UNIVERSITY OF TARTU
1632

Estonian language support in open-source large language models

Kairit Sirts
Aleksei Dorkin

LUMI hackathon
Oslo, May 2025

Project

Project details

Title: Estonian language support in open-source large language models

Project duration: August 2024– July 2025 (with an expected extension until July 2026).

Project lead: University of Tartu (PI + 5 people)

Other participating organizations

- Tallinn University of Technology (2 people)
- Estonian Language Institute (2 people)
- Tallinn University (1 person)

Project goal

Continue training of pretrained open-source LLMs (like LLaMa) on Estonian data to improve:

- Estonian language capabilities and fluency
- Knowledge of Estonian language and culture

Project activities

- Data collection for training
- Benchmark development for model evaluation
- Model training

Project activities: data collection

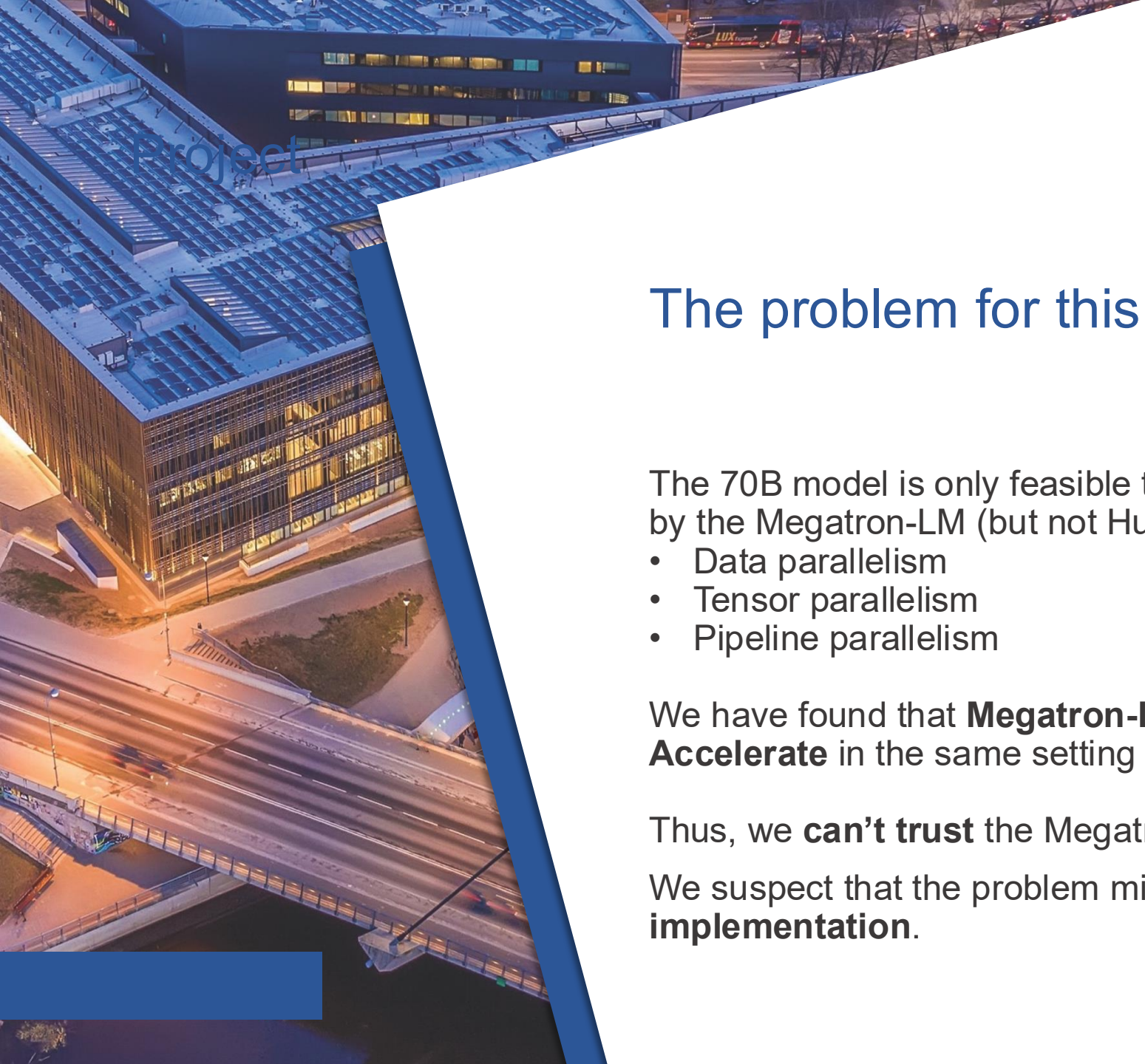
- Curate and preprocess publicly available Estonian corpora
- Obtain copyright protected texts in collaboration with Estonian Language Institute
- Collect data for subsequent training steps:
  - Instruction finetuning and human preferences

Project activities: benchmark creation

- Gather available Estonian benchmarks into LM Evaluation Harness
- Create new benchmarks
- Translate benchmarks (both automatically and manually)

Project activities: model training

- Figure out necessary configurations for model training (data mixes, hyperparameters)
- Continue training LLaMa3.x models of different sizes:
  - 1B and 8B parameter models using Huggingface trainer
  - 70B parameter model with Megatron-LM
- Monitor training effectiveness with benchmark evaluations



Project

The problem for this hackathon

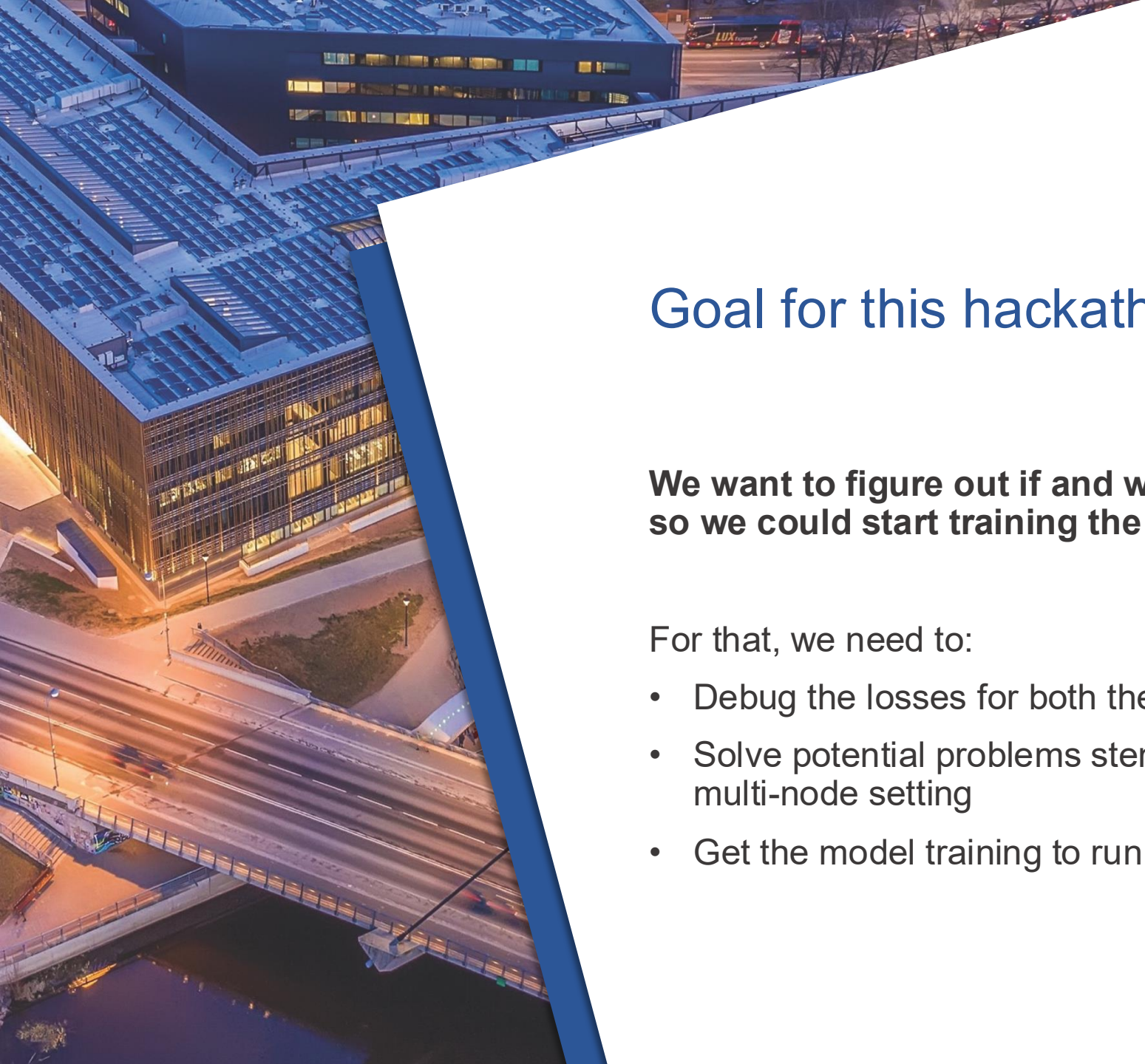
The 70B model is only feasible to train with **at least 3D parallelism** supported by the Megatron-LM (but not Huggingface Accelerate):

- Data parallelism
- Tensor parallelism
- Pipeline parallelism

We have found that **Megatron-LM** produces higher loss than **Huggingface Accelerate** in the same setting (model, configuration)

Thus, we **can't trust** the Megatron-LM training for the 70B model

We suspect that the problem might be caused by the **tensor parallel implementation**.



Goal for this hackathon

We want to figure out if and what is the problem with Megatron-LM so we could start training the 70B parameter model.

For that, we need to:

- Debug the losses for both the Megatron-LM and Huggingface Trainer
- Solve potential problems stemming from running Megatron-LM in the multi-node setting
- Get the model training to run on 100+ nodes in parallel.

Thank you!

Contact

Kairit Sirts (kairit.sirts@ut.ee), PI

Aleksei Dorkin (aleksei.dorkin@ut.ee), technical lead in the hackathon