

The need for problem-first research for European industry and the public sector

LREC 2026 Industry Day



Toms Bergmanis

TildeOpen LLM
principal investigator
toms.bergmanis@tilde.lv

Central Claim

European LLM development is **driven by benchmarks** rather than by the ambition to solve real user problems.

Consequence

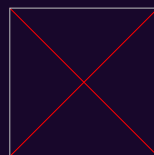


Europe creates research artefacts that are not useful
for the European Industry.

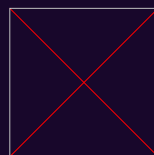
The Benchmarks: ARC/MMLU/MGSM

European LLMs are evaluated on translated variants of multiple-choice questions meant to test American students.

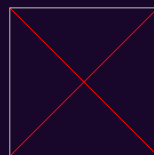
Problems:



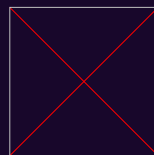
translationese – low-quality translations lacking linguistic diversity



topic bias – some tests are culturally irrelevant



no use-case – nobody ever uses 70b LLM to do a high-school test



distraction – end-user relevant aspects are not evaluated

Benchmark maxing:

Optimising a model to score well on a specific benchmark rather than to actually perform well on the task the benchmark was meant to measure.

[utter-project/EuroBlocks-SFT-Synthetic-1124](#)

[FIN-bench-v2: A Unified and Robust Benchmark Suite for Evaluating Finnish Large Language Models](#)

Benchmark maxing:

- **EuroBlocks-SFT-Synthetic:** provide synthetic ARC and MMLU style MCQ data
-

[utter-project/EuroBlocks-SFT-Synthetic-1124](https://github.com/utter-project/EuroBlocks-SFT-Synthetic-1124)

[FIN-bench-v2: A Unified and Robust Benchmark Suite for Evaluating Finnish Large Language Models](#)

Benchmark maxing:

- **EuroBlocks-SFT-Synthetic:** provide synthetic ARC and MMLU style MCQ data
-
- **EuroLLM and Apertus SFT data:** contains MCQ data directly optimising model performance on test-like benchmarks
-

[utter-project/EuroBlocks-SFT-Synthetic-1124](https://github.com/utter-project/EuroBlocks-SFT-Synthetic-1124)

[FIN-bench-v2: A Unified and Robust Benchmark Suite for Evaluating Finnish Large Language Models](#)

Benchmark maxing:

- **EuroBlocks-SFT-Synthetic:** provide synthetic ARC and MMLU style MCQ data

- **EuroLLM and Apertus SFT data:** contains MCQ data directly optimising model performance on test-like benchmarks

- **FIN-bench-v2 paper:** suggests best training data to max results on this bench.... which turns out to be English data machine-translated into Finnish 🤖

[utter-project/EuroBlocks-SFT-Synthetic-1124](#)

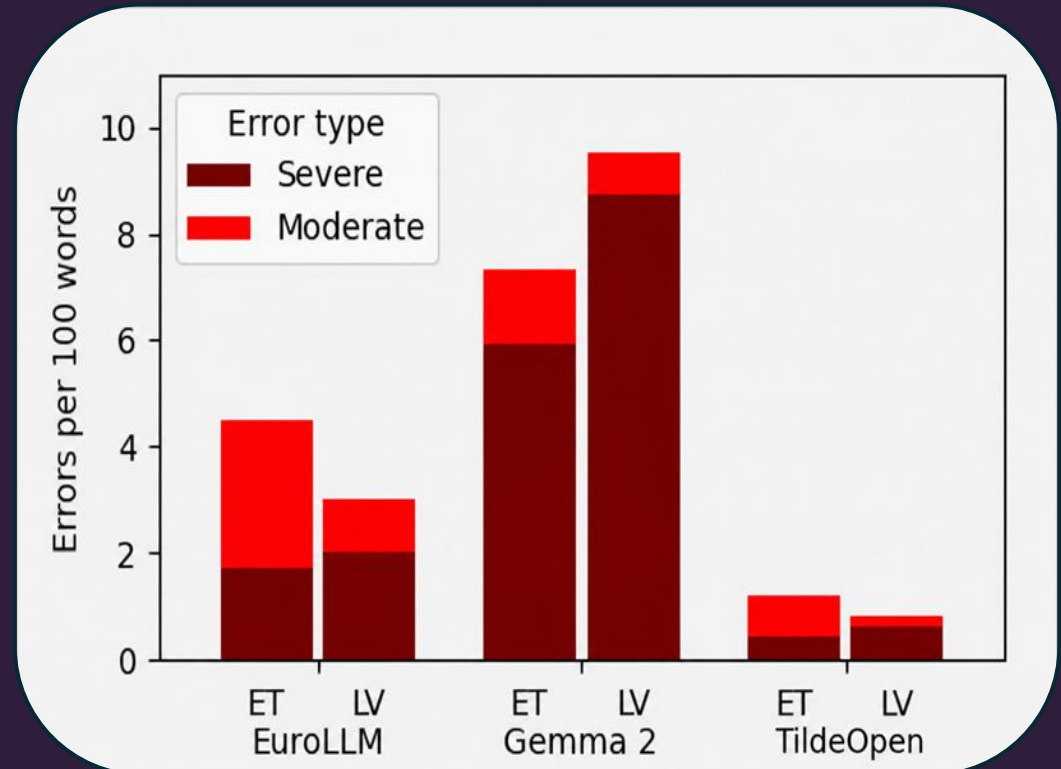
[FIN-bench-v2: A Unified and Robust Benchmark Suite for Evaluating Finnish Large Language Models](#)

Consequences:

The time researchers spend on benchmark-maxing, they could use for real-world problem solving.

Consequences: weak multilingualism

- No need to be proficient in EU languages to do well on translated tests
- Improving proficiency in EU languages might even harm the English-derived benchmark-maxing
- Models produce rudimentary linguistic errors every 20-30 words
- LLM supports 30 languages, can spell in 4
- Many governmental clients can't use such models



TildeOpen LLM: Leveraging Curriculum Learning to Achieve Equitable Language Representation

Consequences: unreliable RAG

When answering questions in a document context, LLM has to:

- 1 Answer according to the document context when information is present

- 2 Abstain when information is missing

All models can answer questions, reasonably well

The radar chart displays the performance of seven LLMs across 12 languages. The languages are FI, ET, EN, CS, BG, UK, SL, SK, RU, MK, LV, and LT. The performance is measured on a scale from 0 to 100. The LLMs are ALIA (red), Apertus (orange), EuroLLM (green), Gemma3 (blue), Gemma4 (purple), Mistral (brown), and TildeOpen (cyan). The chart shows that EuroLLM and Gemma4 generally perform best across most languages, while ALIA and Apertus perform best in BG and UK. TildeOpen and Gemma3 show lower performance across most languages.

Language	ALIA	Apertus	EuroLLM	Gemma3	Gemma4	Mistral	TildeOpen
FI	85	88	95	80	92	90	82
ET	90	92	98	85	95	93	88
EN	88	90	95	82	92	90	85
CS	85	88	92	80	90	88	82
BG	95	98	90	85	92	90	88
UK	95	98	90	85	92	90	88
SL	85	88	92	80	90	88	82
SK	85	88	92	80	90	88	82
RU	85	88	92	80	90	88	82
MK	85	88	92	80	90	88	82
LV	85	88	92	80	90	88	82
LT	85	88	92	80	90	88	82

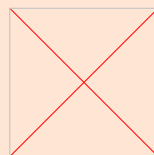
European models in the centre
ignore context and answer
something anyway

[illegible]

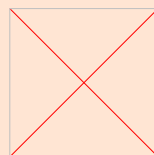
Consequences: opportunity cost

Researchers publish information on models' performance on American high-school tests, translated into European languages.

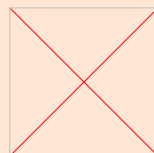
But they don't report:



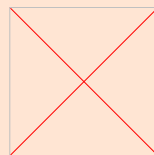
Results on linguistic correctness of generated texts [presentation]



Data on the model's long context ability [reading comprehension]



Adherence to information in context [trustworthiness]



Decoding efficiency in different European languages [cost efficiency]

Conclusions

Useful European AI starts with European user problems:
build for users, not benchmarks!

Thank you!



Toms Bergmanis

TildeOpen LLM principal investigator
toms.bergmanis@tilde.lv