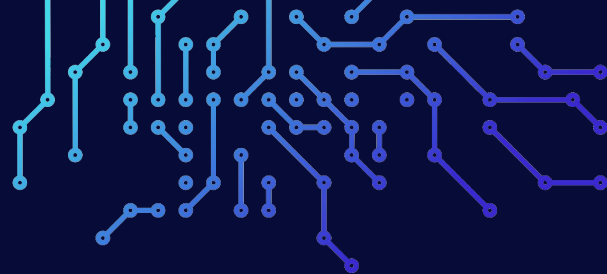




FACULTY
OF ELECTRICAL
ENGINEERING
CTU IN PRAGUE



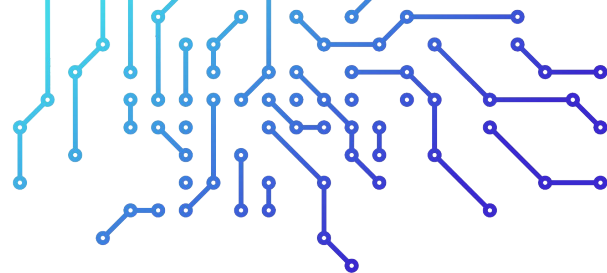
Compute-constrained LLM adaptation to Czech language

Bc. Tomáš Mlynář

Master's Thesis Presentation

Supervisor: Ing. Herbert Ullrich

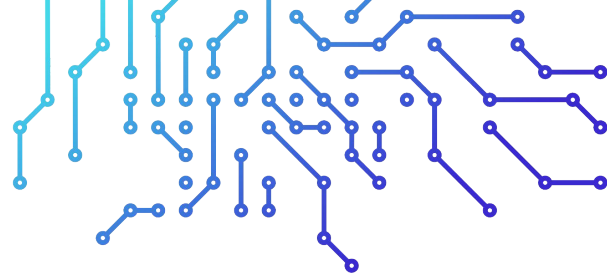
June 19, 2025



Assignment

LLM adaptation for tasks in Czech.

- Analyze SoTA approaches for training and adaptation of LLMs.
- Explore the chosen methods, w.r.t. compute-constraints.
- Assemble and publish appropriate datasets.
- Benchmark the models.
- Release trained models and procedure.



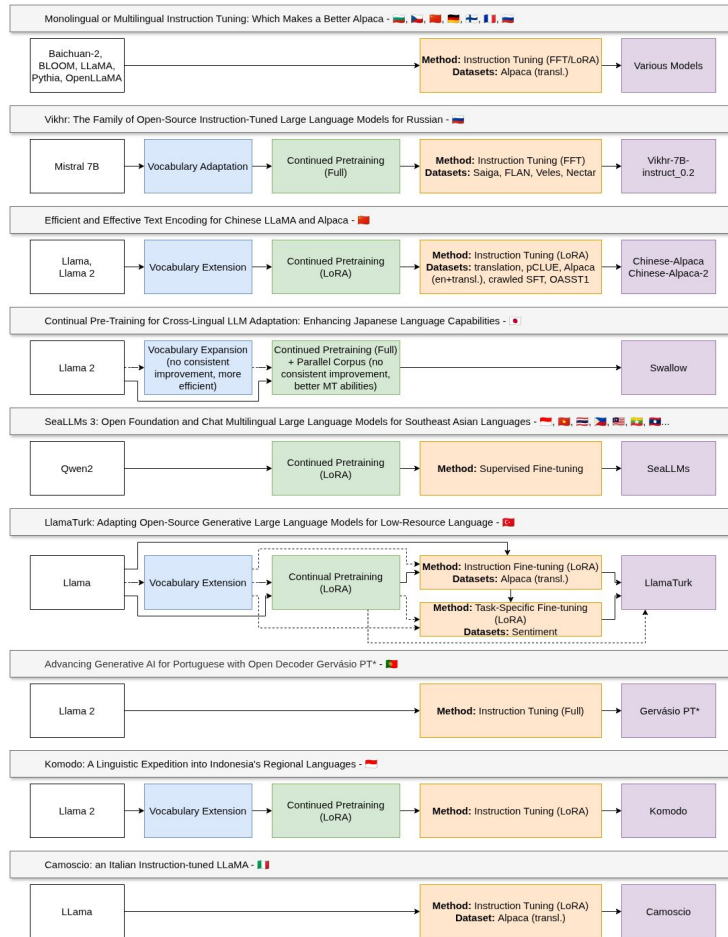
Context & Motivation

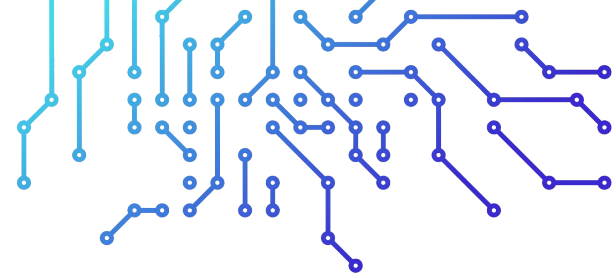
- research in NLP is driven by non-proprietary models
- open-source models (Llama 3.1 8B) **underperforming** in e.g. Czech
- Projects aiming to bridge the gap by full training are: **still in development** (OpenEuroLLM, EuroLingua-GPT) or **underperforming** multilingual models (e.g. Teuken, EuroLLM)
- Related works show that adaptation of multilingual LLMs works in other languages → still no published Czech experiment (proprietary Seznam SeLLMa)



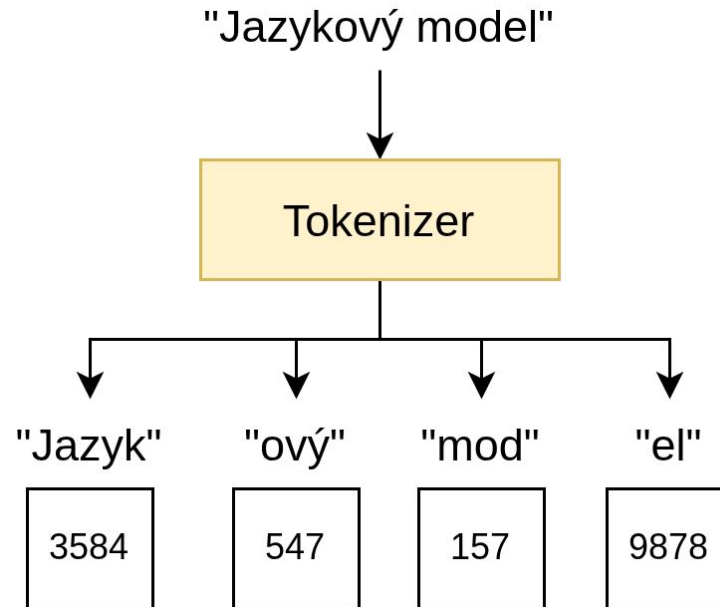
Adaptation Methods

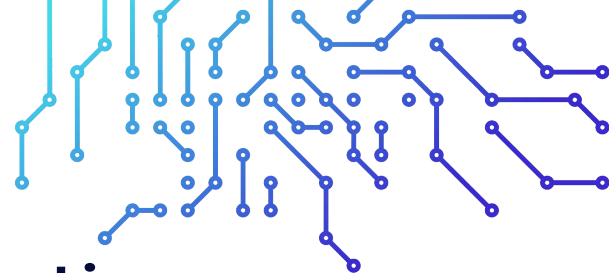
- primarily: **vocabulary adaptation**, **continued pretraining**, **instruction tuning**, human preference alignment
- many approaches and combinations in related research



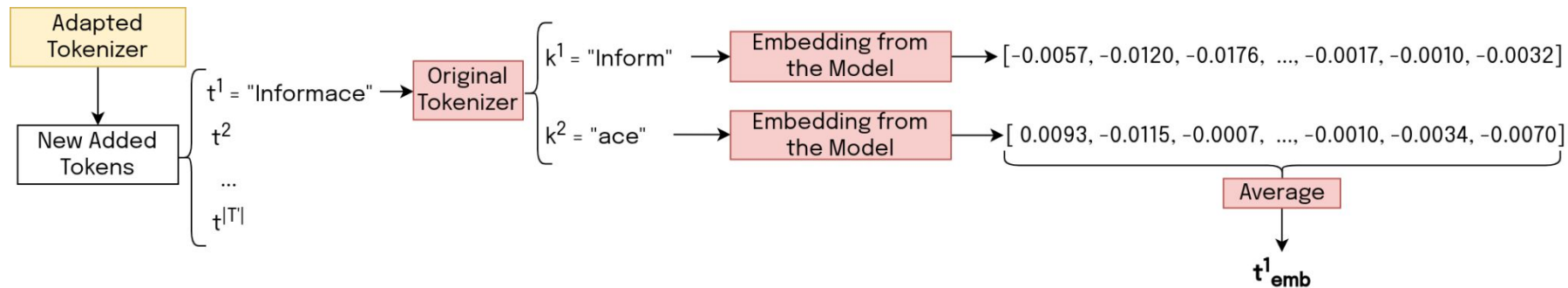


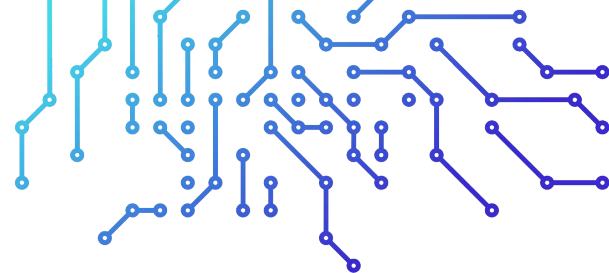
Tokenizer



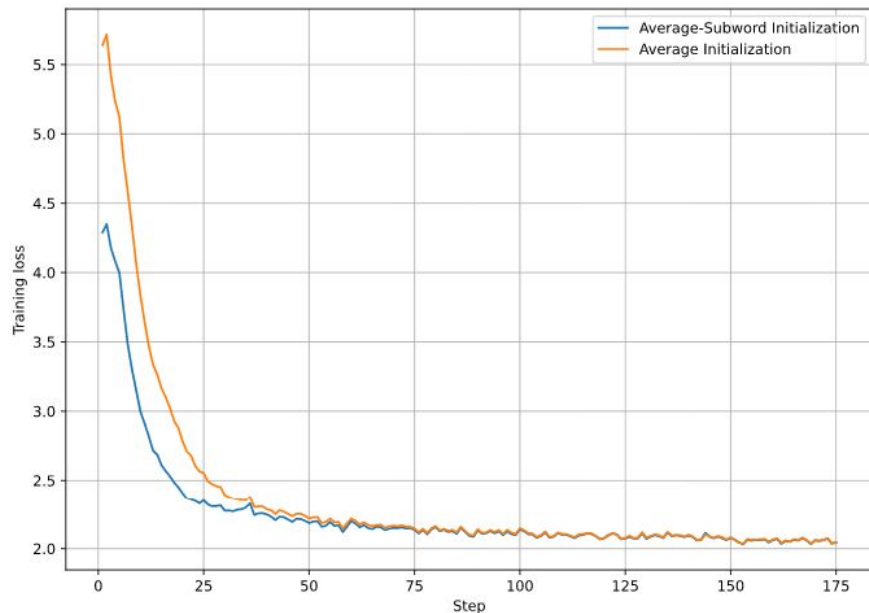


Vocabulary Adaptation – Initialization

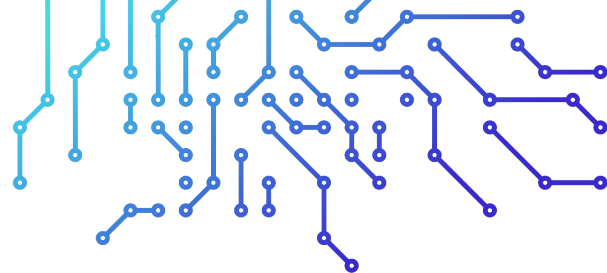




Vocabulary Adaptation

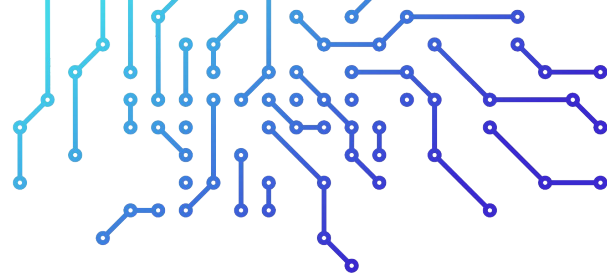


Comparison of training loss curves
between average-subword initialization
and average initialization



Instruction Tuning

- Also called supervised fine-tuning
- “teaching model to follow instructions”
- English datasets: Alpaca, Dolly
- Multilingual datasets: MURI-IT, Bactrian-X, OASST2
- Czech datasets: [Ask Your Library](#), [Questions UJC CAS](#)
- [Collections](#) standardizing available data were published
- FLAN-like [NLI task-specifying tuning collection](#) was published



Example – Ask Your Library

Prosím vysvětli slovo tropář

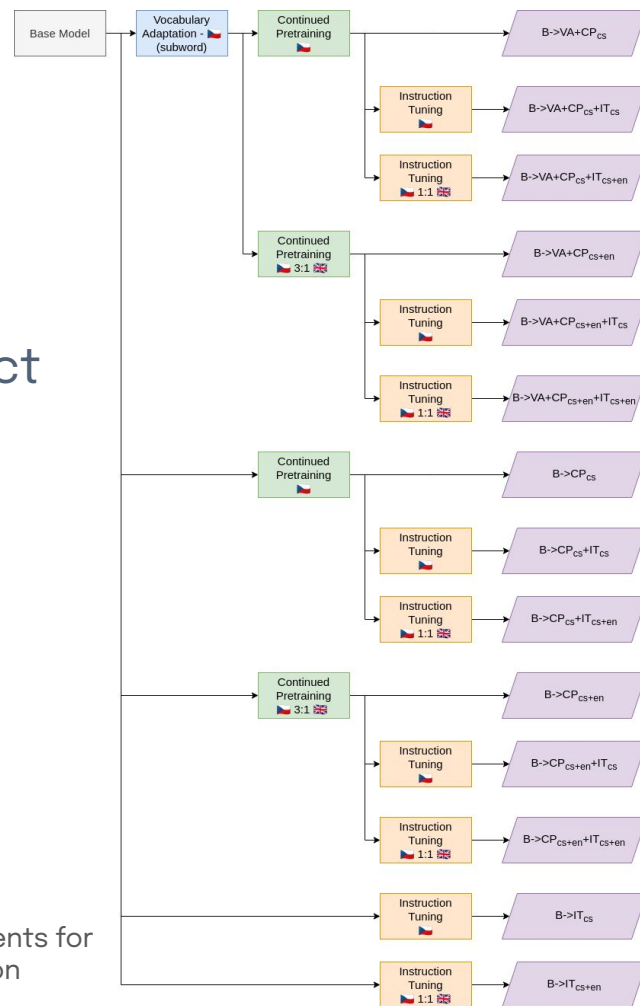
Metadata: title, okres,
library, category, date,
url, id, language,
diacritics_korektor,
spelling_korektor

Tropář je soubor skladeb na liturgické texty, s notovým zápisem. Tropar je církevní hymnus, častý ve východní liturgii – u pravoslavné církve. Existuje např. Velikonoční tropář (soubor velikonočních liturgických skladeb), u nás patří k nejstarším Svatovítský tropář z konce 12. století. Vůbec k nejstarším patří Winchesterský tropář asi z roku 980, který obsahuje přes 150 dvouhlasých skladeb na liturgické texty, s nedokonalou notací.



Training Plan

- + 2 experiments on Llama 3.1 8B Instruct
- + 2 experiments on Dolly & Alpaca

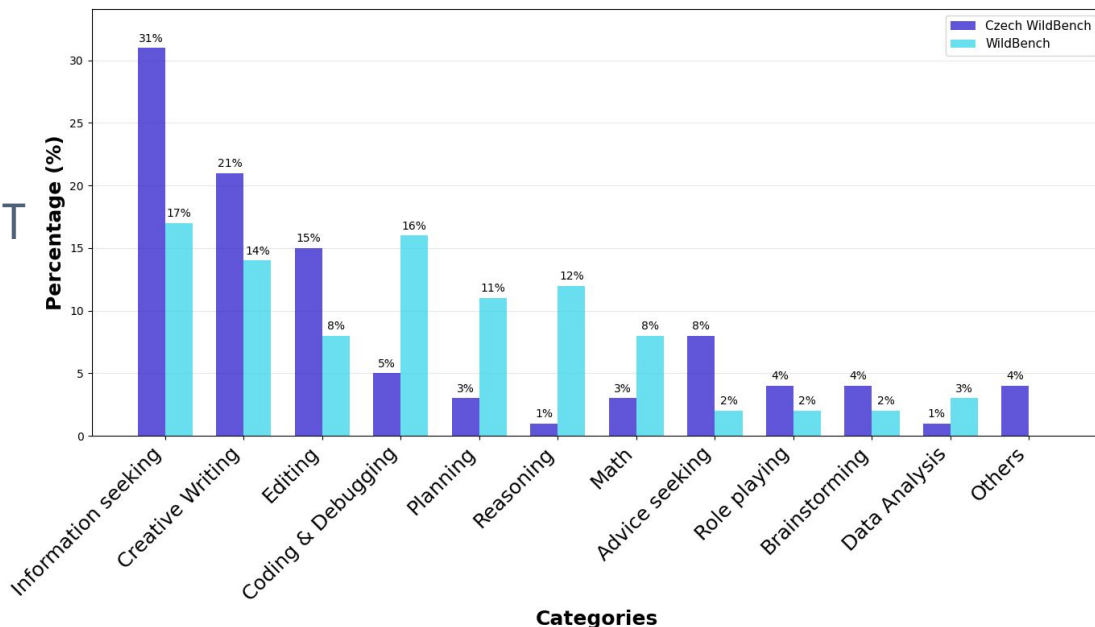


Plan of experiments for
model adaptation



Czech WildBench

- Adaptation of WildBench Benchmark to Czech
- LMSYS, WildChat, ShareGPT datasets
- LLM-as-a-judge
- WB-Score, WB-Reward



Overview of change in category distribution between Czech and original WildBench

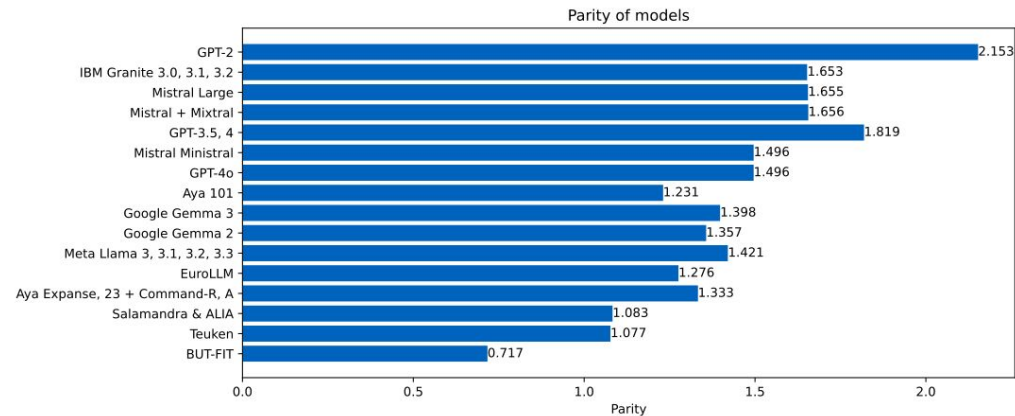
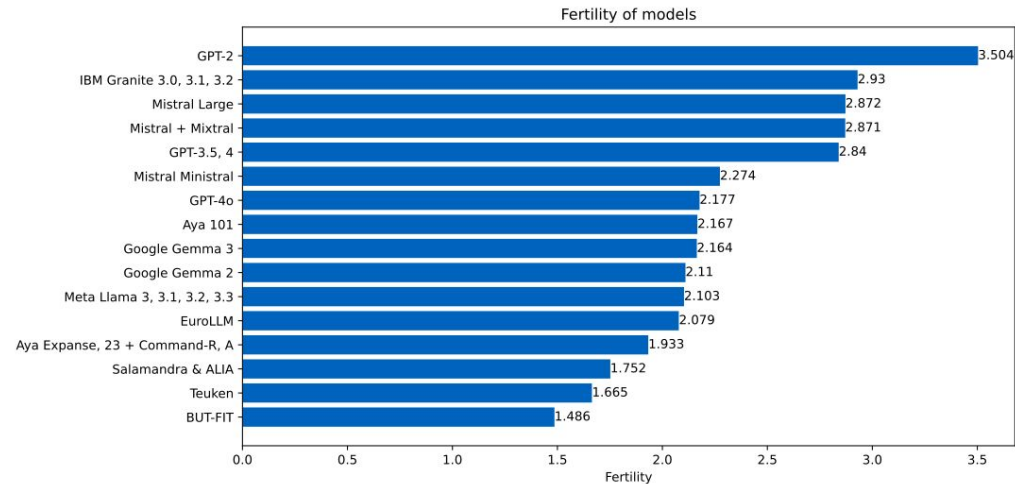


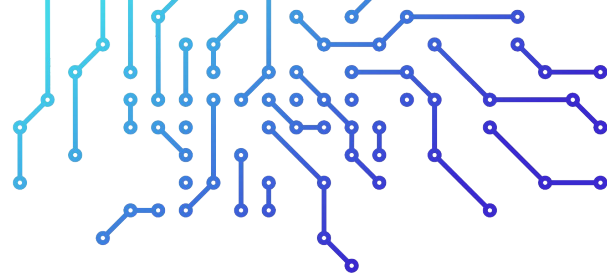
Tokenizer Evaluation

- prepared the evaluator utilizing two metrics:

$$\text{fertility} = \frac{\# \text{ of tokens}}{\# \text{ of words}}$$

$$\text{parity} = \frac{\# \text{ of cs tokens}}{\# \text{ of en tokens}}$$



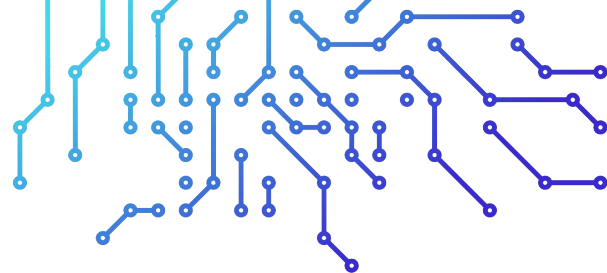


Evaluation & Training Results

- Training with QLoRA (LoftQ) on one NVIDIA A100 80GB GPU
- Tokenizer evaluation, perplexity, preselection prompts, BenCzechMark, Czech WildBench
- **Catastrophic forgetting avoided** – mixtures of Czech and English data

Model	Czech Perplexity ↓	English Perplexity ↓
Llama 3.1 8B – Baseline	12.05	13.26
Baseline → 🇨🇪 Continued pretraining	8.57	22.71
Baseline → 🇨🇪 + 🇬🇧 Continued pretraining	8.87	16.50

Perplexity evaluation results computed using Czech and English hold-out sets

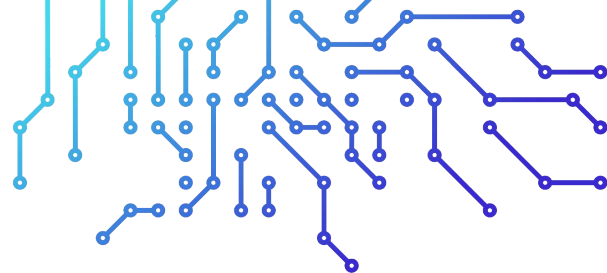


Evaluation & Training Results

- Tokenizer evaluation lead **higher fertility**
- Llama 3.1 8B Instruct **baseline ranked as the best model**
- No specific best-performing model from the trained ones – **promising results in some categories** of BenCzechMark:
 - on half of datasets better than Llama 3.1 8B baseline (e.g. all NLI, Sentiment, Math)
 - on specific tasks better than or comparable to Llama 3.1 8B Instruct baseline
 - Subjectivity, Czech Sentiment CSFD, Umimeto.cz Math, Cermat Math and Czech

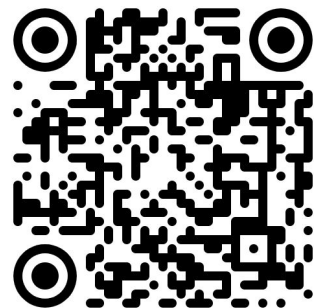


Conclusion



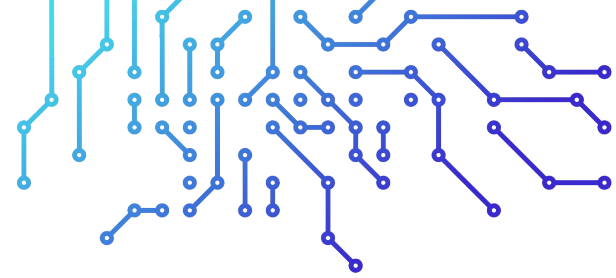
The main thesis accomplishments are:

- overview of adaptation techniques and PEFT methods used in related works, and of Czech-related LLMs,
- examination of available datasets,
- creation and publishing of 2 original instruction-tuning datasets,
- assembling instruction-tuning data collections (🇨🇪, 🇬🇧)
- assembling NLI instruction-tuning collection
- creation of Czech version of WildBench benchmark
- training, and evaluation of various adaptation approaches

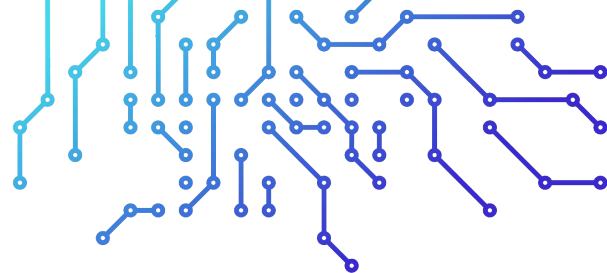




**FACULTY
OF ELECTRICAL
ENGINEERING
CTU IN PRAGUE**

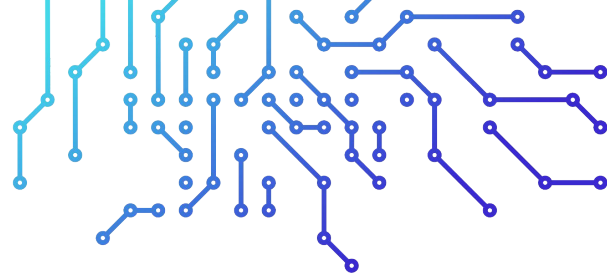


Thank you for you attention!



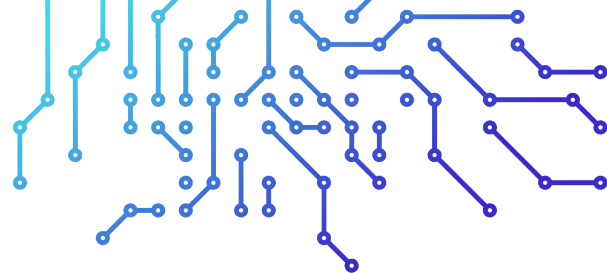
It seems that you have found and identified well the issue behind the fertility increase (page 43). : prioritization of tokens added by the add_tokens method. What other possibilities would there be to add new tokens without causing this issue?

1. Do not add new tokens
2. Retrain the BPE tokenizer from scratch and copy related embeddings
3. Add new tokens directly and add rules which formed them - problem:
 - a. original: p o s l e d n í → **po** s l e d n í → **pos** l e d n í → pos l **ed** n í → **pos** **led** n í
 - b. rules added to end (e.g. s+l→sl, e+d→ed, sl+ed→sled...): sled rule not applied
 - c. rules added to front -> breaking the original tokenizer: s l e d d i n g s →* **sled** d d i n g s (however in original there could be rule sle + dding → sledding leading to **sledding** s)
 - d. merged with frequencies -> frequencies not comparable and **not available**



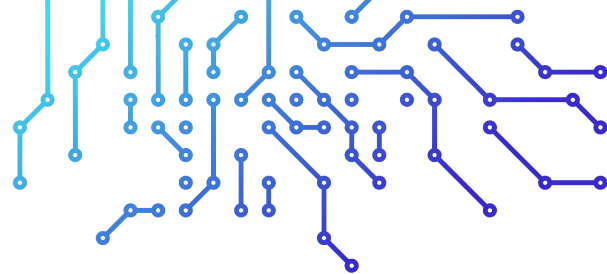
You noted repeating tokens and unwanted artifacts in the outputs. Could these be related to tokenizer adaptation?

- Probably not. The artefacts also occurred for models without tokenizer adaptation.
- They seem to be related to dataset quality and training hyperparameters.



Please explain your choice of the 3:1 ratio of Czech to English data for continued pretraining and the 1:1 ratio for instruction tuning.

- Way to avoid catastrophic forgetting
- These ratios occur most times in related research
- 3:1 – based performed experiments in related research
- 1:1 – often based on data unavailability or motivated by cross-lingual transfer



Continued Pretraining

- Next token prediction task (Causal Language Modelling)
- decoder-only LLMs
- Datasets available (text): BUT-LCC, CTU collection, FineWeb

