

# Improving Estonian Language Capabilities in Open LLMs: Opportunities and Challenges

Kairit Sirts

Institute of CS, UT

DigiTS workshop, 19.11.2025

# Proprietary models



**Good in Estonian**

**but closed**

# Open models



Transparency



Science



Collaboration

# Multilingual open models

Reasons to do specific work on Estonian:

- We have a specific interest in Estonian
- We understand the language
- Research on improving specific language is valuable
- Develop and maintain technical competence



# EstLLM project

Data

Training

Evaluation

# Data insight 1: No reuse

Text A, Text A, Text A → 

Text A, Text B, Text C → 

# Data insight 2: Size $\neq$ Quality

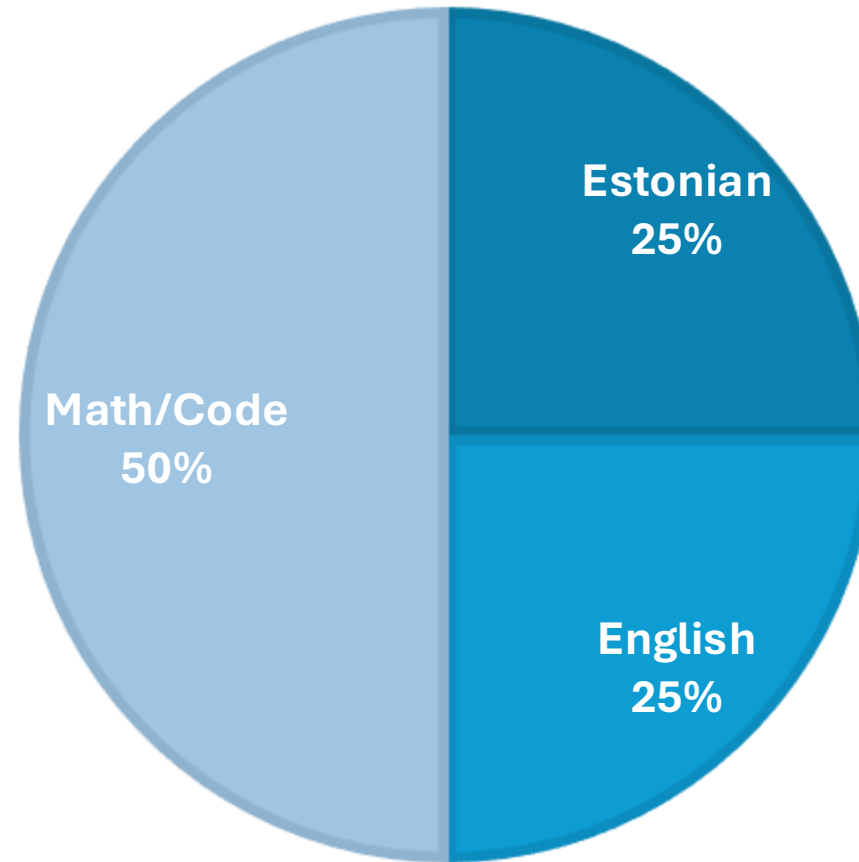


Large amount of similar texts  $\rightarrow$  **not helpful**



Smaller amount of diverse high quality texts  $\rightarrow$  **more helpful**

# Data insight 3: Mixing



# Training overview



# Instruction tuning

Pretraining

Sentence prefix



Next word

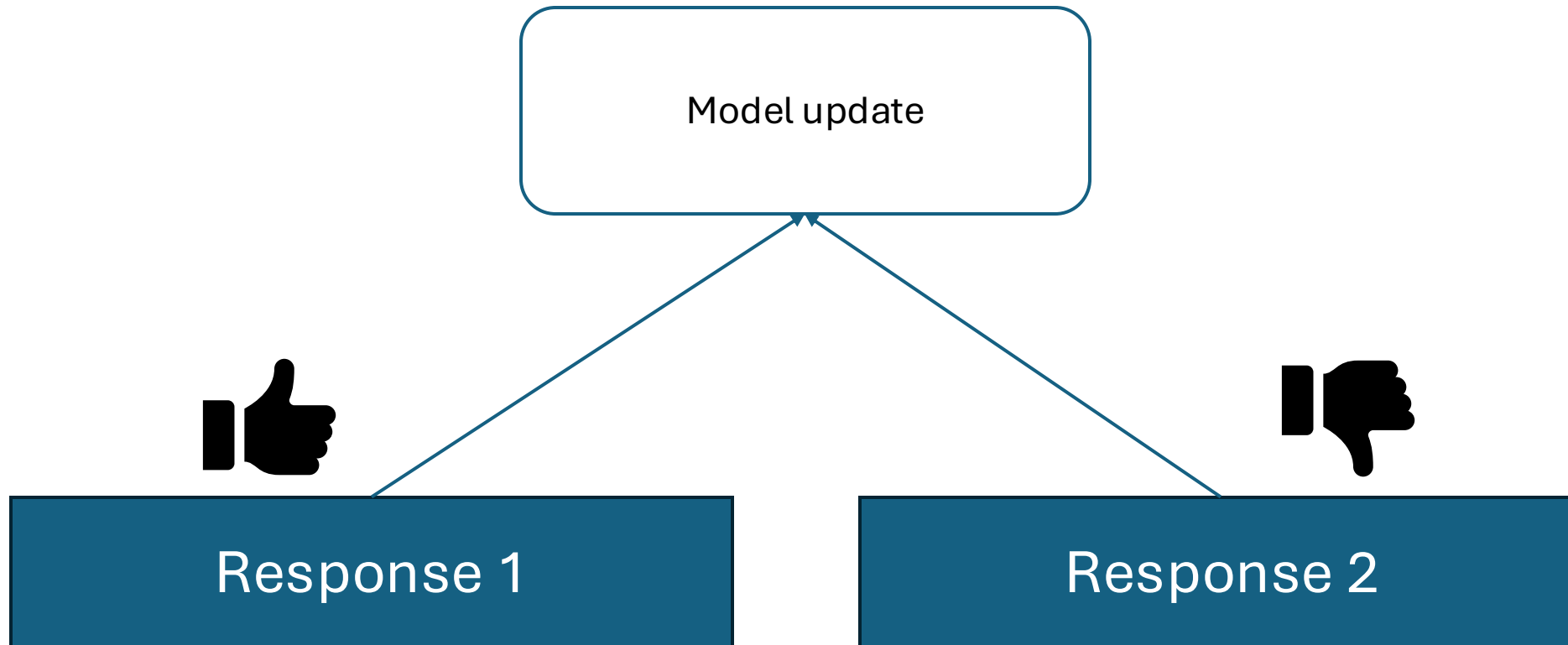
Instruction tuning

Instruction



Response

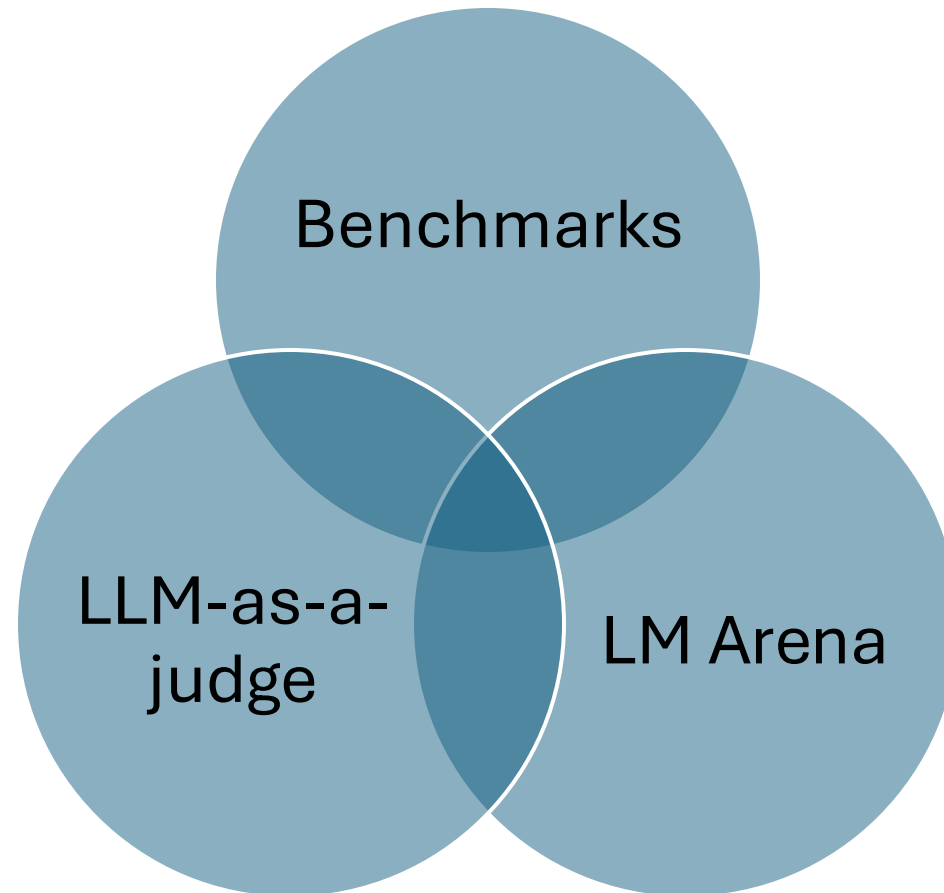
# Preference tuning



# EstLLM Prototype

- Continuously pre-trained from Llama-8B model
- Instruction-tuned and preference-tuned
  
- Available in [Huggingface/TartuNLP](#)
  
- Credits due to:
  - Taido Purason
  - Aleksei Dorkin
  - Emil Kalbaliyev

# Evaluation



# Benchmarks

A collection of Estonian Benchmarks

- Factual knowledge
- Reasoning
- Linguistics
- Information extraction
- QA
- Summarization
- etc

[Huggingface/TartuNLP](https://huggingface.co/collections/tartuNLP/estonian-llm-evaluation): <https://huggingface.co/collections/tartuNLP/estonian-llm-evaluation>  
[github/taltechnlp](https://github.com/taltechnlp): <https://github.com/taltechnlp/lm-eval-harness-tasks-estonian/>

# Benchmarks

Model	Mean Score	Exams Accuracy	Trivia Accuracy	Decl. Accuracy	Words Accuracy	Grammar Levenshtein*	News ROUGE-L	Speakers F1-score
<i>Small open models</i>								
Llama-3.1 8B EstLLM-0825	<b>0.431</b>	0.575	<b>0.425</b>	<b>0.811</b>	0.327	<b>0.275</b>	<b>0.152</b>	0.452
EuroLLM-9B-Instruct	0.338	<b>0.610</b>	0.381	0.644	<b>0.338</b>	0.248	0.104	0.038
Gemma-3-12b-It	0.288	0.736	0.320	0.356	0.289	0.197	0.119	0.000
Mistral Nemo Instruct (12B)	0.279	0.570	0.299	0.274	0.177	0.187	0.034	0.410
Gemma-2-9B-It	0.272	0.556	0.278	0.308	0.178	0.182	0.128	n/a**
Llama-3.1 8B Instruct	0.271	0.542	0.309	0.099	0.132	0.189	0.119	0.505
Qwen3-4B-Instruct-2507	0.251	0.553	0.276	0.056	0.090	0.136	0.105	<b>0.540</b>

# Human judges and LLM-as-a-judge

## Human evaluation dataset of 47 questions:

- Where in Estonia could one find a lot of chanterelles?
- What methods did Socrates use to challenge the common beliefs of his time?

Claude 3.7 Sonnet used as a judge

**Very high correlation (>0.9) between human judges and model judge**

# Issues with machine translated benchmarks

Estonian Winogrande dataset

- Errors in the English original get amplified
- Some questions need cultural adaptation
- Machine translation can shift meaning
- Even the best machine translation cannot improve the quality too much

# baromeeter.ai

🔍 Kliki siia, et näha võrdluses olevaid mudeleid. 🔥 Valikus uued mudelid!

💬 Mudel A

💬 Mudel B

👉 Kirjuta siia enda küsimus ja vajuta ENTER

Saada

# baromeeter.ai leaderboard

 ainult avalikud mudelid peida vanad versioonid

Koht ▲	Mudel ▲	Skoor ▲	95% UI ▲	Hääli ▲	Looja ▲	Litsents ▲	Teadmiste lõpp ▲
1	<a href="#">DeepSeek-V3 (0324)</a>	1456	+28/-34	853	DeepSeek	MIT	Unknown
1	<a href="#">Kimi-K2-Instruct</a>	1448	+29/-32	622	Moonshot AI	Modified MIT	Unknown
1	<a href="#">Llama 4 Maverick</a>	1434	+30/-34	656	Meta	Apache 2.0	Unknown
1	<a href="#">Gemma-3-27B-it</a>	1406	+29/-39	673	Google	Gemma	Unknown
1	<a href="#">llama-estllm-prototype-0825</a>	1377	+44/-34	489	TartuNLP	Llama 3.1	Unknown
2	<a href="#">Llama 4 Scout</a>	1379	+41/-38	658	Meta	Apache 2.0	Unknown
4	<a href="#">Meta-Llama-3.1-405B-Instruct</a>	1358	+33/-37	624	Meta	Llama 3.1 Community	2023/12
4	<a href="#">Qwen3-235B-A22B</a>	1348	+30/-40	659	Alibaba	Apache 2.0	Unknown

**High quality data is crucial!**

Thank you!