INFORMATICA, 2025, Vol. 36, No. 2, 385–406

DOI: https://doi.org/10.15388/25-INFOR592

Open Llama2 Models for the Lithuanian Language

Artūras NAKVOSAS*, Povilas DANIUŠIS, Vytas MULEVIČIUS

Neurotechnology, Laisvės av. 125A, LT-06118, Vilnius, Lithuania
e-mail: arturas@neurotechnology.com, povilasd@neurotechnology.com, vytas.mulevicius@neurotechnology.com

Received: September 2024; accepted: April 2025

Abstract. In this paper, we focus on the problem of whether efficient Lithuanian large language models (LLMs) can be achieved from Llama2 LLMs, which lack Lithuanian-specific components. Although the Llama2 architecture was previously successfully utilised to derive various regional LLMs, we propose and describe the first open Llama2 LLMs for the Lithuanian language (7 and 13 billion parameter versions), an accompanying question/answer (Q/A) dataset, and translations of popular language understanding benchmarks (Arc, Belebele, Hellaswag, MMLU, TruthfulQA, and Winogrande), which contribute to the standardisation of Lithuanian LLM evaluation. We empirically evaluate the proposed models by investigating their perplexity and performance in the translated language understanding benchmarks. The perplexity experiments show that it decreases consistently during pretraining, reflecting enhanced next-token prediction capabilities. Benchmarking the proposed LLMs against language understanding tasks reveals that high-quality pretraining datasets may be essential to achieve models that perform efficiently on these benchmarks. Comparison of the proposed LLMs with the latest open multilingual LLM shows that our model with 13 billion parameters is ranked 4th of 8 models in tasks such as Arc, Hellaswag, and Winogrande, but is generally outperformed in other tasks. These benchmarks allow us to hypothesise that from recent LLMs more efficient Lithuanian language models can be derived in the future. The complete realisations of the LLMs and other contributed components are available in the accompanying open repository https://huggingface.co/neurotechnology.

Key words: Llama2, Regional LLMs, LLMs for the Lithuanian language.

1. Introduction

Large language models (LLMs), relying on the Transformer architecture proposed by Vaswani *et al.* (2017) have shown remarkable effectiveness in many natural language processing (NLP) tasks (Minaee *et al.*, 2024; Naveed *et al.*, 2024). This has primarily been fuelled by increasingly large model parameterisations and training datasets, which are deemed essential according to neural scaling laws (Hernandez *et al.*, 2022). On the other hand, with the consistent advancement of computational linguistics and NLP, open LLMs such as Llama2 (Touvron *et al.*, 2023), Mistral (Jiang *et al.*, 2023), Mixtral (Jiang *et al.*, 2024), Falcon (Almazrouei *et al.*, 2023) were released. The performance characteristics of

*Corresponding author.

these open models are comparable with their commercial counterparts. Such LLMs usually require massive datasets and considerable computational resources. For example, the pretraining of the Llama2 family was carried out using a 2 trillion token set and required 3311616 GPU hours (Touvron *et al.*, 2023). In addition to direct applications, these models can be further trained for various downstream problems (Minaee *et al.*, 2024), including regional language modelling.

For this application, it is important to note that open LLMs are usually trained with largely English texts (e.g. Touvron *et al.* (2023) indicate that $> 89\%$ of the dataset, which was used to pretrain Llama2 consisted of English texts), resulting in a lack of performance for less common languages. Although commercial LLMs usually better support underrepresented languages, as a rule, they are exposed only via APIs, which do not provide access to the model's parameters or its intermediate representations. Since there are ≈ 380 non-English languages with at least 1 million speakers (Eberhard *et al.*, 2021), open regional LLMs constitute an important research direction and there have been multiple recent attempts to achieve efficient open LLMs, tailored for various regional languages (see Section 2, Table 1).

Regional LLM training is a challenging technical task not only computationally but also from the perspective of training data, which should reflect a rich structure of the language of interest, local cultural nuances, and domain-specific knowledge in multiple areas. Although this is only partially solved by massive multilingual datasets, such as CulturaX (Nguyen *et al.*, 2023), it is still an important challenge to collect representative datasets in regional languages.

Open LLMs are also potentially useful for NLP research as their internal mechanism is fully transparent. There are also related applications outside the scope of NLP. For example, successful regional LLMs can significantly impact areas such as education, public services, healthcare, and cultural preservation.

This article describes Neurotechnology's[1] contribution to regional LLM research, consisting of

- Llama2-based 7 and 13 billion parameter LLMs for the Lithuanian language, and their empirical evaluation;
- A new dataset, consisting of 13,848 Q/A pairs primarily about Lithuania and Lithuanian history (in the Lithuanian language);
- Translations of popular LLM benchmarks to the Lithuanian language;
- Open repository, containing all the mentioned components.

In this article, we investigate whether efficient Lithuanian LLMs can be achieved from Llama2 LLMs (which do not have the Lithuanian component, according to Touvron *et al.*, 2023). In our opinion, for this research the Llama2 architecture is potentially advantageous against other similar open LLMs without Lithuanian language support (e.g. Mistral), since it allows experimentation with different model sizes, and its 13 billion parameter version nearly matches the performance of Mistral, as shown by Jiang *et al.* (2023).

[1]Neurotechnology (http://www.neurotechnology.com) is a Lithuanian company, specialising in artificial intelligence, biometrics, computer vision, and deep neural networks.

We structure our paper by starting with a short review of the related work in Section 2. Section 3 describes the proposed LLMs and other contributed components, and Section 4 is devoted to an empirical evaluation. Finally, the conclusive Section 5 summarises the research conducted from different perspectives.

2. Related Work

Llama2 model. Transformer-based Llama2 is available in different parameter sizes (e.g. 7, 13, and 70 billion parameters). The model is first pretrained using a 2 trillion token set, collected from public sources, and utilising a self-supervised autoregressive approach with cross-entropy loss. Afterward, it is fine-tuned using publicly available instruction datasets, augmented with human-annotated data, and Reinforcement Learning with Human Feedback (RLHF) methodologies (Touvron *et al.*, 2023).

This model can support the maximum context length of the 4096 tokens. According to benchmarks, Llama2 generally performs on par with various open alternatives (e.g. Falcon (Almazrouei *et al.*, 2023), Mistral (Jiang *et al.*, 2023) and Mixtral (Jiang *et al.*, 2024)), which also may be advantageous in specific scenarios. For example, Falcon is recognized for its strong performance at higher parameter counts, and Mistral/Mixtral are generally lighter-weight models that emphasize efficiency and specialized use cases. Compared to these models, Llama2 aims to maintain a balance between robust performance and scalability. As is common with large foundational models, it can be further successfully tuned for various downstream tasks, including regional language modelling.

LLMs for regional languages. Table 1 summarises LLMs tailored for common European languages, reflecting the recent contributions from the research and engineering community working in this direction. We include only those regional LLMs, that meet the following criteria:

- The model should be published in an open repository (e.g. Hugging Face[2]),
- It should contain at least a minimal description (architecture, training data, and other details).

According to Table 1, open LLMs are released for the majority of common European languages. Table 1 shows that Llama2 and Mistral are the leading architectures for open LLMs for regional European languages, and 7 billion parameter models are the most common. Table 1 also reveals that full parameter training is conducted in the majority of cases (19 cases from 20), instead of the parameter-efficient fine-tuning (PEFT) based approach. However, in some instances (2 cases from 20) regional LLMs were trained using PEFT methods, such as LoRA (Hu *et al.*, 2022), which may result in less accurate models compared to full-parameter training, although with the lower computational costs. In addition, quite often only the model itself is published (11 out of 20 cases), without an accompanying citable document (e.g. technical report/peer-reviewed publication), or training and

[2] https://huggingface.co/

Table 1
Open LLM models for regional European languages. The F/P column denotes whether the model was full-parameter trained (F), or trained via PEFT (P), and Doc. column shows whether the corresponding model has an accompanying publication.

Language and reference	Architecture	Size	F/P	Doc.
Bulgarian (INSAIT, 2024)	Mistral	7B	F	No
Danish (Mabeck, 2024)	Mistral	7B	F	No
Dutch (Rijgersberg, 2024)	Mistral	7B	F	No
French-English (Faysse *et al.*, 2024)	Llama	1.3B	F	Yes
German (Plüster and Schuhmann, 2024)	Llama2	7B,13B	F	No
Greek (SPAHE, 2024)	Mistral	7B	F	No
Hungarian-English (Csaki *et al.*, 2024)	Llama2	7B	F	Yes
Finnish and other (LumiOpen, 2024)	Llama2	7B–33B	F	No
Icelandic (Snæbjarnarson *et al.*, 2022)	RoBERTa		F	Yes
Italian (Bacciu *et al.*, 2023)	Llama2	7B,13B	P	Yes
Lithuanian (**Ours**)	Llama2	7B,13B	F	Yes
Norwegian (Norallm, 2024)	Mistral	7B	F	No
Serbian, Bosnian, Croatian (Gordić, 2024)	Mistral	7B	F	No
Spanish (Projecte AINA, 2024)	Falcon	7B	F	No
Swedish (Ekgren *et al.*, 2023)	GPT-SW3	126M–40B	F	Yes
Slovenian (Ulčar and Robnik-Šikonja, 2021)	RoBERTa		F	Yes
Polish (Speakleash, 2024)	Mistral	7B	F	No
Ukrainian (Boros *et al.*, 2024)	Mistral	7B	F	Yes
Portuguese (Garcia *et al.*, 2024)	Phi-2B	1.3B–7B	P	Yes
Romanian (Masala *et al.*, 2024)	Llama2	7B	F	Yes

evaluation datasets. In our opinion, the lack of accompanying scientific documentation limits the potential usefulness of the released regional LLMs in various important aspects, including their reproducibility, empirical performance assessment, and establishing a connection to the existing related results.

Multilingual LLMs that support the Lithuanian language. Another way of achieving LLMs with regional language support is to train models for multiple languages simultaneously. Although this approach requires much more computational and data resources (for instance, EuroLLM was pretrained using 256 Nvidia H100 GPU's and 4 trillion token set, as indicated by Martins *et al.*, 2024), compared to learning models only for single language, recent open LLMs that support Lithuanian language are multilingual (e.g. Llama3.X (Grattafiori *et al.*, 2024), and Gemma2 (Riviere and et al., 2024), and EuroLLM). Although these LLMs perform quite similarly on various benchmarks, there are applications in which some of these models are advantageous against other counterparts. For example, Gemma2 is potentially more suitable for general knowledge and reasoning tasks, Llama3.1 is efficient in coding and complex problem-solving tasks, and EuroLLM is optimised for European languages. All these multilingual models with Lithuanian language support were published in later stages or after our research and currently represent state-of-the-art (SOTA) in the field of open LLMs.

Table 2
Overview of Llama-based LLMs.

Model name	Description
Llama2-7B	A second-generation Llama foundational language model with 7 billion parameters (no Lithuanian language support).[3]
LT-Llama2-7B	Proposed Lithuanian LLM derived from Llama2-7B model, according to information, provided in Section 3.
Llama2-13B	A second-generation Llama foundational language model with 13 billion parameters (no Lithuanian language support).[4]
LT-Llama2-13B	Proposed Lithuanian LLM derived from Llama2-13B model, according to information, provided in Section 3.

3. Proposed Open LLMs and Accompanying Components

Proposed open LLMs and their training details. We trained the proposed LLMs (including tokenizers) from Llama2-7B and Llama2-13B, respectively (Table 2).

The training follows a standard two-step approach, consisting of autoregressive pretraining and supervised fine-tuning, which schematically is depicted in Fig. 1.

Autoregressive pretraining was performed on the Lithuanian component of the CulturaX dataset (Nguyen *et al.*, 2023). It is the most intensive step computationally (Table 3), and corresponds to the integration of the Lithuanian language into the model. During this step, the cross-entropy loss for the next token prediction task was minimised (hence, no labelled data are required for pretraining). The complete set of model's parameters was optimised (i.e. no PEFT was used). Figure 2 shows the loss during the pretraining process. From it we see that, although loss minimisation tends to saturate in the end, one may hypothesise that the learning would continue for more than one epoch.

Figure 6 (Appendix A) shows the distribution of the source of the Lithuanian component of the CulturaX dataset. From it we see that this dataset is quite rich in quantity. However, it is collected mainly from common web sites. Figure 7 (Appendix A) shows the distribution of the length of the record in tokens. In order to speed up pretraining, we reduced the context length to 2048 tokens (reflected in the peak near 2048, in Fig. 7 (Appendix A)). See Table 3 for more details on the pretraining process.

Supervised fine-tuning (SFT) explicitly guides the pretrained model toward task-specific outputs using labelled data. It is much less computationally intensive, since the model already has the Lithuanian language integrated (Table 3). We conducted SFT using the Alpaca dataset (Dubois *et al.*, 2024), which has been translated into Lithuanian using the ChatGPT (gpt-4-1106-preview) and dataset (Neurotechnology, 2024). SFT tunes the LLMs to process formatted prompts `"[INST] «SYS» {system_level_instruction} «/SYS»{instruction}[/INST]"`, where parameter `system_level_instruction` sets desired behaviour constraints (e.g. tone, response style), and parameter `instruction` specifies task (see the caption of Table 9

[3] https://huggingface.co/meta-llama/Llama-2-7b
[4] https://huggingface.co/meta-llama/Llama-2-13b

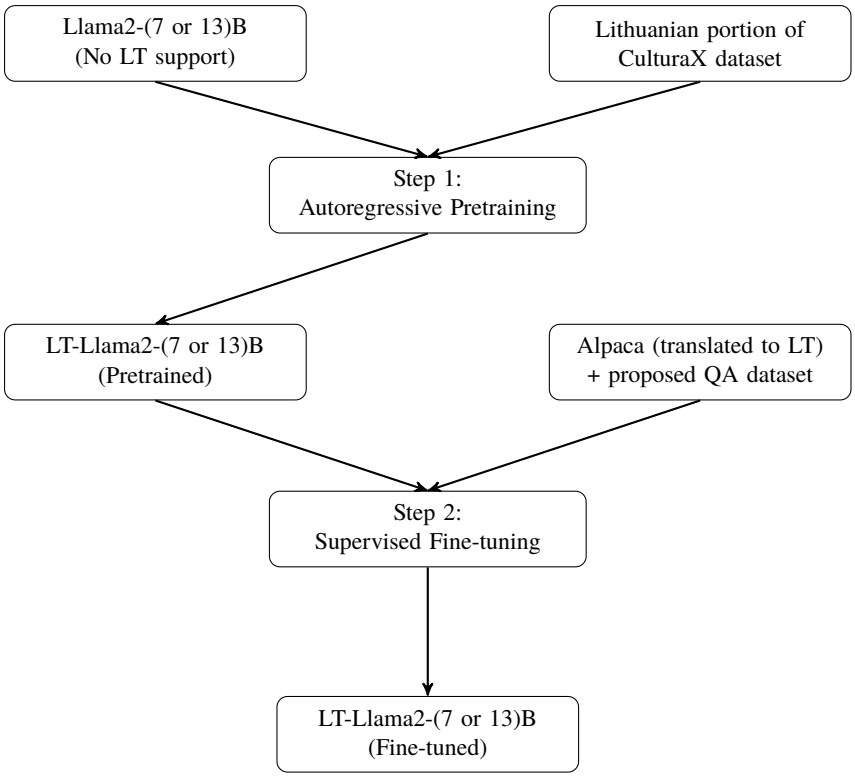

Fig. 1. Overview of the two-step process for creating LT-Llama2-7B/LT-Llama2-13B.

(Appendix A), for an example). SFT was conducted with the same parameters as in Table 3, except for the learning rate, which was set to 0.00001, and the context length was restored to 4096.

Hyperparameters, such as learning rate, warmup ratio, weight decay provided in Table 3 were selected according to Touvron *et al.* (2023) guidelines, but slightly adjusting the values provided to ensure faster and more stable loss minimisation. During pretraining we also observed gradient exploding effects. We mitigated them by tuning gradient accumulation steps (see Table 3).

Table 10 (Appendix A) provides text generation examples (pretrained models), and Table 9 (Appendix A) provides examples of answers to questions (pretrained and fine-tuned models). If not stated otherwise, in all experiments and benchmarks with the proposed LLMs, we used pretrained-only models, which corresponds to the common practice. The download links for the proposed LLMs are provided in Table 8 (Appendix A).

Proposed open Q/A dataset. This dataset was constructed from the ChatGPT (OpenAI *et al.*, 2024) summarisations of a subset of Lithuanian Wikipedia using the procedure described below. First, the Lithuanian Wikipedia was downloaded and the titles of its pages were filtered with the following prompt: `"I will provide a list of titles in Lithuanian language. From the list provide me the`

Table 3
Hyperparameters and other details.

Learning parameter	Llama2-7B	Llama2-13B
Number of epochs	1	1
Learning rate	0.0002	0.00004
Warmup ratio	0.05	0.05
Weight decay	0.07	0.05
Per-device batch size	8	4
Gradient accumulation steps	2	4
Duration of pretraining in hours for a single H100 GPU	1722.0	2980.5
Duration of fine-tuning in hours for a single H100 GPU	< 1	< 1
Total number of tokens	14761219995	
Records in dataset	13339785	
Mean number of tokens per record	1106.5560	
Standard deviation of tokens per record	697.0089	
Optimiser	AdamW	
Hardware	8xH100 GPUs	

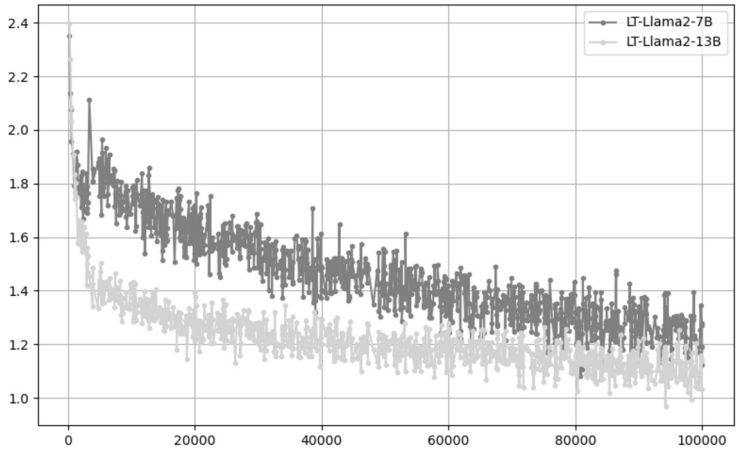

Fig. 2. Losses (y-axis) vs training steps (x-axis) during the model's pretraining.

`titles without any explanation which are directly or indirectly related with Lithuania except fauna and flora. List: {list}`", where variable `list` represents a list of titles of all pages. After this filtering and manual check, the resulting list of Lithuanian Wikipedia pages represented by pair (`title`, `text`) was transformed into Q/A pairs using the prompt, returned by Algorithm 1.

The proposed dataset consists of 13,848 such pairs, and represents various facts about Lithuania and Lithuanian history. Note that it was not used in the pretraining process. Table 4 presents a set of examples from the proposed Q/A dataset, which can be accessed through the download links provided in Table 8 (Appendix A).

Proposed open translations of language understanding benchmarks. Language model evaluation harness (LMEH, Gao *et al.*, 2023) are language understanding benchmarks

Algorithm 1 Generate prompt for Q/A summarisation

Require: Title of the page (`title`), text of the page (`text`)

```
wordCount ← word_count(text) // number of words in a given text
if wordCount > 400 then
    count ← min(10, wordCount/200)
    P := "Iš duoto teksto sudaryk {count}, kuo įvairesnių klausimų atsakymų porų, kuriems nereikėtų konteksto. Atsakymai turi būti išsamūs. Pirmas klausimas turi būti tiesiogiai susijęs su teksto tema {title}. Tekstas: {text}"
else
    P := "Iš duoto teksto sudaryk klausimo-atsakymo porą. Teksto tema: {title}. Tekstas: {text}"
end if
return P
```

Table 4
Examples from the accompanying Q/A dataset.

Question	Answer
Koks yra Vilniaus miesto statusas Lietuvoje?	Vilnius yra Lietuvos sostinė.
Kur yra Gedimino pilis?	Gedimino pilis yra Vilniuje, ant Gedimino kalno.
Kas buvo vadinamas „Lito tėvu“?	Vladas Jurgutis buvo vadinamas „Lito tėvu“, nes jam buvo patikėta spręsti visus naujos valiutos įvedimo niuansus.
Kokios upės teka per Vilnių?	Per Vilnių teka Neris ir Vilnia.
Kada buvo įkurtas Vilniaus universitetas?	Vilniaus universitetas buvo įkurtas 1579 metais, Vilniuje, po Lietuvos didžiojo kunigaikščio Stepono Batoro privilegijos suteikimo jėzuitų ordino kolegijai.
Kada ir kur įvyko Žalgirio mūšis?	Žalgirio mūšis įvyko 1410 m. liepos 15 d. netoli Tanenbergo ir Griunvaldo (Žalgirio) kaimelių, dabartinės Lenkijos teritorijoje, į pietvakarius nuo Olštyno.

which are created for the evaluation of LLMs across a wide range of tasks. LMEH includes a set of popular LLM evaluation benchmarks:

- Arc (Lai *et al.*, 2023) benchmark consists of multiple choice science questions at school level.
- GSM8K (Cobbe *et al.*, 2021) benchmark consists of linguistically diverse mathematical problems.
- Hellaswag (Zellers *et al.*, 2019) benchmark consists of common-sense inference challenge dataset.
- Massive multitask language understanding (MMLU) (Hendrycks *et al.*, 2021) benchmark covers different tasks from a diverse set of academic disciplines and is designed to measure the accuracy of the model in a multitask setting.

- Truthful-qa (Lai *et al.*, 2023) benchmark is designed to measure whether an LLM is truthful in generating answers to questions that span different categories (health, law, finance, and politics).
- Winogrande (Sakaguchi *et al.*, 2019) is a set of pronoun resolution problems originally designed to be unsolvable for statistical models that are based on selectional preferences or word associations.

These benchmarks produce prompts consisting of question and answer options, and evaluate the accuracy of the responses of LLMs. The accuracy can be measured conveniently because of the structured prompt, which asks the LLM to select an option (e.g. "a", "b" or "c"). We translated the LMEH benchmarks into Lithuanian using GPT-4. The download links are provided in Table 8 (Appendix A).

4. Empirical Evaluation

4.1. *Perplexity During Pretraining*

We analysed the LLMs by examining their perplexity (measured on the proposed Q/A dataset), which is defined as

$$P(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i \mid w_{<i})\right), \tag{1}$$

where

- $W = w_1, \ldots, w_N$ is the sequence of tokens,
- $p(w_i \mid w_{<i})$ is the conditional probability of the token w_i given all the previous tokens $w_{<i}$ (if $i = 1$, probability $p(w_1 \mid w_{<1})$ is defined as $p(w_1)$).

The perplexity can be interpreted as the model's ability to predict the next token, given the previous ones. From the definition (Eq. (1)), the lower perplexity values indicate better performance, and for any input sequence W, $P(W) \geqslant 1$. The selection of input perplexity was motivated by Gonen *et al.* (2023), where the authors reveal that for a wide range of tasks, the lower the perplexity of the prompt, the better the prompt can perform the task.

We investigated the association of average perplexity (averaged over all Q/A concatenations from the proposed Q/A dataset), and the percentage of data from CulturaX Lithuanian component, exposed to the model in the pretraining process. We conducted this experiment using both LT-Llama2-7B and LT-Llama2-13B models, measuring perplexity every 10% of the total number of iterations during a pretraining epoch. Figure 3 reveals that with the inclusion of additional pretraining data, perplexity tends to decrease, although, in the end, increasing saturation is visible in both cases. The initial and final perplexities in Table 5 reflect the integration of the Lithuanian language component into the proposed Llama2 models. Note that the proposed Q/A dataset was not used in the pretraining.

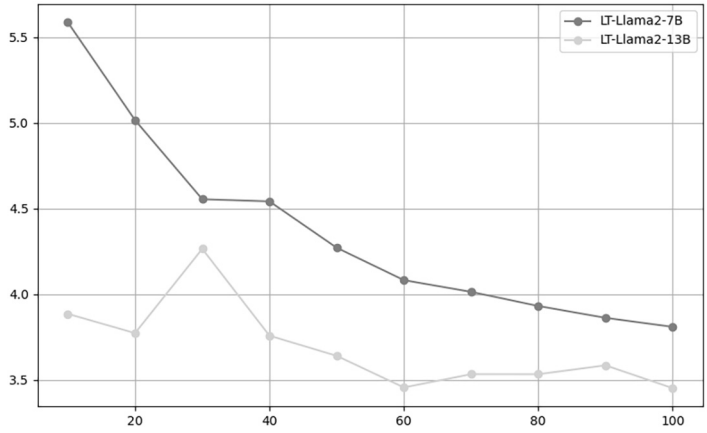

Fig. 3. Percentage of the Lithuanian component of the CulturaX dataset used in the pretraining (x-axis) vs. corresponding average perplexity (y-axis).

Table 5
Average perplexities before (Llama2-7B and Llama2-13B) and after (LT-Llama2-7B and LT-Llama2-13B) pretraining.

Model	Average perplexity
Llama2-7B	17.4613
LT-Llama2-7B	3.8096
Llama2-13B	13.8849
LT-Llama2-13B	3.4520

4.2. *Language Understanding During Pretraining*

We evaluated the proposed open LLMs with the proposed open translations of LMEH benchmarks, using the same scheme as in perplexity experiments.

Figures 4 and 5 showcase the accuracies for a sequence of checkpoints, which correspond to the percentage of the pretraining data from CulturaX Lithuanian component, starting with 0% (which corresponds to the initial Llama2-7B), with the step of 10%. Similarly, Fig. 8 (Appendix A) and Fig. 9 (Appendix A) provide information about individual benchmarks from the MMLU set.

Although for some tasks (e.g. `Arc`, `Hellaswag`, `Winogrande`) we see consistent improvement throughout the entire pretraining process, this benchmark surprisingly reveals that in most cases of MMLU (see Fig. 8 (Appendix A) and Fig. 9 (Appendix A)), there is no improvement compared to the initial model. We hypothesise that this may be because the Lithuanian component of CulturaX is almost exclusively collected through web crawling of common websites (see Fig. 6 (Appendix A)), which does not include data that is relevant to those specific tasks. Therefore, extension of regional components of CulturaX with high-quality data may improve LLMs, tailored for the corresponding regional languages.

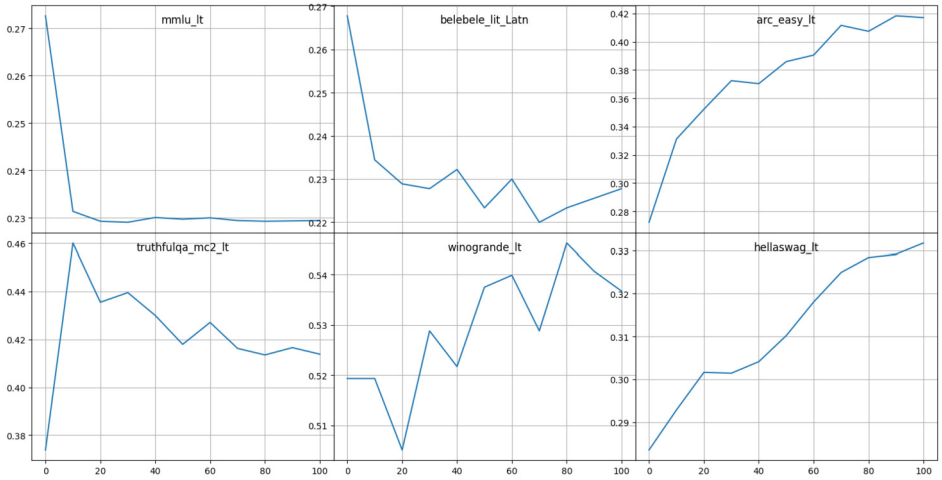

Fig. 4. Accuracies (y-axis) of LMEH benchmarks for LT-Llama2-7B model, pretrained with different proportions of Lithuanian component of CulturaX dataset (x-axis). The MMLU benchmarks are summarized in mmlu_lt.

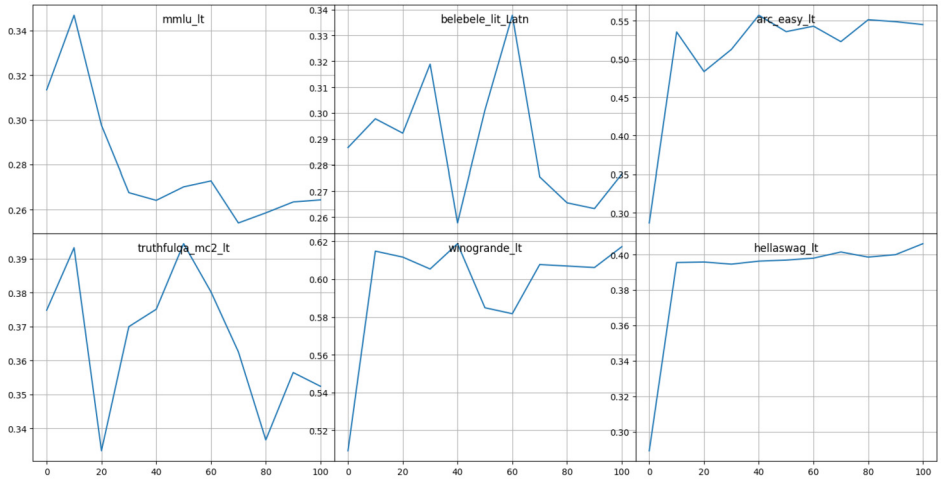

Fig. 5. Accuracies (y-axis) of LMEH benchmarks for LT-Llama2-13B model, pretrained with different proportions of Lithuanian component of CulturaX dataset (x-axis). The MMLU benchmarks are summarized in mmlu_lt.

4.3. *Comparison with Recent Multilingual LLMs that Support the Lithuanian Language*

Language understanding benchmarks. We compared the proposed LLMs and the latest multilingual models that support the Lithuanian language (Gemma2, EuroLLM, Llama3.1, and Llama3.2) in the translated LMEH benchmarks. We provide these evaluations in Table 6. Table 7 summarises Table 6 by reflecting the rankings of the proposed LT-Llama2-7B and LT-Llama2-13B LLMs in these benchmarks. It shows that in 3 of 6

Table 6
Accuracies of LLMs in LMEH benchmarks.

Model	MMLU	Arc	Winogrande	TruthfulQA	Hellaswag	Belebele
Gemma2-27B	64.82	77.4	66.77	42.06	50.82	89.22
Gemma2-9B	60.09	68.31	65.15	39.69	45.32	86.78
EuroLLM-9B	51.95	71.55	64.17	42.13	46.32	69.44
Llama3.1-8B	44.86	48.65	54.22	37.61	35.19	67.56
Gemma2-2B	35.84	45.45	51.85	54.78	34.8	52.44
Llama3.2-3B	36.41	39.39	51.85	38.87	31.51	46.22
LT-Llama2-13B	26.44	54.5	61.72	35.23	40.61	27.67
LT-Llama2-7B	26.01	43.18	53.67	41.38	33.17	27.23

Table 7
Rankings of the proposed LLMs in LMEH benchmarks (1 means that the model was the most accurate, and 8 means that it was the least accurate).

Model	MMLU	Arc	Winogrande	TruthfulQA	Hellaswag	Belebele
LT-Llama2-13B	7	4	4	8	4	7
LT-Llama2-7B	8	7	6	4	7	8

benchmarks (`Arc`, `Hellaswag`, and `Winogrande`), our LT-Llama2-13B model was ranked as 4th of 8, although the more recent SOTA open LLMs generally performed better than our models.

Quality of text generation. The paper by Kapočiūtė-Dzikienė *et al.* (2025) provides an empirical evaluation of recent LLMs (GPT-4o, Llama3.1, Gemma2, and ours). Based on their findings, our LT-Llama2-13B model outperformed its competitors (GPT-4o, Llama 3.1, and Gemma2) in benchmarks for text generation quality in the Lithuanian language. It achieved an error rate of 0.98%, and the closest competitor (GPT-4o) achieved an error rate of 3.44%. However, in a benchmark of the accuracy of the answer, the other models were more accurate than ours.

5. Conclusions

We presented the first Llama2-based open LLMs tailored especially for the Lithuanian language, the accompanying Q/A dataset, and the translated LMEH benchmarks, which contribute to the standardisation of the evaluation of Lithuanian language models.

We also provided an overview of the existing LLMs for common European languages. It shows that most regional models follow the Llama2 or Mistral architecture. In addition, some authors do not train a full parameter set, but instead rely on PEFT approaches, which are less computationally demanding but also potentially less efficient in performance. On the other hand, PEFT methods partially allow one to retain the original parameter structure, and thereby they may be beneficial for achieving more efficient regional LLMs from the perspective of language understanding benchmarks. Our findings also reveal a lack of scientific documentation of the published open regional LLMs.

We evaluated the proposed LLMs based on perplexity and translated LMEH benchmarks. During the pretraining epoch, we evaluated average perplexities (measured with independent dataset) every 10% of the training iterations. These benchmarks show that perplexity decreases consistently during pretraining, reflecting enhanced next-token prediction capabilities. The initial and final perplexities (17.4613 versus 3.8096 for LT-Llama2-7B and 13.8849 versus 3.4520 for LT-Llama2-13B) show the integration of the Lithuanian language component in the proposed Llama2 models. Using the same scheme, we also evaluated our models with the translated LMEH set, which includes a conceptually diverse set of language model benchmarks. The results of these experiments hint that the Lithuanian component of CulturaX may not be sufficiently rich for modern LLM architectures. Although we positively answer the question of whether efficient Lithuanian LLMs (which were non-existent at the beginning and during most of this research) can be achieved from Llama2 LLMs, which lack Lithuanian components, the latest open multilingual models (Llama3.1, Llama3.2, Gemma2, and EuroLLM) already have a strong Lithuanian component. According to our benchmarks, these open SOTA LLMs generally performed better than our models, however, the proposed LT-Llama2-13B was ranked average (4/8) in half of the LMEH benchmarks. This also leads to the hypothesis that by deriving Lithuanian LLMs from these recent models, one may obtain more efficient Lithuanian LLMs. In our opinion, the good performance of our model in the external benchmark by Kapočiūtė-Dzikienė *et al.* (2025) may be due to the fact that it was trained in a single language and the other LLMs were multilingual.

In the context of regional LLMs, the proposed models open up further research perspectives not only for NLP, but also for other directions, since LLM representations are potentially useful in various scenarios (e.g. sentiment analysis (Zhang *et al.*, 2024), robotics (Kim *et al.*, 2024)). The important limitations of our contribution are related to the rapid progress of LLM research, leading to the continuous emergence of more advanced models. In addition, we used automatically translated and generated data in the contributed components, which may also cause negative effects. To achieve stable loss minimisation during pretraining we faced and solved several challenges related to the selection of hyperparameters (learning rates, batch size, gradient accumulation steps). Our future work will include fully trained small language models tailored for Baltic languages and English.

A. Appendix

Table 8
Download links for proposed LLMs and data.

URL	Description
https://tinyurl.com/3vrjt5u3	Proposed LLM LT-Llama2-7B (pretrained version).
https://tinyurl.com/236mab8b	Proposed LLM LT-Llama2-7B (pretrained and fine-tuned version).
https://tinyurl.com/bdzcae84	Proposed LLM LT-Llama2-13B (pretrained version).
https://tinyurl.com/2wr9npfh	Proposed LLM LT-Llama2-13B (pretrained and fine-tuned version).
https://tinyurl.com/5y88x7ym	Proposed open Q/A dataset.
https://tinyurl.com/3srtmv46	LT-Arc is the Lithuanian translation of Arc dataset (Lai *et al.*, 2023), which consists of a set of genuine grade-school level, multiple-choice science questions assembled to encourage research in advanced question-answering.
https://tinyurl.com/3s2khf7f	LT-GSM8K is a Lithuanian translation of GSM8K dataset (Cobbe *et al.*, 2021) that consists of linguistically diverse mathematical problems.
https://tinyurl.com/bdzcayrw	LT-Hellaswag is a Lithuanian translation of Hellaswag benchmark (Zellers *et al.*, 2019), consisting of a common sense inference challenge dataset.
https://tinyurl.com/38w2m94c	LT-MMLU is a Lithuanian translation of MMLU benchmark (Hendrycks *et al.*, 2021) that measures a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, history, computer science, law, and more.
https://tinyurl.com/5e53mmjh	LT-Truthful-qa is a Lithuanian translation of Truthful-qa benchmark (Lai *et al.*, 2023) to measure whether an LLM is truthful in generating answers to questions that span different categories (health, law, finance and politics).
https://tinyurl.com/2s5wcatx	LT-Winogrande is a Lithuanian translation of Winogrande dataset (Sakaguchi *et al.* (2019)), which consists of expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations.

Table 9
Examples of answers from the proposed LLMs (pretrained and fine-tuned versions) to the given prompts. The prompt format: `[INST] «SYS» Pateik lakoniška, bet išsamų atsakymą į klausimą. «/SYS» Question given in the first column of this table [/INST].`

Question	LT-Llama2-7B	LT-Llama2-13B
Kuo užsiima bendrovė Neurotechnology?	Bendrovė Neurotechnology yra įmonė, kuri specializuojasi neuroninių tinklų ir dirbtinio intelekto srityse, teikianti paslaugas įvairioms pramonės šakoms ir organizacijoms.	Bendrovė Neurotechnology yra dirbtinio intelekto ir kompiuterinės regos technologijų kūrėja, teikianti sprendimus robotikai, kompiuterinei regai, mobiliosioms aplikacijoms ir kitoms pramonės šakoms.
Kokių svarbiausių tarptautinių organizacijų narė yra Lietuva?	Lietuva yra Europos Sąjungos (ES) narė nuo 2004 m. ir NATO narė nuo 2004 m.	Lietuva yra Jungtinių Tautų, Europos Sąjungos, NATO, Ekonominio bendradarbiavimo ir plėtros organizacijos, Pasaulio prekybos organizacijos ir kitų tarptautinių organizacijų narė.
Ką žinai apie lietuvių kalbą?	Atsiprašau, bet aš negaliu pateikti konkrečios informacijos apie lietuvių kalbą, nes neturiu prieigos prie jūsų pateiktos informacijos. Jei norite sužinoti apie lietuvių kalbą, galite kreiptis į patikimus šaltinius, pavyzdžiui, Lietuvių kalbos institutą arba Lietuvių kalbos žodyną.	Lietuvių kalba yra viena iš dviejų išlikusių gyvų gyvų baltų kalbų, priklausanti indoeuropiečių kalbų šeimai. Ji yra oficiali Lietuvos Respublikos kalba ir turi apie 3,5 mln. kalbančiųjų. Lietuvių kalba turi turtingą žodyną, įskaitant daugiau nei 5000 skirtingų žodžių, ir yra žinoma dėl savo archajiškumo bei išlaikytų archajiškų formų. Lietuvių kalba taip pat yra svarbi Lietuvos kultūros ir tapatybės dalis, ir ji yra įtraukta į UNESCO pasaulio paveldo sąrašą.
Koks yra 50 km/h važiuojančio automobilio greitis?	50 km/h yra maksimalus greitis, kurį automobilis gali pasiekti važiuodamas 100 km/h greičiu. Tai yra maksimalus greitis, kurį automobilis gali pasiekti, kai jis važiuoja 100 km/h greičiu.	50 km/h važiuojantis automobilis yra 50 km per valandą.

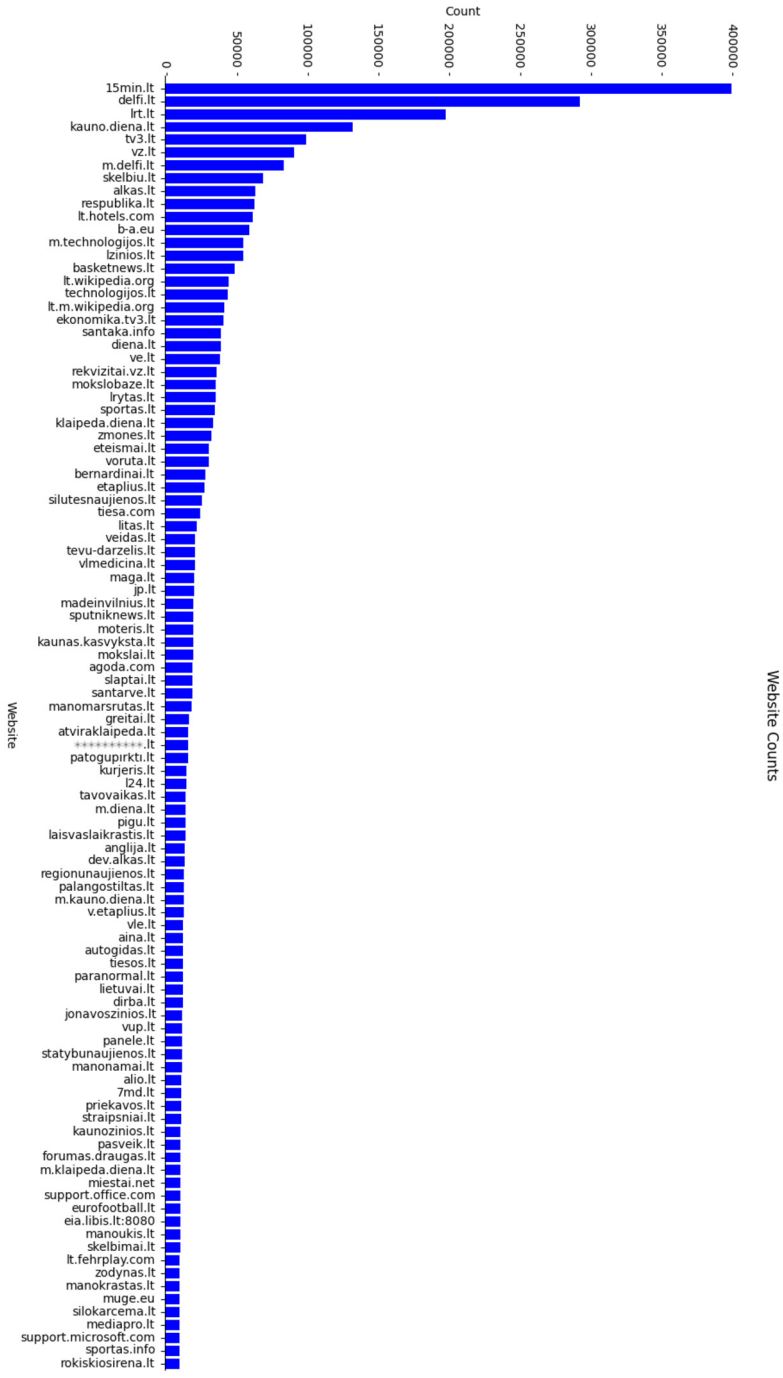

Fig. 6. Source distribution of the Lithuanian component of the CulturaX dataset.

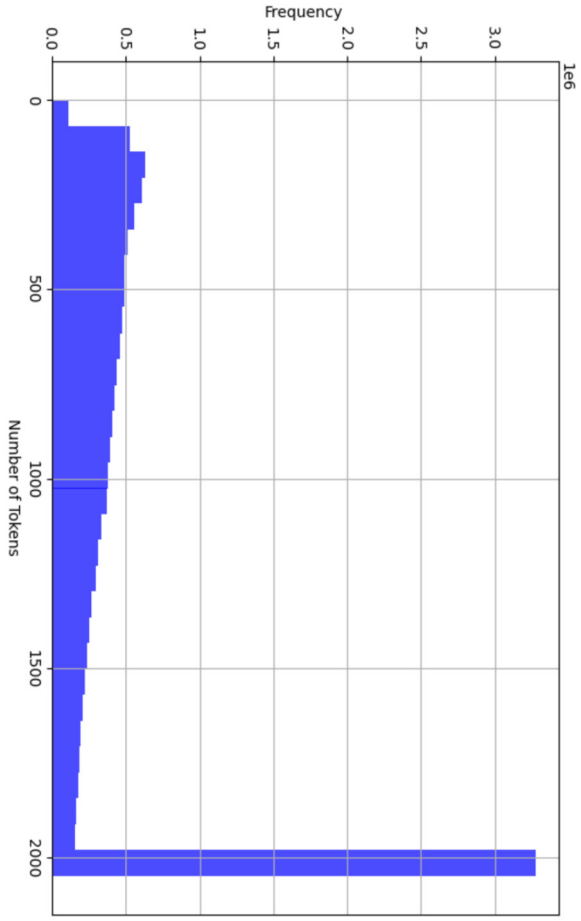

Fig. 7. Distribution of the record length of the Lithuanian component of the CulturaX dataset (in tokens), which was used in pretraining of the proposed LLMs.

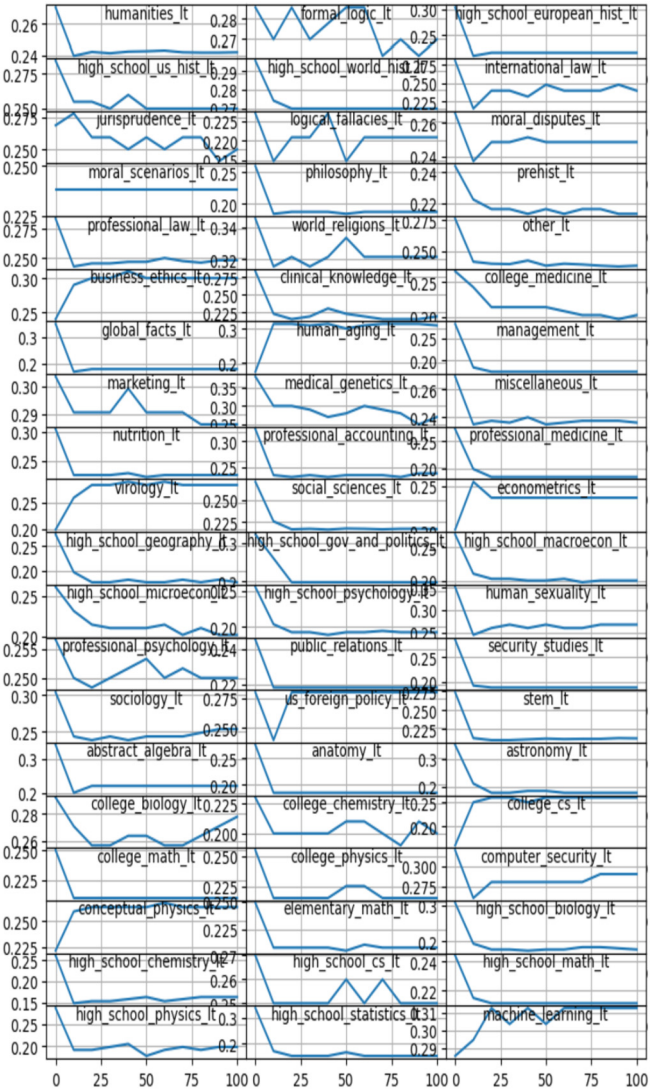

Fig. 8. Accuracies (y-axis) of individual MMLU benchmarks for LT-Llama2-7B model, pretrained with different proportions of Lithuanian component of CulturaX dataset (x-axis).

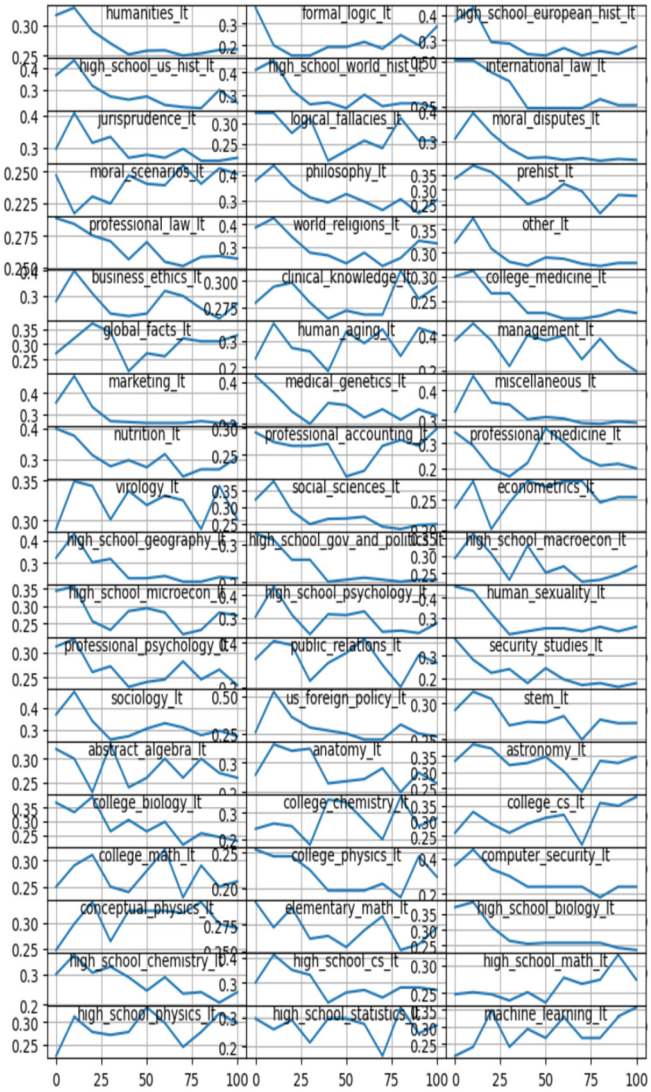

Fig. 9. Accuracies (y-axis) of individual MMLU benchmarks for LT-Llama2-13B model, pretrained with different proportions of Lithuanian component of CulturaX dataset (x-axis).

Table 10
Examples of text generation with the proposed LLMs (pretrained versions). Prompts are indicated in gray text.

LT-Llama2-7B
Kartą gyveno senelis ir senelė. Senelis buvo labai geras žmogus, tai jis visada pasiūlydavo kokį nors darbelį, kad tik jį padarytų. Senelė buvo labai gera, tai ji visada viską išbandydavo, kad tik jos vaikai būtų sveiki. Vieną dieną senelis išėjo į mišką, o senelė pasiėmė savo vaikus ir išėjo. Senelis sako: "O kodėl aš
Jungtinės Valstijos pirmadienį išreiškė susirūpinimą dėl Vengrijos ministro pirmininko Viktoro Orbano susitikimo su Kinijos prezidentu Xi Jinpingu Pekine aptarti karo Ukrainoje. "Mes atidžiai stebime padėtį Ukrainoje ir raginame Vengriją gerbti tarptautinę teisę ir laikytis savo įsipareigojimų", – sakoma JAV valstybės sekretoriaus Antony Blinkeno pareiškime. JAV valstybės sekretorius pabrėžė, kad Jungtinės Valstijos "palaiko Ukrainos teritorinį vientisumą ir suverenitetą". "Mes raginame Vengriją gerbti tarptautinę teisę...
Jau saulelė vėl atkopdama budino svietą. Ir žiemos šaltos triūsus pargriaudama juokės. Šalčių pramonės vėjai, Šaltųjų karų pelenai, Tamsos dulkės, Sniego gniūžtės, Tylos siena, Kurią, rodos, kadaise sukūrė žmonės. O kadaise – šaltis, Sniegas, Šaltis, Šaltis, Šaltis, Šaltis, Šaltis, Šaltis, Šaltis, Šaltis...
Broliai, seserys, imkiat mani ir skaitykiat, Ir tatai skaitydami permanykiat. Mokslo šito tėvai jūsų trokšdavo turėti, Ale to negalėjo nė vienu būdu gauti. Ir tatai skaitydami permanykiat. Mokslo šito tėvai jūsų trokšdavo turėti, Ale to negalėjo nė vienu būdu gauti. Broliai, seserys, imkiat mani ir skaitykiat, Ir tatai skaitydami permanykiat. Mokslo šito tėvai jūsų trokšdavo turėti, Ale to negalėjo nė vienu būdu gauti. Ir tatai skaitydami...
LT-Llama2-13B
Kartą gyveno senelis ir senelė. Jie turėjo vieną vaiką. Vaikas buvo labai smalsus. Jis klausinėjo savo tėvų, kodėl jie turi tiek daug vaikų. Kartą senelis sugalvojo nuvykti į kitą miestą. Jis norėjo nusipirkti daugiau maisto ir kitų reikalingų daiktų. Senelis išsiruošė į kelionę...
Jungtinės Valstijos pirmadienį išreiškė susirūpinimą dėl Vengrijos ministro pirmininko Viktoro Orbano susitikimo su Kinijos prezidentu Xi Jinpingu Pekine aptarti karo Ukrainoje. JAV prezidento Joe Bideno atstovas spaudai Ridas Bouchardas sakė, kad susitikimas įvyks kitą savaitę, bet nepatikslino, ar jame dalyvaus JAV prezidentas. V. Orbanas ir Xi Jinpingas susitiks aptarti karo Ukrainoje, sakė Vengrijos užsienio reikalų ministras Peteris Szijjarto. Kinijos valstybinė naujienų agentūra „Xinhua“ pranešė, kad per susitikimą bus aptarta padėtis Ukrainoje, bet nenurodė, ar V. Orbanas vyks į Pekiną
Jau saulelė vėl atkopdama budino svietą. Ir žiemos šaltos triūsus pargriaudama juokės.Šalčių pramonės darbuotojai, atsidėkodami už tą jų šventę, ėmė švęsti visi kartu. Ir šventė jie septynias dienas. Paskui prasidėjo kalėdinės atostogos. O tada prasidėjo tikros linksmybės. Ir šventės tęsėsi ištisus metus. Kaip gi šventė? Ogi šventė visi kartu. Ir kiekvienas kartu. Ir visi kartu. Ir kiekvienas kartu. Ir visi kartu
Broliai, seserys, imkiat mani ir skaitykiat, Ir tatai skaitydami permanykiat. Mokslo šito tėvai jūsų trokšdavo turėti, Ale to negalėjo nė vienu būdu gauti. Ale dabar tatai jums atsivėrė. Jūs esate gavę tai, kas anksčiau buvo paslėpta nuo jūsų akių; taip pat ir tatai, kas buvo paslėpta nuo tėvų akių, dabar yra jums atskleista. O kadangi šitas mokslas yra jums atskleistas, tai dabar jūs, broliai, seserys, imkitės to, kad skaitytumėtės jį dieną naktį, kad tiktai jūsų širdys būtų atvertos, kad tiktai

Acknowledgements

This research was funded by Neurotechnology. We are grateful to Neurotechnology for providing resources and support for this research. We thank Rasa Kundrotaitė and Greta Tikužytė for editing the English language, Ignas Mataitis, and other colleagues for useful remarks and discussions. We also extend our thanks to anonymous reviewers for their valuable feedback.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., Penedo, G. (2023). *The Falcon Series of Open Language Models*. https://arxiv.org/abs/2311.16867.

Bacciu, A., Trappolini, G., Santilliand, A., Rodolà, E., Silvestri, F. (2023). *Fauno: The Italian Large Language Model that will Leave you Senza Parole*! https://arxiv.org/abs/2306.14457.

Boros, T., Chivereanu, R., Dumitrescu, S.D., Purcaru, O. (2024). Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In: *Proceedings of the Third Ukrainian Natural Language Processing Workshop, LREC-COLING*. European Language Resources Association, Torino, Italy.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems*. https://arxiv.org/abs/2110.14168.

Csaki, Z., Li, B., Li, J., Xu, Q., Pawakapan, P., Zhang, L., Du, Y., Zhao, H., Hu, C., Thakker, U. (2024). *SambaLingo: Teaching Large Language Models New Languages*. https://arxiv.org/abs/2404.05829.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., Hashimoto, T.B. (2024). *AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback*. https://arxiv.org/abs/2305.14387.

Eberhard, D.M., Simons, G.F., Fennig, C.D. (2021). *Ethnologue: Languages of the World*, 24th ed. SIL International. Accessed: 2024-01-04. https://www.ethnologue.com/.

Ekgren, A., Gyllensten, A.C., Stollenwerk, F., Öhman, J., Isbister, T., Gogoulou, E., Carlsson, F., Heiman, A., Casademont, J., Sahlgren, M. (2023). *GPT-SW3: An Autoregressive Language Model for the Nordic Languages*. https://arxiv.org/abs/2305.12987.

Faysse, M., Fernandes, P., Guerreiro, N.M., Loison, A., Alves, D.M., Corro, C., Boizard, N., Alves, J., Rei, R., Martins, P.H., Bigata Casademunt, A., Yvon, F., Martins, A.F.T., Viaud, G., Hudelot, C., Colombo, P. (2024). *CroissantLLM: A Truly Bilingual French-English Language Model*. https://arxiv.org/abs/2402.00786.

Gao, L., Tow, J., Abbasi, B., et al. (2023). A framework for few-shot language model evaluation. Zenodo. https://doi.org/10.5281/zenodo.10256836. https://zenodo.org/records/10256836.

Garcia, G.L., Paiola, P.H., Morelli, L.H., Candido, G., Cândido, A., Jodas, D.S., Afonso, L.C.S., Rizzo Guilherme, I., Penteado, B.E., Papa, J.P. (2024). *Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task*. https://arxiv.org/abs/2401.02909.

Gonen, H., Iyer, S., Blevins, T., Smith, N., Zettlemoyer, L. (2023). Demystifying prompts in language models via perplexity estimation. In: Bouamor, H., Pino, J., Bali, K. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pp. 10136–10148. https://doi.org/10.18653/v1/2023.findings-emnlp.679.

Gordić, A. (2024). YugoGPT Model. https://huggingface.co/gordicaleksa/YugoGPT. Accessed: 2024-07-15.

Grattafiori, A., Dubey, A., Jauhriand, A., et al. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding*. https://arxiv.org/abs/2009.03300.

Hernandez, D., Brown, T.B., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C., Amodei, D., Joseph, N., Kaplan, J., McCandlish, S. (2022). *Scaling Laws and Interpretability of Learning from Repeated Data*. https://arxiv.org/abs/2205.10487.

Hu, E.J., Yelong S., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2022). LoRA: low-rank adaptation of large language models. In: *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9.

INSAIT (2024). BgGPT-7B-Instruct-v0.1 model. https://huggingface.co/INSAIT-Institute/BgGPT-7B-Instruct-v0.1. Accessed: 2024-07-17.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh Chaplot, D., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., El Sayed, W. (2023). *Mistral 7B*. https://arxiv.org/abs/2310.06825.

Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.-A.,

Stock, P., Subramanian, S., Yang, S., Antoniak, S., Le Scao, T., Gervet, T., Lavril, T., Wang, T., Lacroix, T., El Sayed, W. (2024). *Mixtral of Experts*. https://arxiv.org/abs/2401.04088.

Kapočiūtė-Dzikienė, J., Bergmanis, T., Pinnis, M. (2025). *Localizing AI: Evaluating Open-Weight Language Models for Languages of Baltic States*. https://arxiv.org/abs/2501.03952.

Kim, Y., Kim, D., Choi, J., Park, J., Oh, N., Park, D. (2024). *A Survey on Integration of Large Language Models with Intelligent Robots*. https://arxiv.org/abs/2404.09228.

Lai, V.D., Nguyen, C.V., Ngo, N.T., Nguyen, T., Dernoncourt, F., Rossi, R.A., Nguyen, T.H. (2023). *Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback*. https://arxiv.org/abs/2307.16039.

LumiOpen (2024). Viking 13B Model. https://huggingface.co/LumiOpen/Viking-13B. Accessed: 2024-07-15.

Mabeck (2024). Heidrun-Mistral-7B-chat model. https://huggingface.co/Mabeck/Heidrun-Mistral-7B-chat. Accessed: 2024-07-20.

Martins, P.H., Fernandes, P., Alves, J., Guerreiro, N.M., Rei, R., Alves, D.M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J.G.C., Birch, A., Martins, A.F.T. (2024). *EuroLLM: Multilingual Language Models for Europe*. https://arxiv.org/abs/2409.16235.

Masala, M., Ilie-Ablachim, D.C., Corlatescu, D., Zavelca, M., Leordeanu, M., Velicu, H., Popescu, M., Dascalu, M., Rebedea, T. (2024). *OpenLLM-Ro – Technical Report on Open-source Romanian LLMs*. https://arxiv.org/abs/2405.07703.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J. (2024). *Large Language Models: A Survey*. https://arxiv.org/abs/2402.06196.

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A. (2024). *A Comprehensive Overview of Large Language Models*. https://arxiv.org/abs/2307.06435.

Neurotechnology (2024). Lt-QA-V1 dataset. https://huggingface.co/datasets/neurotechnology/lithuanian-qa-v1.

Nguyen, T., Nguyen, C.V., Lai, V.D., Man, H., Trung Ngo, N., Dernoncourt, F., Rossi, R.A., Nguyen, T.H. (2023). *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*. https://arxiv.org/abs/2309.09400.

Norallm (2024). Normistral-7B-Warm. https://huggingface.co/norallm/normistral-7b-warm. Accessed: 2024-07-20.

OpenAI, Achiam, J., Adler, S., et al. (2024). *GPT-4 Technical Report*. https://arxiv.org/abs/2303.08774.

Plüster, B., Schuhmann, C. (2024). *LAION LeoLM: Linguistically Enhanced Open Language Model*. https://huggingface.co/LeoLM. Accessed: 2024-07-17.

Projecte AINA (2024). Aguila 7B Model. https://huggingface.co/projecte-aina/aguila-7b. Accessed: 2024-07-15.

Rijgersberg, L. (2024). GEITje. https://github.com/Rijgersberg/GEITje. Accessed: 2024-07-17.

Riviere, M., Pathak, S., Sessa, P.G. et al. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. https://arxiv.org/abs/2408.00118.

Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y. (2019). *WinoGrande: An Adversarial Winograd Schema Challenge at Scale*. https://arxiv.org/abs/1907.10641.

Snæbjarnarson, V., Símonarson, H.B., Ragnarsson, P.O., Ingólfsdóttir, S.L., Jónsson, H.P., Þorsteinsson, V., Einarsson, H. (2022). *A Warm Start and a Clean Crawled Corpus – A Recipe for Good Language Models*. https://arxiv.org/abs/2201.05601.

SPAHE (2024). Meltemi-7B-Instruct-v1-GGUF. https://huggingface.co/SPAHE/Meltemi-7B-Instruct-v1-GGUF. Accessed: 2024-07-17.

Speakleash (2024). Bielik-7B-Instruct-v0.1. https://huggingface.co/speakleash/Bielik-7B-Instruct-v0.1. Accessed: 2024-07-20.

Touvron, H., Martin, L., Stone, K., et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://arxiv.org/abs/2307.09288.

Ulčar, M., Robnik-Šikonja, M. (2021). SloBERTa: Slovene Monolingual Large Pretrained Masked Language Model. In: *24th International Multiconference Information Society 2021, Volume C. Data Mining and Data Warehouses*. Ljubljana.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017). Attention Is All You Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zhang, W., Deng, Y., Liu, B., Pan, S., Bing, L. (2024). Sentiment analysis in the Era of Large Language Models: a reality check. In: Duh, K., Gomez, H., Bethard, S. (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, pp. 3881–3906. https://aclanthology.org/2024.findings-naacl.246.

A. Nakvosas was born in Lithuania in 1986. He received a bachelor's degree in 2009 and a master's degree in 2012 from Šiauliai University. Since 2013, he has been working in R&D at Neurotechnology. His work spans machine learning and biometric systems, including fingerprint, iris, and facial recognition, and has been recognized in NIST evaluations. In recent years, his research has expanded to natural language processing, with a focus on speech-to-text, text-to-speech, and large language models. His interests include deep learning, transformer architectures, signal processing, and software engineering.

P. Daniušis was born in Lithuania in 1983. He received a bachelor's degree (mathematics) from Šiauliai University in 2005, a master's degree (mathematics) from Vilnius University in 2007, and a PhD (computer science) from Vilnius University in 2012. He has been working at Neurotechnology since 2010. His research interests include AI, artificial neural networks, adaptive robotics, causal inference, and statistical dependence estimation.

V. Mulevičius (born in 1997) earned his bachelor's degree in computer science from the University of Birmingham in 2020. During his studies, he developed a strong interest in artificial intelligence and natural language processing (NLP), which led him to join the NLP team at Neurotechnology in 2018. Since then, he has been actively involved in the development of language technologies, contributing to various projects ranging from speech recognition and text analysis to the creation of large-scale language models.