on ablation studies where diacritics degraded the model performance. The resulting tokenizer achieved a fertility rate of 1.38 for Romanian text, representing a substantial improvement over existing tokenizers, as shown in Figure 1.

V. Architecture

The architecture of LLMic is a classical decoder-only transformer [33]. Following an approach similar to OLMo [15], we surveyed the leading families of models at this time, including Gemma [34], OLMo [15]) and Llama2 [35]) through ablation studies. Based on this, we selected Llama2 as the underlying architecture. All the details about LLMic can be consulted in Table IIa. We use grouped query attention (GQA) [36], for faster inference speed, rotary positional embeddings (RoPE) [37] to enable context extension and SiLU activation [38].

Parameter	Value
Sequence Length	2048
Number of Layers	24
Embedding Size	2,560
FFN Hidden Size	10,240
Number of Heads	20
Number of KV Heads	5
Activation Function	SiLU
Position Encodings	RoPE ($\Theta = 500{,}000$)
Layer Norm	RMSNorm ($\epsilon = 10^{-5}$)
Tied Embeddings	No

(a) Model Architecture

Parameter	Value
Batch Size (per GPU)	8
Warmup	3000 steps
Gradient Accumulation	2
Sequence Length	2048
Weight Decay	0.1
Learning Rate Scheduler	Cosine with Min LR
Learning Rate	4×10^{-4}
Adam β_1	0.9
Adam β_2	0.95
Adam ϵ	1×10^{-5}

(b) Training Hyper-parameters

TABLE II: Model Specifications and Training Parameters

VI. Training

We analyzed existing models' training regimes, as well as conducted learning rate ablations, in order to find the optimal set of hyper-parameters for training. For the most part, we followed the simple approach of increasing the learning rate as much as possible until the model diverges, then slightly lowering it and finding the best stable batch size.

We experimented both with constant learning rates, as well as cosine-decay learning rates. We've tried learning rates as high as 4×10^{-3}, though we did not find these stable enough (as evidenced by the model's outputs becoming overly repetitive and showing signs of text memorization). Subsequently, we implemented a cosine-decay learning rate scheduler with a maximum learning rate of 4×10^{-4}, and a minimum learning rate 10 times lower. The complete set of hyper-parameters is detailed in Table IIb.

For distributed training, we implemented a custom framework on top of Hugging Face's Transformers [39][4]. The training process utilized Fully Sharded Data Parallel (FSDP) [40] parallelism with full sharding. Additionally, we used `bfloat16` mixed precision training alongside the optimized Liger kernels [41].

The training procedure followed a multi-phase approach. During the initial 50B tokens, we used a 50:50 split between Romanian and English, as we reasoned the model should first learn both the languages equally. Subsequently, we used the 30:70 Romanian-English split, as dictated by the language ratio in the corpus. Towards the end of the pretraining phase, we reused a high-quality subset of Romanian documents multiple times.

VII. Evaluation

In this section, we provide a comprehensive evaluation of our two main contributions: the FuLG corpus and the LLMic model. We evaluate both the quality of our data curation efforts and the effectiveness of training a bilingual model.

A. FuLG Corpus Evaluation

To evaluate FuLG, we conducted ablation studies using a 1B decoder-only model based on OLMo [15], with a sequence length of 2048 and the OLMo tokenizer trained on the OSCAR-Ro [42] corpus. We used a global batch size of 256. We made several changes to the hyper-parameters to adjust to the 1B model size: weight decay of 0.001 and a max learning rate of 4e-4 annealed by a cosine schedule, with an initial warmup of 1000 steps.

We trained three identical models on different pretraining corpora: Oscar [42], mC4 [43] and FuLG. We trained to completion on each corpus (290k steps for FuLG, 75k for mC4 and 20k for OSCAR). Since FuLG was much bigger than both OSCAR and mC4 (about 4 times bigger than mC4), we also included an earlier checkpoint from FuLG at 70k steps. We do note that this early checkpoint slightly underperforms, but we believe this naturally happens because of the cosine decay learning rate schedule, as the learning rate is still quite high at this point.

Perplexity. A commonly employed method to assess the quality of a pretraining corpus is to fix a model, train it with different corpora, and measure perplexity against a curated evaluation set. While perplexity is not a definitive measure of corpus quality [10], it is a good indicator if something is off. We constructed a perplexity dataset covering multiple domains, sourced from Wikipedia, news articles, textbooks, research papers, and books, totaling 74M tokens. Although

[4] https://github.com/nets-cs-pub-ro/RoptimusNet

Dataset	Perplexity
FuLG (290k)	16.06
FuLG (70k)	21.00
mC4	15.38
OSCAR	23.57

TABLE III: Perplexity comparison across different pretraining corpora.

we did not specifically perform a separate decontamination step, we do note that the collection methods of the evaluation dataset were manual, trying to use data that succeed the date of the CommonCrawl snapshots used to source the three corpora. Before computing perplexity, the evaluation dataset was processed using the *clean-text* Python package [5]. The results are shown in Table III. Both ablations of FuLG show perplexity values similar to those of existing corpora, confirming that we are on the right track.

Dataset	Grammar	Creativity	Complexity
FuLG (290k)	8.5	7.1	4.6
FuLG (70k)	7.25	5.875	4.0
mC4	7.5	5.2	3.1
OSCAR	6.3	4.2	2.8

TABLE IV: Results for the story generation task.

Qualitative Evaluation Inspired by the TinyStories [44] work, we used the models to generate stories in Romanian from given prompts and asked GPT-4 to rate the creativity, grammar, and overall complexity of each story. We only considered responses that were coherent stories, discarding anomalous outputs. In Table IV, we present our findings, which suggest that FuLG may enable better performance for specific tasks. We note that both OSCAR and the early checkpoint of FuLG (70k) generated significantly more anomalies than the other ablations, showing the need for more and better training data.

B. LLMic Model Evaluation

We developed LLMic as a foundation model optimized for Romanian language processing tasks. As models at this size typically used for specific tasks, we evaluated its performance on English-to-Romanian translation by fine-tuning the model and testing it on the WMT [45] translation benchmark. Since LLMic operates on uncased text without diacritics, we modified the WMT data by removing diacritics and converting to lowercase, while leaving everything else unchanged.

The results, presented in Table Vb, show that LLMic outperforms both established open models like LLaMA-2 [35], Mistral [8], and Gemma [18], as well as previous work on fine-tuning open models for Romanian [19], confirming that small specialized models are competitive in the world of LLMs, as well as the need for language specific training corpora.

Method	Score
Full Precision (FP16)	41.01
GPTQ INT8 (calibrated on C4)	41.01
GPTQ INT4 (calibrated on C4)	40.29
BitsAndBytes INT4	40.44
LoRA (r = 8)	37.05

(a) Quantization Methods Performance

Model	Score
LLMic	41.01
mBART [46]	38.50
Llama-3.1-8B-Instruct	29.02
RoMistral-7b-Instruct	27.70
RoLlama3-8b-Instruct	27.31
Mistral-7B-Instruct-v0.2	26.19
RoGemma-7b-Instruct	25.96
Gemma-1.1-7b-it	25.48

(b) Model Performance

TABLE V: Performance Comparison of Quantization Methods and Language Models

A common approach to reduce the size of the model is quantization. This reduces the memory footprint and enables the use of lower precision arithmetic, which can achieve much higher numbers of floating point operations per second on newer GPU architectures. For the translation task, we noticed only a marginal impact on performance, while being able to reduce the model size significantly as shown in Table Va.

VIII. Related Work

Open corpora are essential components for truly open models and play a vital role in democratizing future LLMs. Currently, numerous open corpora exist, many derived from CommonCrawl. Notable examples include: Dolma [10] (3T tokens), C4 [47] (175B tokens), The Pile [48] (387B tokens), ROOTS [49] (400B tokens), RefinedWeb [50] (600B tokens), RedPajama v2 [12] (30T tokens), FineWeb [11] (15T tokens), Zyda [51] (1.3T tokens), LLM360 Amber [16] (1.2T tokens), MAP-Neo [52] (4.5T tokens), OSCAR [42]. While these corpora provide sufficient quality and size for English centric models, they often lack adequate representation of less commonly spoken languages. For example, after training a Romanian GPT-NeoX tokenizer on the Romanian part of OSCAR, we obtain a number of only 10B tokens on OSCAR, and 41B tokens on mC4 [43] (slightly higher number of tokens if using an English tokenizer, similar to Llama or OLMo, due to worse compression ratio). Considering the scaling laws of optimal loss [13] during training given corpus size and model size, these quantities are insufficient for optimal LLM training. Beyond these, several smaller corpora exist for various Natural Language Processing (NLP) tasks in Romanian [53]–[58], but they generally contain fewer than 500M tokens, which is far from adequate for LLM development.

Fine-tuning has been employed to adapt models for specific languages [19], [20]. However, fine-tunning yield un-

[5] https://github.com/jfilter/clean-text

derwhelming results compared to the models' performance in their original pretraining language, A more resource-intensive approach is to pretrain the model on a corpus with multiple languages [5], [14], but the performance is often degraded for the less represented languages — a problem that is particularly pronounced in smaller models.

IX. Discussion and Future Work

The availability of high-quality corpora for less commonly spoken languages is crucial for the democratization of LLMs. While proprietary models often demonstrate proficiency across a wide range of languages, open models frequently underperform in this aspect. By developing large, high-quality open corpora for diverse languages, we can expand the open ecosystem around LLMs.

FuLG represents an initial effort to improve pretraining corpus size and quality for the Romanian language, while LLMic shows that it makes sense to look into training smaller models for language specific tasks. There are multiple ways to build upon this research, with interesting tracks being:

- Data processing: We currently work with Common Crawl snapshots in WET format. However, as previous research indicates [11], using a better HTML parser can improve both the quality of extracted text and the quantity of usable data.
- Language-specific optimization: Romanian language particularities could be leveraged to adapt and refine quality filters. Currently, we use thresholds designed for the English language, but developing language-specific criteria and identifying new criteria for the Romanian language may yield better results after filtering for quality.
- Enhanced instruction dataset for Romanian: We plan to leverage our translation model to translate existing instruction datasets (for tasks such as summarization and sentiment analysis) into Romanian, advancing the state-of-the-art for Romanian NLP tasks.

X. Conclusion

In this paper, we document the complete process of training a bilingual large language model for Romanian, from constructing a comprehensive Romanian language corpus to training a 3B parameter LLM. Our evaluation shows that developing language-specific models is worthwhile and that expanding open corpora to include underrepresented languages yields significant benefits. Through fine-tuning for English-Romanian translation tasks, we achieve state-of-the-art performance with a substantially smaller parameter count than existing approaches. This efficiency enables us to bootstrap additional datasets and models for Romanian by effectively transferring resources from English, creating a foundation for further development. Through this work we hope to advance the representation of Romanian in the open LLM ecosystem.

Acknowledgment

We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call. Some of the experiments were performed on the Luxembourg national supercomputer MeluXina. The authors gratefully acknowledge the LuxProvide teams for their expert support. We acknowledge the computational resources provided by the PRACE award granting access to Discoverer in SofiaTech, Bulgaria. The authors gratefully acknowledge the HPC RIVR consortium and EuroHPC JU for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science. We acknowledge VSB – Technical University of Ostrava, IT4Innovations National Supercomputing Center, Czech Republic, for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (grant ID: 90254). This research was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906. UPB authors were partly supported by VMWare gift funding.

References

[1] H. Hesse. (2024) Gpt-4 architecture, datasets, costs and more leaked. [Online]. Available: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*.

[4] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[6] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[7] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[9] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez *et al.*, "Chatbot arena: An open platform for evaluating llms by human preference," *arXiv preprint arXiv:2403.04132*, 2024.

[10] L. Soldaini, R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar *et al.*, "Dolma: an open corpus of three trillion tokens for language model pretraining research," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 725–15 788.

[11] G. Penedo, H. Kydlíček, A. Lozhkov, M. Mitchell, C. A. Raffel, L. Von Werra, T. Wolf *et al.*, "The fineweb datasets: Decanting the web for the finest text data at scale," *Advances in Neural Information Processing Systems*, vol. 37, pp. 30 811–30 849, 2024.

[12] M. Weber, D. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams *et al.*, "Redpajama: an open dataset for training large language models," *Advances in neural information processing systems*, vol. 37, pp. 116 462–116 492, 2024.

[13] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 30 016–30 030.

[14] W. Ali and S. Pyysalo, "A survey of large language models for european languages," *arXiv preprint arXiv:2408.15040*, 2024.

[15] D. Groeneveld, I. Beltagy, E. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. Jha, H. Ivison, I. Magnusson, Y. Wang *et al.*, "Olmo: Accelerating the science of language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 789–15 809.

[16] Z. Liu, A. Qiao, W. Neiswanger, H. Wang, B. Tan, T. Tao, J. Li, Y. Wang, S. Sun, O. Pangarkar *et al.*, "Llm360: Towards fully transparent open-source llms," *arXiv preprint arXiv:2312.06550*, 2023.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[18] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[19] M. Masala, D. Ilie-Ablachim, A. Dima, D. G. Corlatescu, M.-A. Zavelca, O. Olaru, S.-M. Terian, A. Terian, M. Leordeanu, H. Velicu *et al.*, ""vorbești românește?" a recipe to train powerful romanian llms with english instructions," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 11 632–11 647.

[20] G. L. Garcia, P. H. Paiola, L. H. Morelli, G. Candido, A. C. Júnior, D. S. Jodas, L. Afonso, I. R. Guilherme, B. E. Penteado, and J. P. Papa, "Introducing bode: A fine-tuned large language model for portuguese prompt-based task," *arXiv preprint arXiv:2401.02909*, 2024.

[21] T. Tang, W. Luo, H. Huang, D. Zhang, X. Wang, X. Zhao, F. Wei, and J.-R. Wen, "Language-specific neurons: The key to multilingual capabilities in large language models," *arXiv preprint arXiv:2402.16438*, 2024.

[22] Common Crawl, "Language detection on CommonCrawl datasets," Common Crawl Foundation, 2024, accessed: 2024-01-09. [Online]. Available: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

[23] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023.

[24] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora *et al.*, "Datacomp-lm: In search of the next generation of training sets for language models," *arXiv preprint arXiv:2406.11794*, 2024.

[25] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4003–4012.

[26] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[27] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2021.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[29] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 830–839.

[30] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn *et al.*, "Paracrawl: Web-scale acquisition of parallel corpora." Association for Computational Linguistics (ACL), 2020.

[31] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How good is your tokenizer? on the monolingual performance of multilingual language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3118–3135.

[32] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang *et al.*, "Gpt-neox-20b: An open-source autoregressive language model," *arXiv preprint arXiv:2204.06745*, 2022.

[33] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[34] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[35] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[36] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

[37] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[38] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks*, vol. 107, pp. 3–11, 2018.

[39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," *EMNLP 2020*, p. 38, 2020.

[40] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer *et al.*, "Pytorch fsdp: Experiences on scaling fully sharded data parallel," *Proceedings of the VLDB Endowment*, vol. 16, no. 12, pp. 3848–3860, 2023.

[41] P.-L. Hsu, Y. Dai, V. Kothapalli, Q. Song, S. Tang, S. Zhu, S. Shimizu, S. Sahni, H. Ning, and Y. Chen, "Liger kernel: Efficient triton kernels for llm training," *arXiv preprint arXiv:2410.10989*, 2024.

[42] P. J. Ortiz Suárez, B. Sagot, and L. Romary, "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures," ser. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen, and C. Iliadi, Eds. Mannheim: Leibniz-Institut für Deutsche Sprache, 2019, pp. 9 – 16. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[44] R. Eldan and Y. Li, "Tinystories: How small can language models be and still speak coherent english?" *arXiv preprint arXiv:2305.07759*, 2023.

[45] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz *et al.*, "Findings of the 2016 conference on machine translation (wmt16)," in *First conference on machine translation*. Association for Computational Linguistics, 2016, pp. 131–198.

[46] Y. Liu, "Multilingual denoising pre-training for neural machine translation," *arXiv preprint arXiv:2001.08210*, 2020.

[47] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[48] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima *et al.*, "The pile: An 800gb dataset of diverse text for language modeling," *arXiv preprint arXiv:2101.00027*, 2020.

[49] H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. Villanova del Moral, T. Le Scao, L. Von Werra, C. Mou, E. González Ponferrada, H. Nguyen *et al.*, "The bigscience roots corpus: A 1.6 tb composite multilingual dataset," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 809–31 826, 2022.

[50] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023.

[51] Y. Tokpanov, B. Millidge, P. Glorioso, J. Pilault, A. Ibrahim, J. Whittington, and Q. Anthony, "Zyda: A 1.3 t dataset for open language modeling," *arXiv preprint arXiv:2406.01981*, 2024.

[52] G. Zhang, S. Qu, J. Liu, C. Zhang, C. Lin, C. L. Yu, D. Pan, E. Cheng, J. Liu, Q. Lin *et al.*, "Map-neo: Highly capable and transparent bilingual large language model series," *arXiv preprint arXiv:2405.19327*, 2024.

[53] T. Váradi, S. Koeva, M. Yamalov, M. Tadić, B. Sass, B. Nitoń, M. Ogrodniczuk, P. Pęzik, V. B. Mititelu, R. Ion *et al.*, "The marcell legislative corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 3761–3768.

[54] V. B. Mititelu, D. Tufiş, and E. Irimia, "The reference corpus of the contemporary romanian language (corola)," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[55] R. Ion, E. Irimia, D. Stefanescu, and D. Tufis, "Rombac: The romanian balanced annotated corpus." in *LREC*, 2012, pp. 339–344.

[56] L. Midrigan-Ciochina, V. Boyd, L. Sanchez-Ortega, D. Malancea_Malac, D. Midrigan, and D. P. Corina, "Resources in underrepresented languages: Building a representative romanian corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 3291–3296.

[57] C. Schöch, T. Erjavec, R. Patras, and D. Santos, "Creating the european literary text collection (eltec): Challenges and perspectives," *Modern Languages Open*, 2021.

[58] M. Manolescu and Ç. Çöltekin, "Roff-a romanian twitter dataset for offensive language," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 895–900.