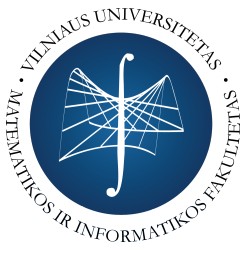


Large Language Models For Lithuanian Language

Student: Dominykas Plesevičius
Supervisor: Prof. Aistis Raudys

16 January 2025

Goal

The goal of the project is to fine-tune a small/medium size LLM for Lithuanian language and release it openly.

Outline

1. Data processing & analysis
2. Model selection
3. Model training
   - 2 existing multilingual models trained with additional Lithuanian data
   - 1 model trained from scratch, with Lithuanian tokenizer
4. Evaluation
5. Conclusions & improvements

Data

The Lithuanian partition of Multilingual Colossal Clean Crawled Corpus (mC4) dataset will be used for model training.

- Contains filtered web scraped data from the Common Crawl archive.
- 11.2 million entries of Lithuanian text.
- 160,000 websites.

Data

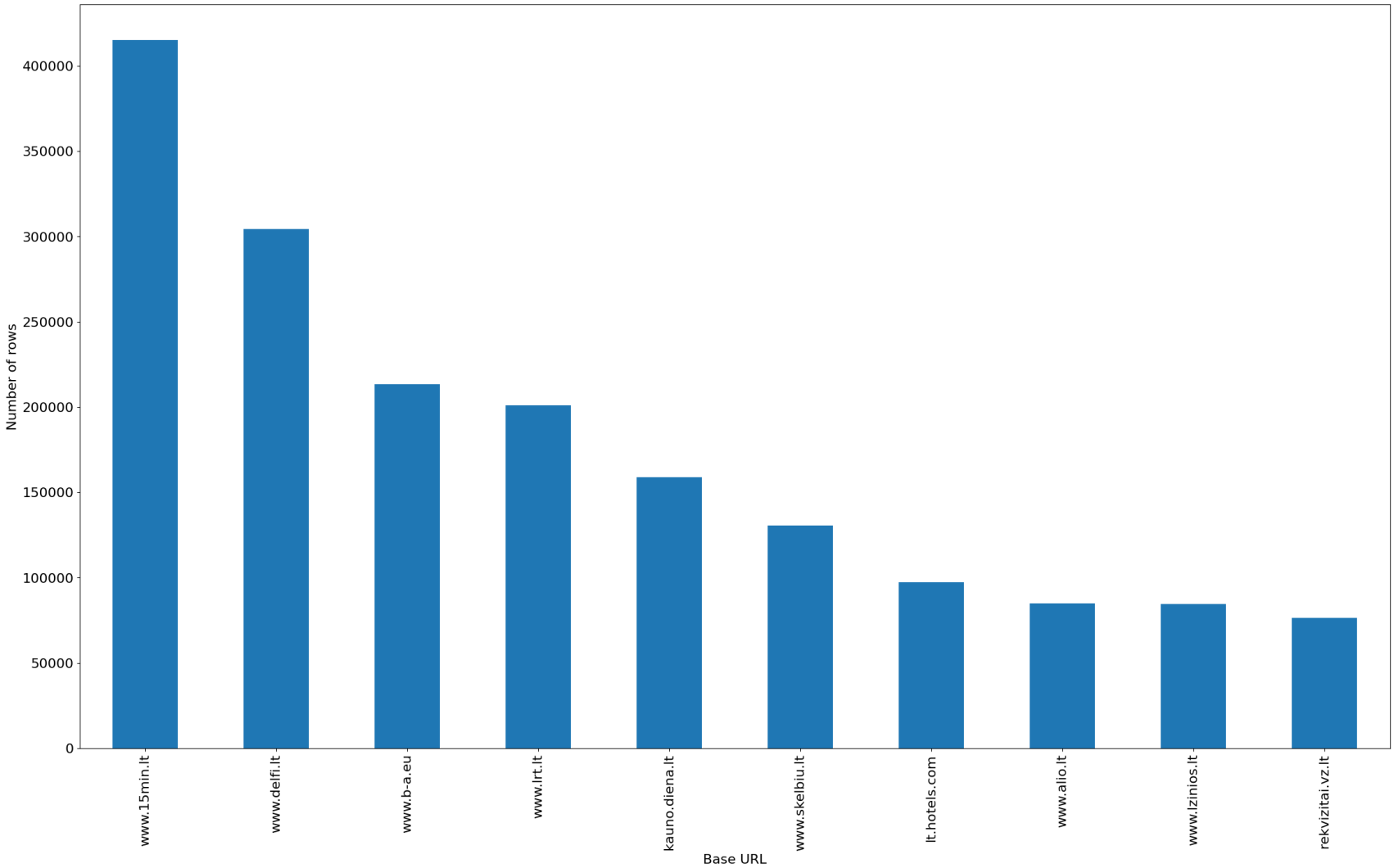


Figure: Top 10 most common URL sources

Data processing

Processing steps:

1. Filter out low quality entries based on perplexity scores from a 5-gram model.
2. Detect language and remove non-Lithuanian text.
3. Remove entries containing Lithuanian specific bad words.
4. Deduplicate dataset.
5. Remove short entries.

Data processing

Table: Summary of how many entries and characters were filtered

	Original	Perplexity filtering	Language filtering	Bad words filtering	Deduplication	Short entries filtering	Filtering rate (%)
Entries (count)	11274295	8455721	8449676	8440990	8437780	8391515	25.6%
Entries (%)	100%	75.0%	74.9%	74.9%	74.8%	74.4%	
Characters (count)	36017009225	30309623778	30163573418	30072147904	27362635039	27359506060	24.0%
Characters (%)	100%	84.2%	83.7%	83.5%	76.0%	76.0%	

Existing Lithuanian LLM's

Currently two open Lithuanian large language models exist:

- neurotechnology/Lt-Llama-2 (7B)
- neurotechnology/Lt-Llama-2 (13B)

From the "Open Llama2 Model for the Lithuanian Language"[1] article.

We will compare our results to these models.

[1] Nakvosas, A., Daniušis, P., Mulevičius, V. (2024). Open Llama2 Model for the Lithuanian Language. https://arxiv.org/abs/2408.12963

Multilingual models

Many multilingual models exist:

- Llama 3.2
- Bloomz
- XLM-RoBERTa
- ...

We select the models for additional training based on the following:

1. Model must be a causal language model.
2. Model size less than 2 billion parameters.
3. How well the model tokenizes Lithuanian text.
4. Has the model seen Lithuanian data.

Tokenization experiment

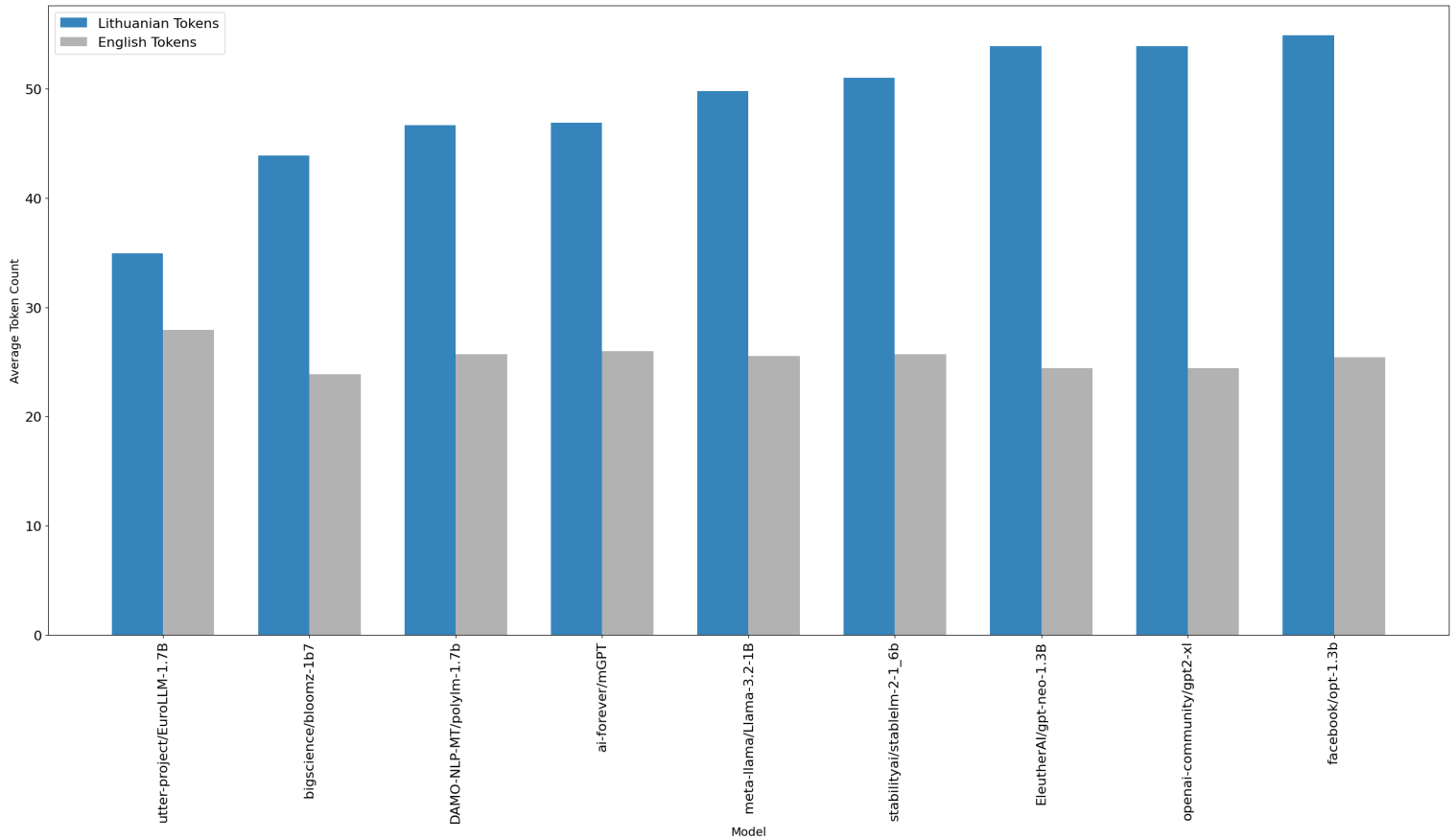


Figure: Average token counts of 1000 Lithuanian and English sentence pairs for each candidate model

Chosen models

Based on the criteria we choose two multilingual models for additional training:

- ai-forever/mGPT
- utter-project/EuroLLM-1.7B

Lithuanian specific model

We also introduce a new Lithuanian specific model:

- Based on the architecture of mGPT (which itself is based on GPT-2)
- Tokenizer parameters (vocabulary size, special tokens) also mirror mGPT
- Tokenizer is trained our Lithuanian data
- Model is trained only on Lithuanian data

We intentionally mirror the mGPT model to ensure stability and reduce the risk of unexpected issues.

Tokenization experiment

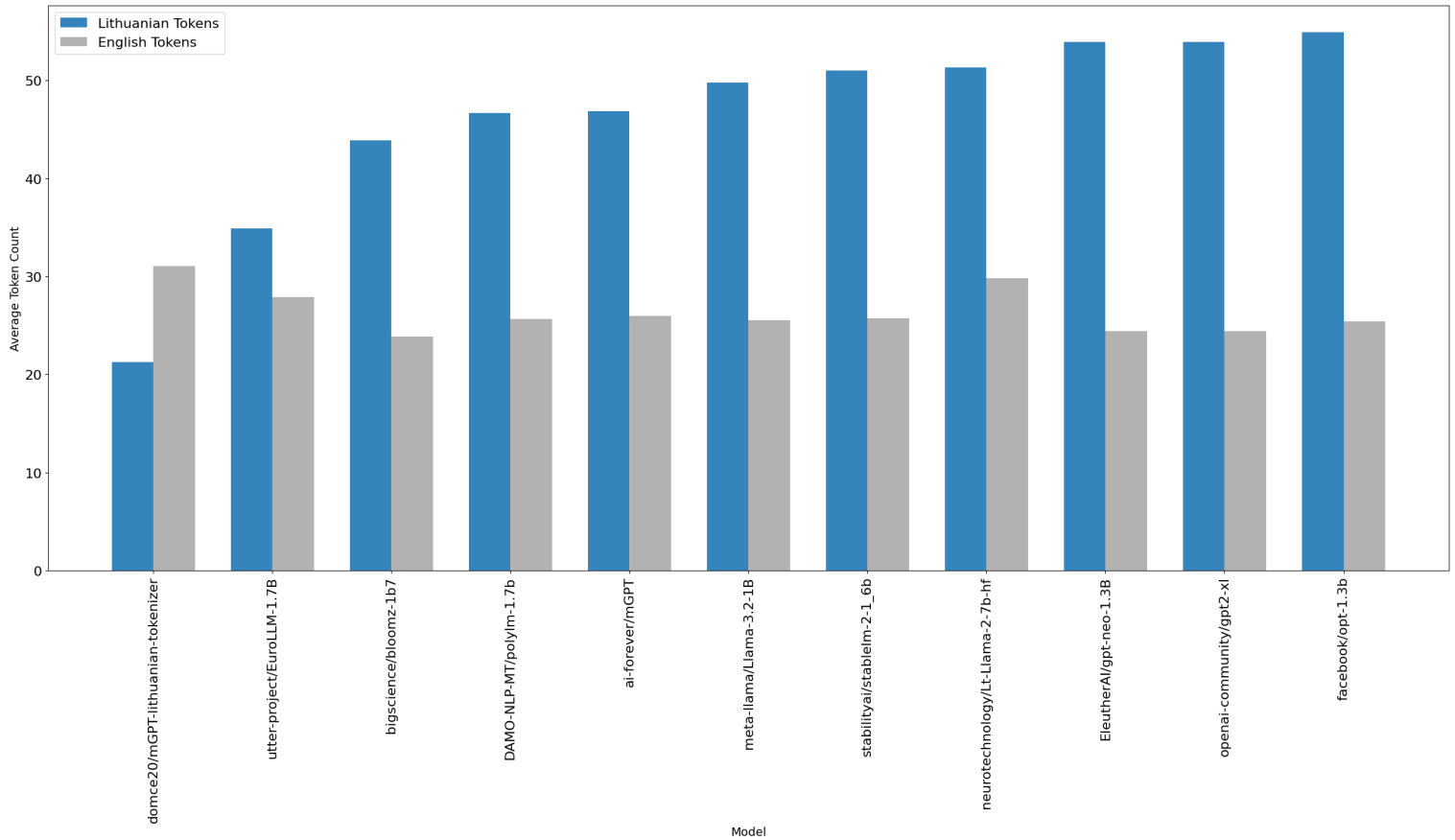


Figure: Average token counts of 1000 Lithuanian and English sentence pairs for candidate, GPT-2 Lithuanian and Lt-Llama models

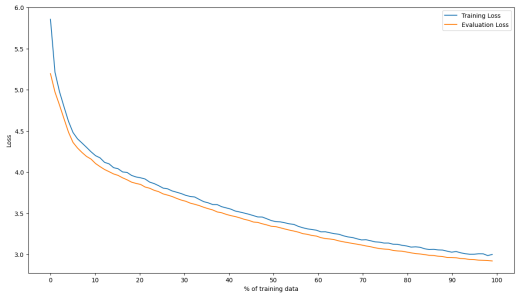
Training strategy

Training was done in 3 steps on the VU supercomputer:

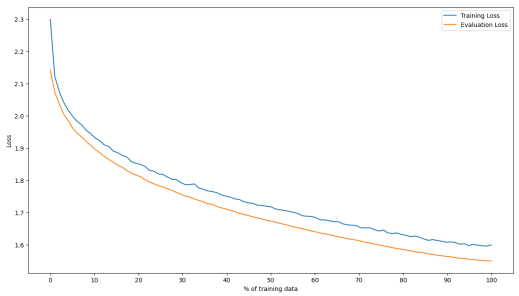
1. Context length 512; 94% data
2. Context length 1024; 5% data
3. Context length 2048; 1% data

This was done in order to train faster and use less computational resources.

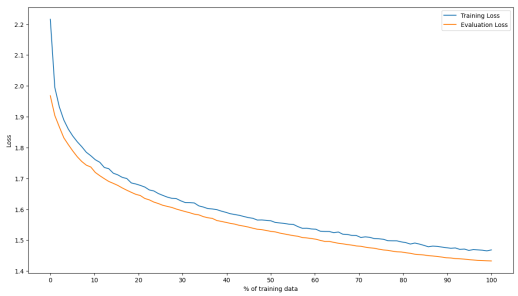
Training - step 1



(a) GPT2 training loss



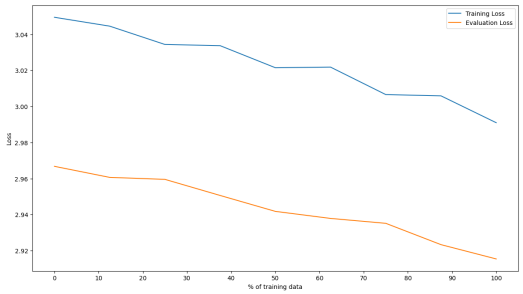
(b) EuroLLM training loss



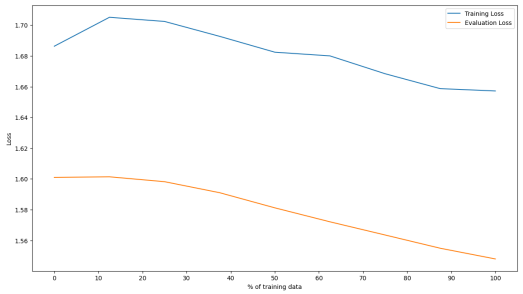
(c) mGPT training loss

Figure: 1st step training losses of the models

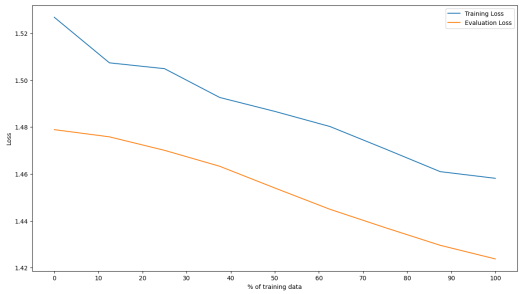
Training - step 2



(a) GPT2 training loss



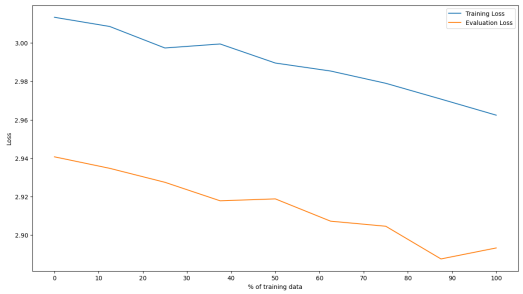
(b) EuroLLM training loss



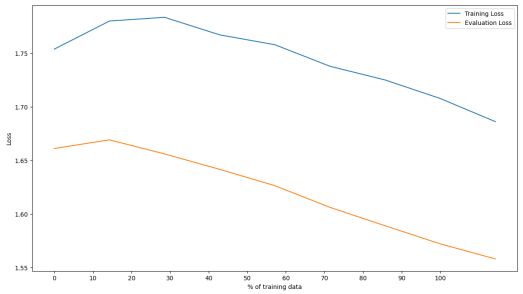
(c) mGPT training loss

Figure: 2nd step training losses of the models

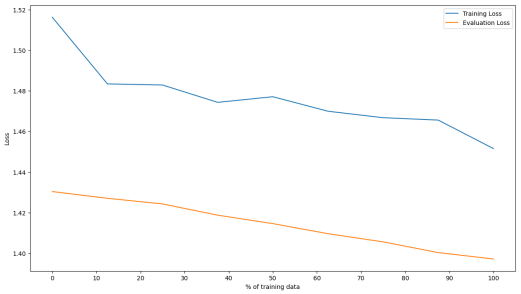
Training - step 3



(a) GPT2 training loss



(b) EuroLLM training loss



(c) mGPT training loss

Figure: 3rd step training losses of the models

Training results

Table: Final losses and training times for each model

Model	Training loss	Validation loss	Training time (hours)
ai-forever/mGPT	1.451	1.397	58.674
utter-project/EuroLLM-1.7B	1.658	1.557	56.486
GPT-2 Lithuanian	2.970	2.888	54.179

Evaluation criteria

Evaluation metrics chosen to evaluate the models:

- Truthful QA - factual correctness
- ARC - scientific reasoning
- WinoGrande - commonsense reasoning
- MMLU - knowledge in different domains
- HellaSwag - contextual understanding
- GSM8K - math problems

We use translated Lithuanian versions of these metrics that are publicly available on huggingface.

Evaluation results

Table: Evaluation of the original and trained models

Benchmark	Original			Trained		
	mGPT	EuroLLM	GPT2-LT	mGPT	EuroLLM	GPT2-LT
arc_challenge	0.1962	0.2619	-	0.2125	0.1928	**0.2190**
arc_easy	0.3392	0.4937	-	0.3994	0.2719	**0.3996**
gsm8k	0.0053	0.0091	-	0.0076	0.0136	**0.0142**
hellaswag	0.2883	0.2755	-	0.3062	0.2785	**0.3070**
mmlu	0.2344	0.2482	-	0.2314	**0.2567**	0.2306
truthful_mc1	0.2509	0.2546	-	0.2558	**0.2913**	0.2705
truthful_mc2	0.4222	0.4094	-	0.4417	**0.4666**	0.4556
winogrande	0.4972	0.5022	-	**0.5257**	0.5028	0.5146

Evaluation results

Comments on evaluation results:

- The results across all models are very similar (except EuroLLM on ARC)
- mGPT model sees a slight increase in most of its benchmark scores.
- EuroLLM has a significant drop on the ARC metrics.
- EuroLLM shows stronger performance on MMLU and TruthfulQA.
- GPT-2 Lithuanian model outperforms the others on ARC and HellaSwag by a narrow margin.
- GPT-2 also shows pretty good results in TruthfulQA.

Evaluation results

Table: Evaluation of the final and Lt-Llama models

Metric	mGPT	EuroLLM	GPT2-LT	Lt-Llama-7B	Lt-Llama-13B
arc_challenge	0.2125	0.1928	0.2190	0.2176	**0.2491**
arc_easy	0.3994	0.2719	0.3996	0.4158	**0.5072**
gsm8k	0.0076	0.0136	0.0142	**0.0197**	0.0144
hellaswag	0.3062	0.2785	0.3070	0.3316	**0.4054**
mmlu	0.2314	**0.2567**	0.2306	0.2314	0.2303
truthful_mc1	0.2558	**0.2913**	0.2705	0.2717	0.2681
truthful_mc2	0.4417	**0.4666**	0.4556	0.4378	0.4222
winogrande	0.5257	0.5028	0.5146	0.5359	**0.6164**

Evaluation results

Comments on evaluation results:

- Lt-Llama models outperform our models by 10% on ARC, HellaSwag, and Winogrande.
- Lt-Llama models are about 10x bigger.
- GSM8K is still a challenge for all models.
- EuroLLM outperforms all models on MMLU and TruthfulQA.

Evaluation example 1

Table: ARC-Easy example 1 with model answers

Question:		
Augalai ir gyvūnai susideda iš organinių junginių. Kokie iš šių elementų dažniausiai randami organiniuose junginiuose? A: Geležis, deguonis, nikelis, varis. B: Natris, kaliumas, auksas, vandenilis. C: Helis, neonas, argonas, kriptonas. D: Anglis, vandenilis, deguonis, azotas.		
Correct answer	D	
Model answer	Original	Trained
mGPT	A	D
EuroLLM	D	D
GPT2-LT	-	D

Evaluation example 2

Table: ARC-Easy example 2 with model answers

Question:		
Kuri savybė apibūdina kačiuko kailio tekstūrą? A: Pilkas. B: Šiltas. C: Ilgas. D: Minkštas.		
Correct answer	D	
Model answer	Original	Trained
mGPT	D	A
EuroLLM	D	A
GPT2-LT	-	A

Improvements

There are many possible improvements:

- Gather or generate more high quality training data.
- Adopt a training strategy similar to EuroLLM, where the final stages are dedicated solely to high quality resources.
- Improving the translation quality of evaluation datasets.
- Use larger base models.
- Smaller, dedicated Lithuanian models.
- Fine-tuning for question answering or specific tasks.

Conclusions

- **Impact of additional Training.** The mGPT and EuroLLM models both benefited from further fine-tuning, however, EuroLLM had a significant drop on ARC after training, suggesting possible risks and importance of training strategies.
- **Results of a dedicated Lithuanian model.** The dedicated GPT-2 Lithuanian model, performed competitively across most tasks, marginally surpassing other models on some benchmarks. Though the gains were small, this highlights the potential for specialized, language-specific designs.
- **Role of model size.** Larger models tend to outperform smaller ones on certain benchmarks like ARC, HellaSwag, and Winogrande. However, these gains are not as pronounced as one might expect for a model that is nearly ten times larger.
- **Importance of high quality Lithuanian data.** The similar evaluation results across all models, point toward the need for bigger, higher quality Lithuanian datasets. Improvements in data quality appear more pivotal than simply increasing model size or changing the architecture.

Thank you for your attention