# 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 EvaluatingFinnish Large Language Models*

# Benchmark maxing:

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

*utter-project/EuroBlocks-SFT-Synthetic-1124*

*FIN-bench-v2: A Unified and Robust Benchmark Suite for EvaluatingFinnish 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*

*FIN-bench-v2: A Unified and Robust Benchmark Suite for EvaluatingFinnish 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*

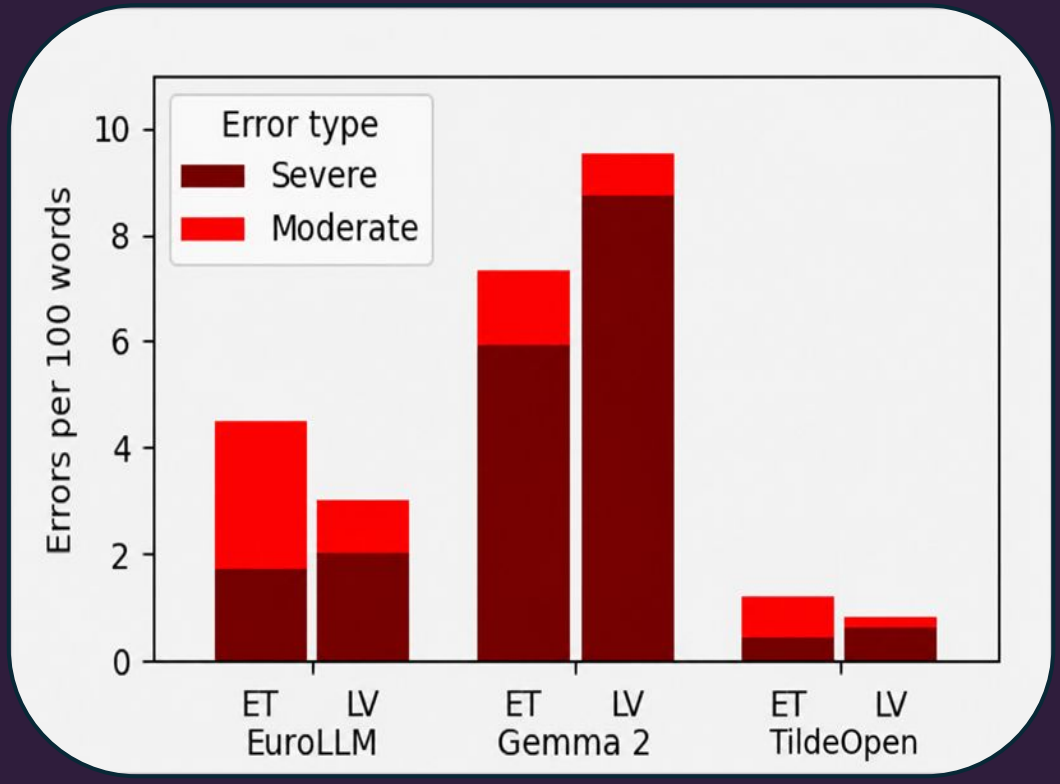
*FIN-bench-v2: A Unified and Robust Benchmark Suite for EvaluatingFinnish Large Language Models*

# 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 AchieveEquitable Language Representation

# Consequences: unreliable RAG

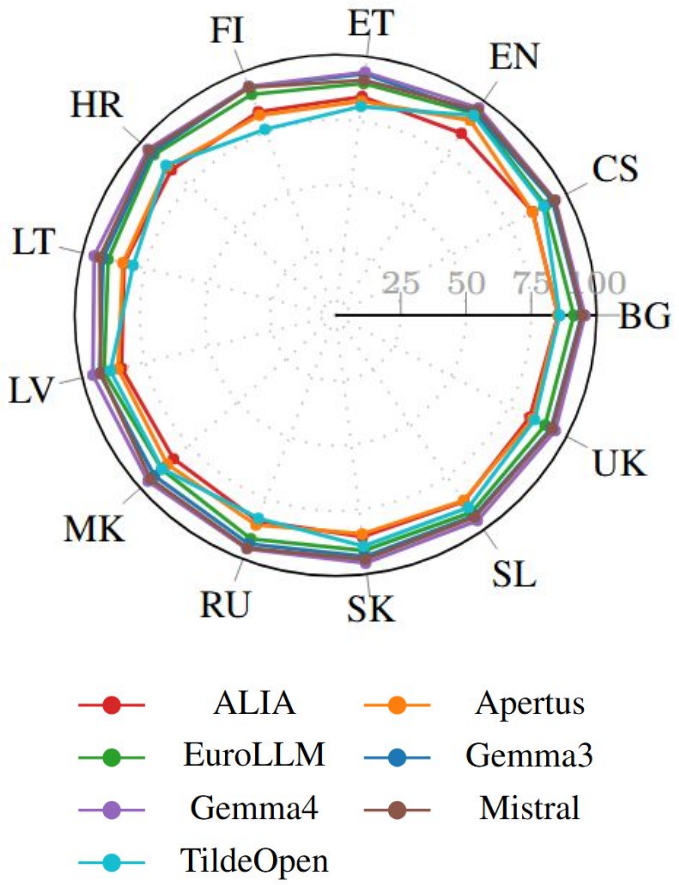
When answering questions in a document context, LLM has to:

1. Answer according to the document context when information is present
2. Abstain when information is missing

# Consequences: unreliable RAG

All models can answer questions, reasonably well

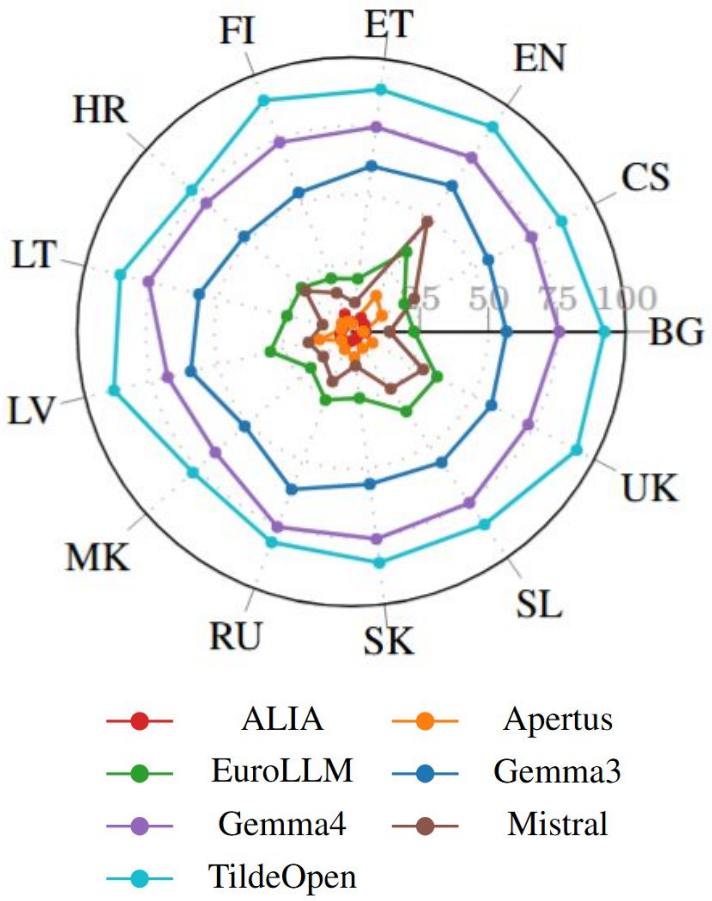
Question answering accuracy:



# Consequences: unreliable RAG

European models in the centre ignore context and answer something anyway

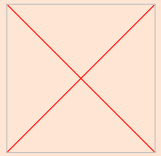
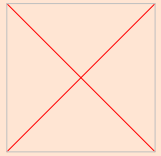
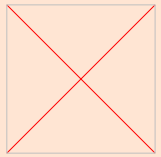
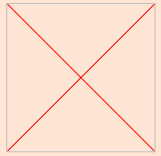
**Abstention** answering accuracy:



# 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