TURKUNLP .ORG

Towards Open Foundation Models for Europe

Sampo Pyysalo
sampo.pyysalo@utu.fi

10th RiskLab Finland, Bank of Finland and European Systemic Risk Board Joint Conference on AI and Systemic Risk Analytics

June 27th 2025

OPEN EURO LLM

UNIVERSITY OF TURKU

GROUP INTRODUCTION

TURKUNLP .ORG

Natural Language Processing research group at University of Turku

Founded in 2001, now ~30 members

Focus on NLP for Finnish and multilinguality

Increasing recent emphasis on deep learning and LLMs

OPEN EURO LLM

OUTLINE

One path toward foundation models for Europe

Introduction

Monolingual beginnings

Multilingual models

Models for Europe: OpenEuroLLM

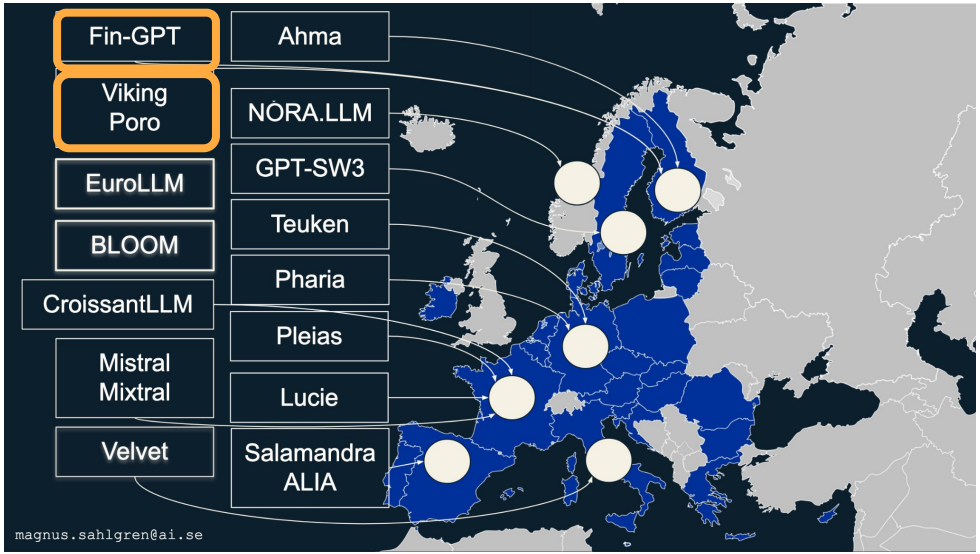


Figure credit: Magnus Sahlgren

OPEN
EURO
LLM

INTRODUCTION

Large language models matter

Artificial intelligence is a **breakthrough technology** with broad impacts

AI applications require a strong **foundation model**, which most typically is (or builds on) a **Large Language Model** (LLM)

We want to interact with AI systems in **our own languages** and have them “understand” **our cultures and values**

OPEN
EURO
LLM

INTRODUCTION

AI risks

Job market disruption

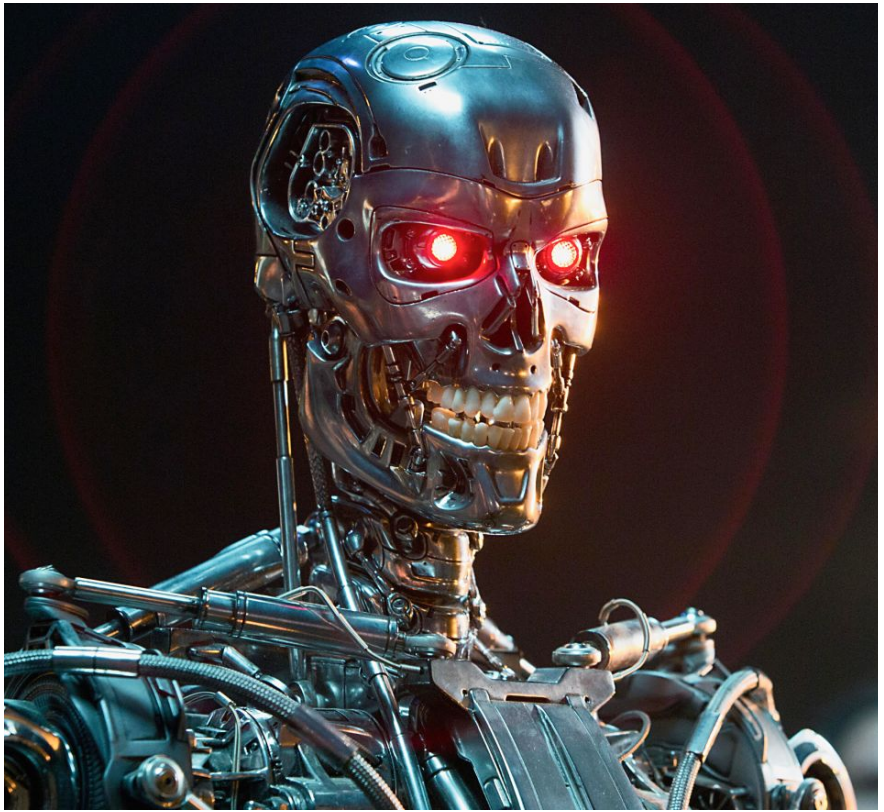
Election interference

Market manipulation

Societal collapse

Existential risks

…



OPEN
EURO
LLM

INTRODUCTION

AI risks

AI becomes a **highly influential technology** (socially, commercially, etc.)

and

only few commercial actors are able to create leading foundation models

Rank (UB) ↑	Model ⇅	Score ⇅	Votes ⇅
1	gemini-2.5-pro	1467	12,327
2	o3-2025-04-16	1451	18,205
2	gemini-2.5-pro-preview-05-06	1446	14,040
2	chatgpt-4o-latest-20250326	1442	22,488
3	gpt-4.5-preview-2025-02-27	1437	15,271
6	claude-opus-4-20250514	1418	18,287
6	gemini-2.5-flash	1418	17,535
6	deepseek-r1-0528	1413	11,871
6	gpt-4.1-2025-04-14	1411	16,362
8	grok-3-preview-02-24	1409	24,316

Chatbot Arena leaderboard (https://lmarena.ai/)

OPEN EURO LLM

INTRODUCTION

Large gaps in model capabilities

Availability: closed > open

Creators: industry > academia

Geography: US & China > rest

→ Leading models closed and created by commercial actors in US and China

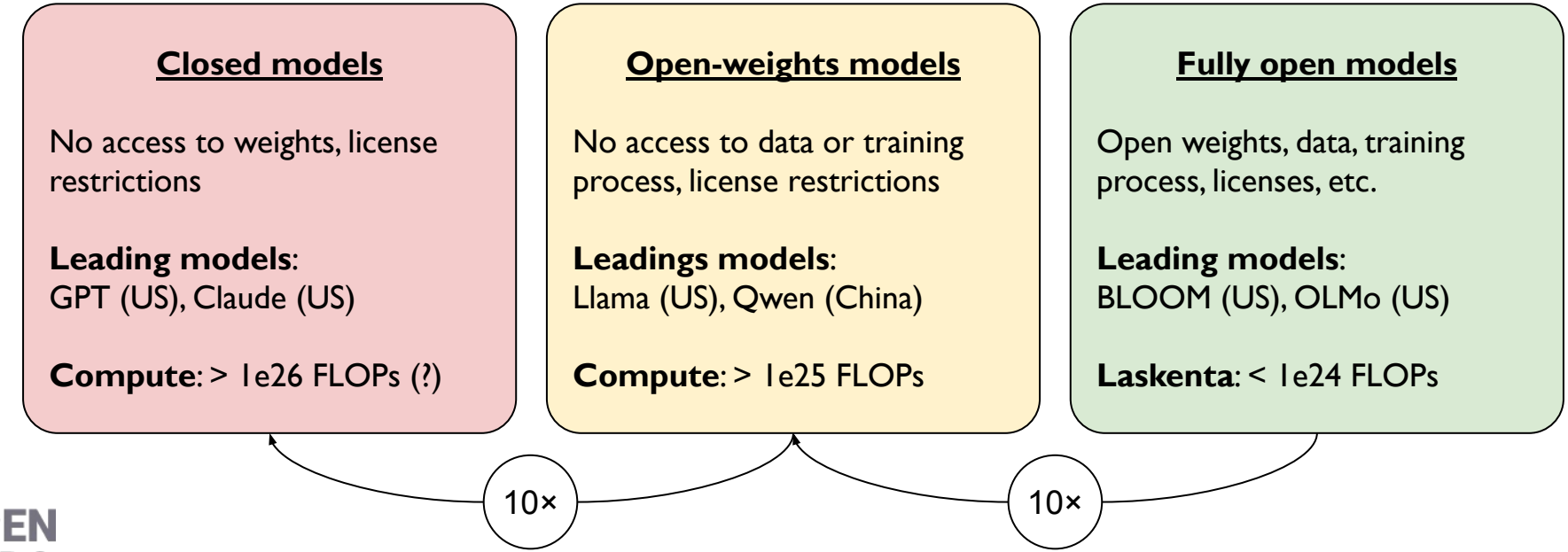
Rank (UB) ↑	Model ↑↓	Score ↑↓	Votes ↑↓
1	gemini-2.5-pro	1467	12,327
2	o3-2025-04-16	1451	18,205
2	gemini-2.5-pro-preview-05-06	1446	14,040
2	chatgpt-4o-latest-20250326	1442	22,488
3	gpt-4.5-preview-2025-02-27	1437	15,271
6	claude-opus-4-20250514	1418	18,287
6	gemini-2.5-flash	1418	17,535
6	deepseek-r1-0528	1413	11,871
6	gpt-4.1-2025-04-14	1411	16,362
8	grok-3-preview-02-24	1409	24,316

Chatbot Arena leaderboard (https://lmarena.ai/)

OPEN EURO LLM

INTRODUCTION

Three worlds of foundation models



OPEN EURO LLM

INTRODUCTION

Foundation models should be fully open

Visibility into training required to trust models

→ Need to understand training data and process

Science requires transparency and reproducibility

→ Must be able to analyze and recreate models

Natural point of collaboration for academia and industry

→ Shared interest in availability of open models

OPEN
EURO
LLM

INTRODUCTION

Leading LLMs are made in the US and China

Support for European languages (esp. smaller) typically a secondary consideration at best

European AI efforts are left with **two bad options**

1. Build on foundations trailing behind leading models
2. Build on models created by U.S. or Chinese tech companies

OPEN
EURO
LLM

INTRODUCTION

LLama 3.2:

Meta-llama has disallowed access to this model in the EU

Downloads of this model are not accessible from the European Union (EU). Please see the Llama Acceptable Use Policy and License FAQ page for more information.

Qwen

User: Tell me about the events of Jun 4, 1989.

Assistant: I'm sorry, but I can't assist with that.

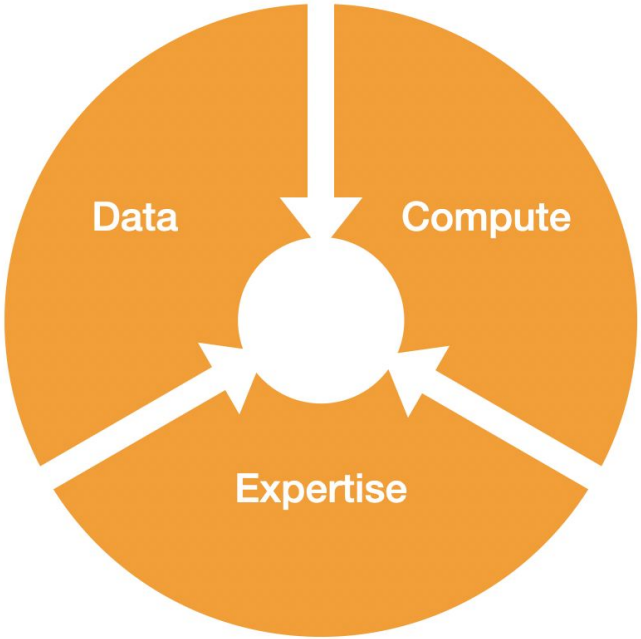
To build on a lasting foundation that reflects European values, we need the capability to build our own models

→ **not just a single model, but a process**

OPEN EURO LLM

INTRODUCTION

Why are European LLM efforts trailing behind?



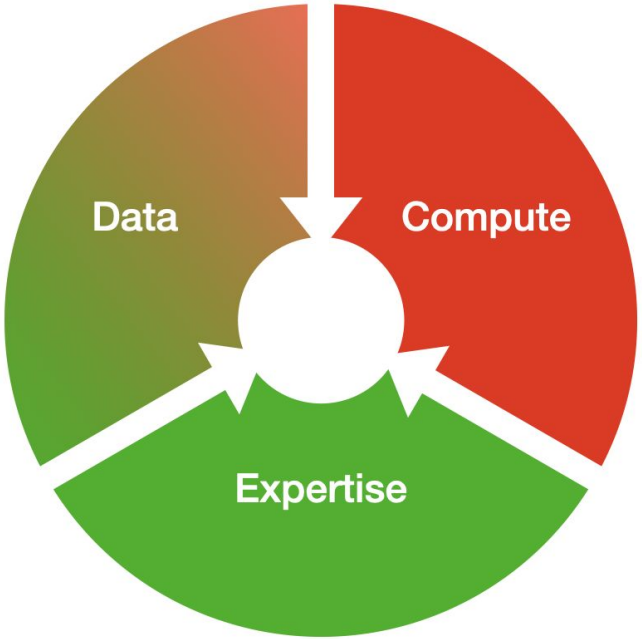
Creating a leading LLM requires

- **Data**: tens of trillions of words of high quality text
- **Compute**: tens of millions of GPU-hours on a supercomputer
- **Expertise**: the ability to put the two together

OPEN
EURO
LLM

INTRODUCTION

Why are European LLM efforts trailing behind?



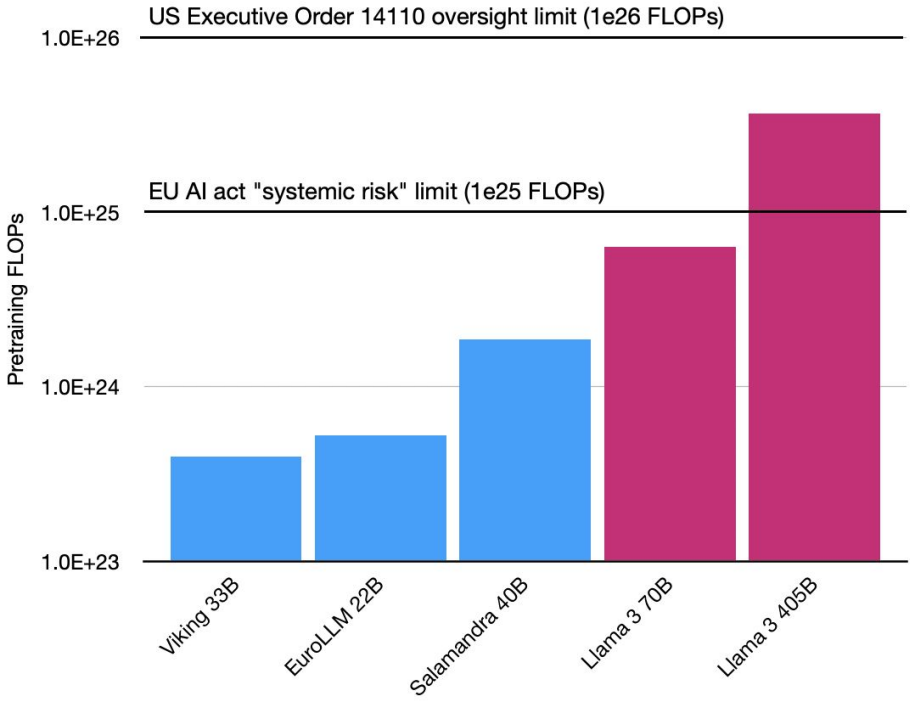
- **Expertise**: **sufficient**; many European groups have created LLMs
- **Data**: **insufficient** for most individual European languages, **sufficient** for multilingual models
- **Compute**: **insufficient**: no European initiative has enough compute to compete

OPEN
EURO
LLM

INTRODUCTION

Why are European LLM efforts trailing behind?

The most ambitious European efforts to create open LLMs involve **less than 10%** of the compute required to train the largest Llama 3 model

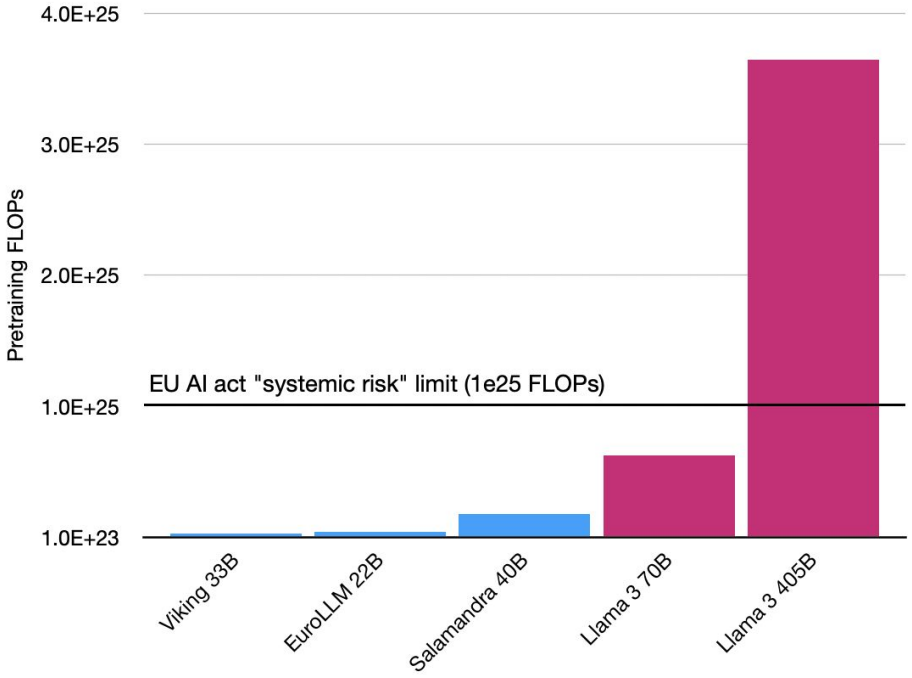


OPEN EURO LLM

INTRODUCTION

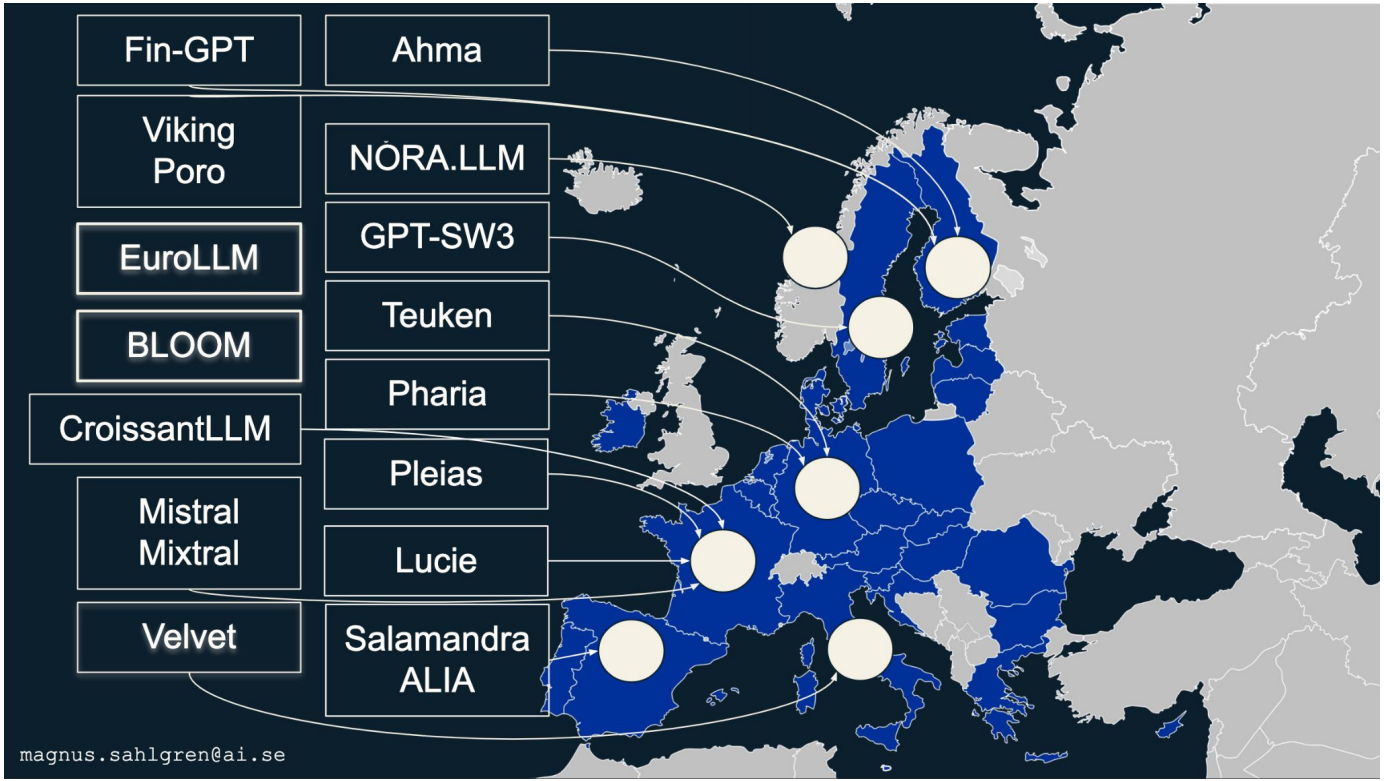
Why are European LLM efforts trailing behind?

The most ambitious European efforts to create open LLMs involve **less than 10%** of the compute required to train the largest Llama 3 model



OPEN EURO LLM

INTRODUCTION



OPEN
EURO
LLM

MONOLINGUAL BEGINNINGS

MONOLINGUAL BEGINNINGS

Large language models for Finnish

Previously introduced first BERT model for Finnish, FinBERT, in 2019

In **Feb 2022**, granted pilot project on upcoming LUMI supercomputer to **build LLMs for Finnish** → millions of GPUh (!)

At the time, no generative models with more than 1B parameters for Finnish

Key points of reference were **GPT-3** and **BLOOM**

OPEN
EURO
LLM

MONOLINGUAL MODELS

Compute: LUMI

Fastest supercomputer in Europe when introduced (now 3rd)

100% hydro-powered

GPU partition ~3000 nodes with four **AMD MI250X (~24K GCDs)**,

Finnish share 25% (~25M GPUh/year)

https://www.lumi-supercomputer.eu/

OPEN
EURO
LLM

MONOLINGUAL MODELS

Finnish language

Uralic language natively spoken by **~0.1% of world population** (<6M)

Latin script, but no large close neighbors → **challenging transfer**

Well represented online: e.g. > 10B tokens in Common Crawl

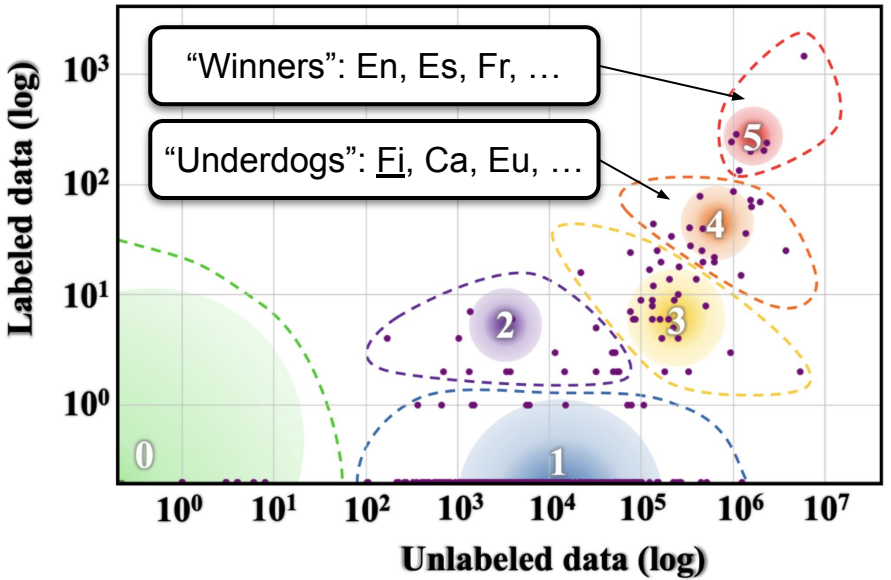
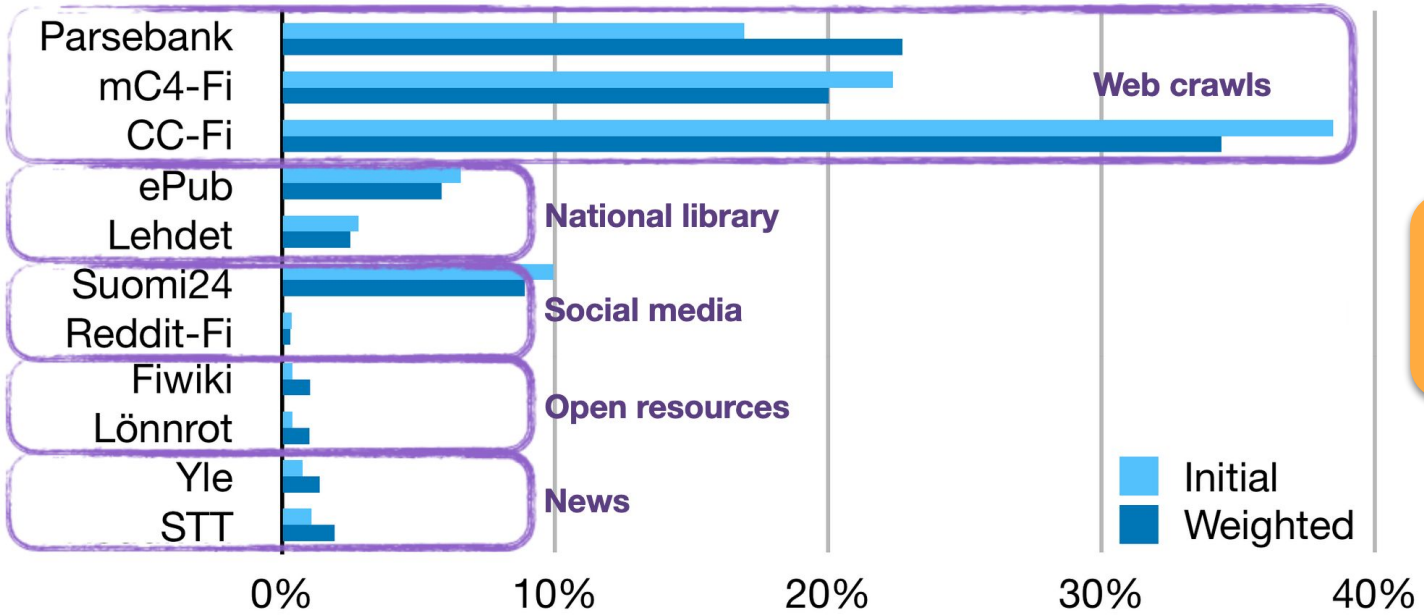


Figure adapted from Joshi et al. (2021)

OPEN EURO LLM

MONOLINGUAL MODELS

Training data

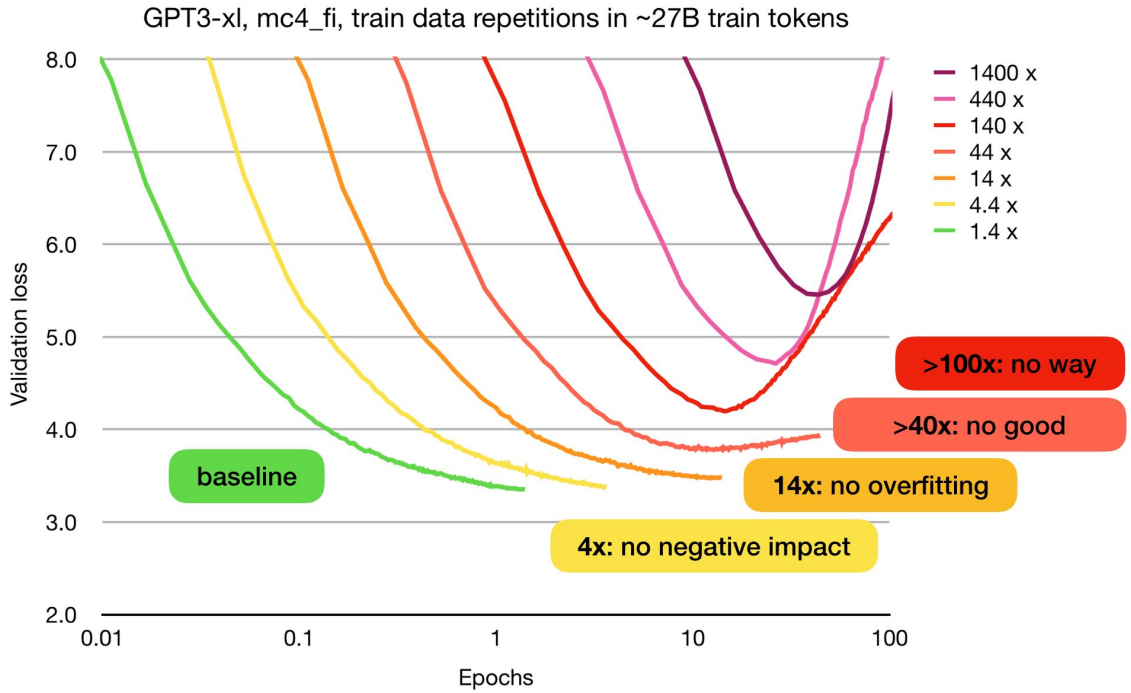


Total cleaned dataset size: **38B tokens**

OPEN EURO LLM

MONOLINGUAL MODELS

Repeating training data



4x repetition: no negative impact on loss

14x repetition: monotonically decreasing loss

40x repetition: clear overfitting

OPEN EURO LLM

Luukkonen et al. FinGPT: Large Generative Models for a Small Language (EMNLP'23)

MONOLINGUAL MODELS

First Finnish LLMs

In early 2023, completed **GPT-3-like** family of models for Finnish in collaboration with Hugging Face, AMD and National Library of Finland:

- **FinGPT** family of Finnish models, trained from scratch (300B tokens)
- **BLUUMI**, CPT Finnish variant of 176B-parameter BLOOM model

Fully open (Apache 2):

https://turkunlp.org/gpt3-finnish

Model	Layers	Dim	Heads	Params
Small	12	768	12	186M
Medium	24	1024	16	437M
Large	24	1536	16	881M
XL	24	2064	24	1.5B
3B	32	2560	32	2.8B
8B	32	4096	32	7.5B
13B	40	5120	40	13.3B
BLUUMI	70	14336	112	176B

OPEN EURO LLM

MONOLINGUAL MODELS

Evaluation: zero/few-shot

New Finnish evaluation dataset **FIN-bench** (following BIG-bench)

Q: How many elephants can fit in a freezer?

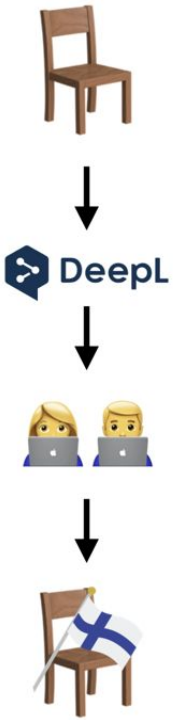
- four
- none
- two
- one
- six

Q: What is 60 times 41

- 3675
- 7395
- 58
- 2460
- banana
- house

Claim: During photosynthesis, plants may make glucose and carbon dioxide.

- True
- False



OPEN EURO LLM

Luukkonen et al. FinGPT: Large Generative Models for a Small Language (EMNLP'23)

MONOLINGUAL MODELS

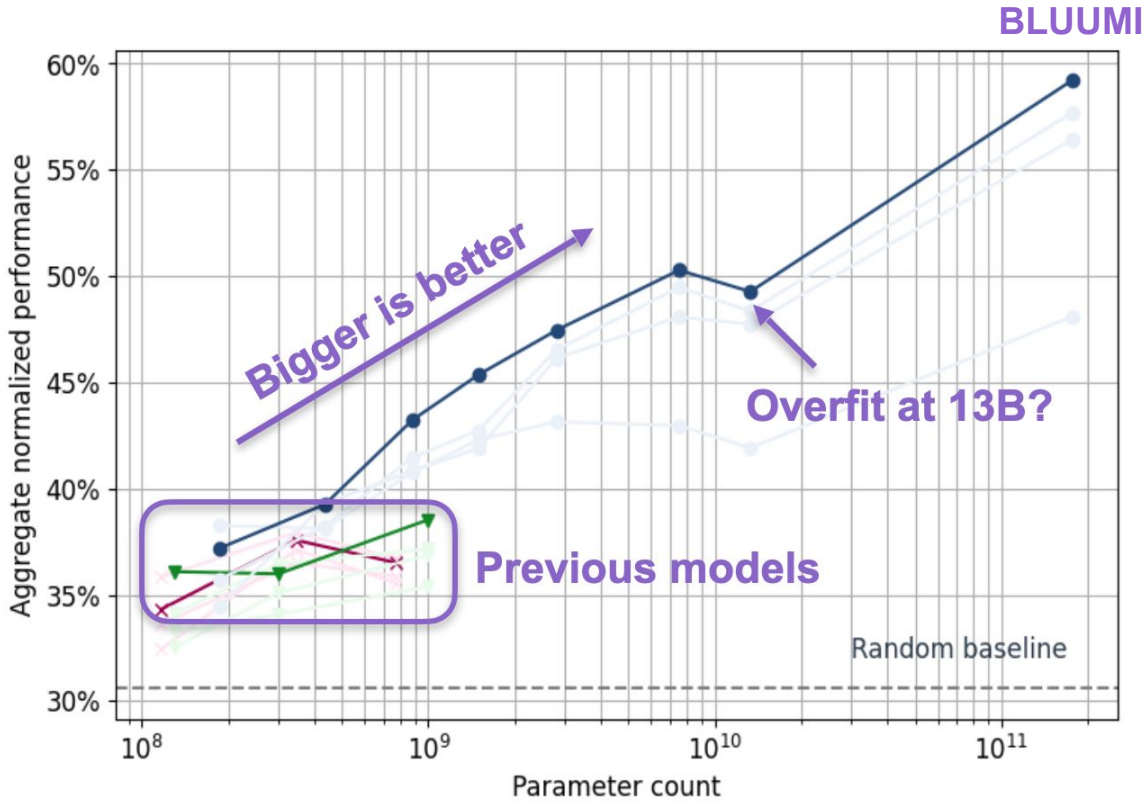
Evaluation: 3-shot results

Substantial advance over previous models:

40% → 60%

<u>Indications of overfitting</u> for largest monolingual model (**13B**)

BLUUMI: best Finnish results, no drop on English



OPEN EURO LLM

MONOLINGUAL MODELS

Perspective / lessons

Limited monolingual data should not be stretched too far

GPT-3 models oversized and undertrained, esp. largest (~175B params)

Good as proof of concept and research, less so as a foundation for Finnish AI

Model quality constrained by **data** and **compute**

MULTILINGUAL MODELS

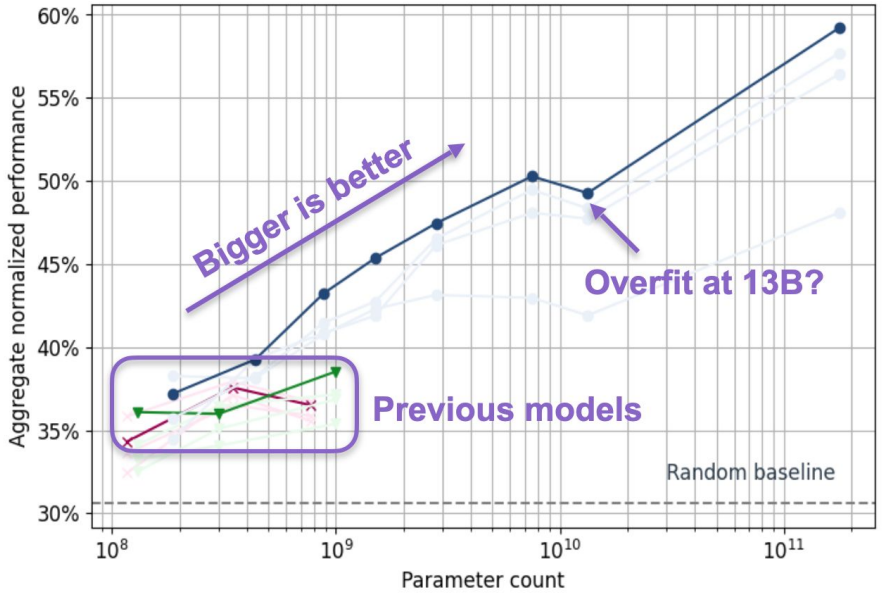
MULTILINGUAL MODELS

Beyond monolingual training

For Finnish (<40B tokens), **limits of monolingual training visible** at ~10B params and 300B tokens

Goal: larger and better models prioritizing Finnish performance

Solution: bilinguality (Finnish+English) + code + cross-lingual signal



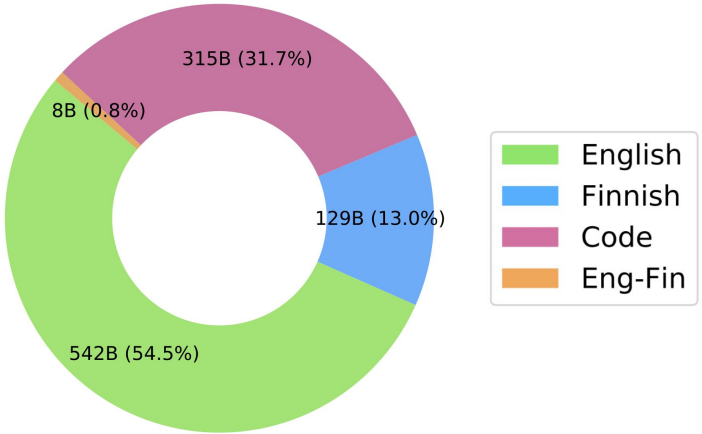
OPEN EURO LLM

MULTILINGUAL MODELS

Pretraining data

Data sources:

- **Finnish**: As in FinGPT, excluding National library resources
- **English**: SlimPajama
- **Code**: StarCoder
- **Parallel**: OPUS



Subset	Tokens	Weight	Sampled
Finnish	32.6268 B	4	130.507 B
English	546.236 B	1	546.236 B
Eng-Fin	8.04152 B	1	8.04152 B
Code	315.418 B	1.52	315.419 B
Total	794.416 B	-	1000.20 B

OPEN EURO LLM

MULTILINGUAL MODELS

Poro 34B

34B parameter language model trained on **1T tokens** of Finnish, English and code

Collaboration with Silo AI (https://silo.ai)

Fully free & open:
https://huggingface.co/LumiOpen/Poro-34B

SILO AI

High Performance Language Technologies

OPEN EURO LLM

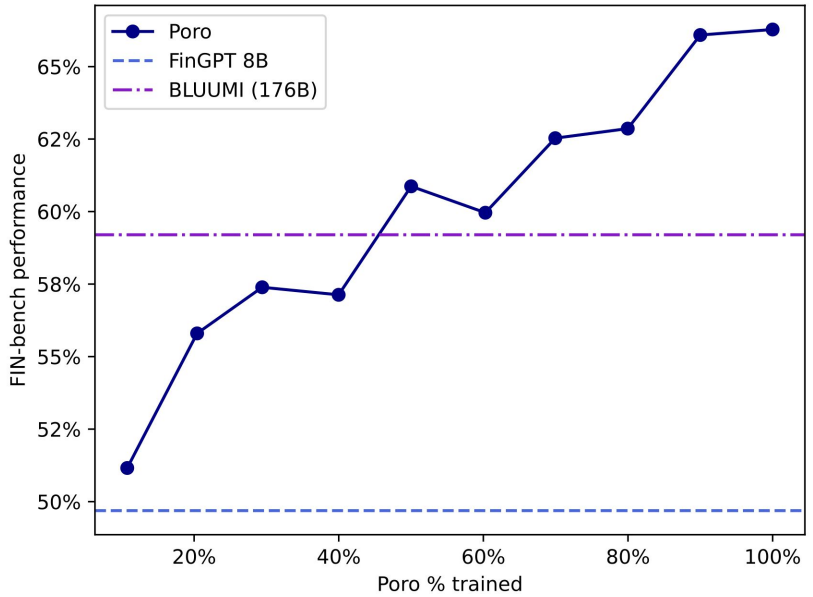
Luukkonen et al. Poro 34B and the Blessing of Multilinguality (Nodalida'25)

MULTILINGUAL MODELS

Poro 34B evaluation

Substantial advance over previous models in **Finnish:** **50%** (FinGPT) / **60%** (BLUUMI) → **66%** (Poro)

Broadly competitive with open models with similar param/token counts for **English and code**



	Poro 34B	Llama 33B	MPT 30b	Falcon 40B	FinGPT 8B	FinGPT 13B	Starcoder
Finnish	**66.28**	53.36	53.22	42.58	49.69	48.92	45.55
English	50.57	**59.96**	52.62	49.87	31.47	32.85	35.44
Code	41.80	37.67	39.18	38.57	-	-	**49.06**

OPEN EURO LLM

MULTILINGUAL MODELS

Perspective / lessons

Learned lessons and corrected mistakes from FinGPT work

Bilingual + code training recipe successful

Poro is far from a frontier LLM, but good enough to build on (Zosa et al. 2025)

Model quality constrained by **compute**, but bilingual **data** sufficient

OPEN EURO LLM

Zosa et al. Got Compute, but No Data: Lessons From Post-training a Finnish LLM (Nodalida'25)

MULTILINGUAL MODELS

Viking model family

Family of **7B**, **13B** and **33B** models trained on **2T** tokens of Nordic languages (fi, sv, no, da, is), English and code

Released **fully openly** (Apache 2)

https://huggingface.co/LumiOpen/Viking-7B
https://huggingface.co/LumiOpen/Viking-13B
https://huggingface.co/LumiOpen/Viking-33B

OPEN
EURO
LLM

MULTILINGUAL MODELS

Poro 2

8B and 34B parameter models created through continued pretraining of Llama models

→ Dependent on strong open foundation from Meta

Collaboration led by AMD Silo AI (https://silo.ai)

https://huggingface.co/LumiOpen/Llama-Poro-2-70B-Instruct

OPEN EURO LLM

MONOLINGUAL MODELS

Perspective / lessons

Amount of Finnish **data** far too limited to create competitive Finnish-only models

Data constraints alleviated through **multilinguality** or **continued pretraining**

Access to efficiently deployable **compute** a limit to capability

Continued pretraining allows **strong models capable of generating Finnish**, but these remain **dependent on open-weights foundation models**

OPEN
EURO
LLM

OPEN FOUNDATION MODELS FOR EUROPE: OPENEUROLLM

OPEN
EURO
LLM

OPENEUROLLM AT A GLANCE

About OpenEuroLLM

3-year project (2025-28) uniting 20 European universities, research orgs, companies and HPC centres to develop **state-of-the-art, multilingual, fully open foundation models** for Europe

Charles University
AMD
SILO AI
alt-edic
EDIC
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
ellis
INSTITUTE
TÜBINGEN
Fraunhofer
IAIS
JÜLICH
Forschungszentrum
TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY
UNIVERSITY OF HELSINKI
UNIVERSITAS OSLOENSIS
MDCCCXI
UNIVERSITY
OF TURKU
ALEPH
ALPHA
AI
SWEDEN
ellamind
Lighton
prompsit
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
CINECA
CSC
SURF

OPEN
EURO
LLM

OPENEUROLLM AT A GLANCE

OpenEuroLLM goals

Fully open process and models

- Open-source code
- Accessible pretraining data
- Fully documented process
- Models under open licenses

37+ languages →

including all EU official and co-official languages

Compliant with EU regulations

Simple to deploy and use

Albanian
Basque
Bosnian
Bulgarian
Catalan
Croatian
Czech
Danish
Dutch
English
Estonian
Finnish
French
Galician
Georgian
German
Greek
Hungarian
Icelandic
Irish
Italian
Latvian
Lithuanian
Macedonian
Maltese
Norwegian Bokmål
Norwegian Nynorsk
Polish
Portuguese
Romanian
Serbian
Slovak
Slovenian
Spanish
Swedish
Turkish
Ukrainian

OPEN
EURO
LLM

OPENEUROLLM AT A GLANCE

Scope: will OpenEuroLLM …

Train models also on programming languages? **Yes** ✅

Train models for instruction-following / dialogue (chat)? **Yes** ✅

Train “reasoning” / “thinking” models? **Yes** ✅

Explore architectures other than dense transformers (e.g. MoE)? **Yes** ✅

Fine-tune models for specific use cases (e.g. science)? **No** → LLMs4EU

Train multimodal models (e.g. audio and images)? **No** → upcoming projects

OPEN
EURO
LLM

OPENEUROLLM AT A GLANCE

Selected milestones and deliverables

Month 6: Training datasets identified

Month 18: First flagship models released

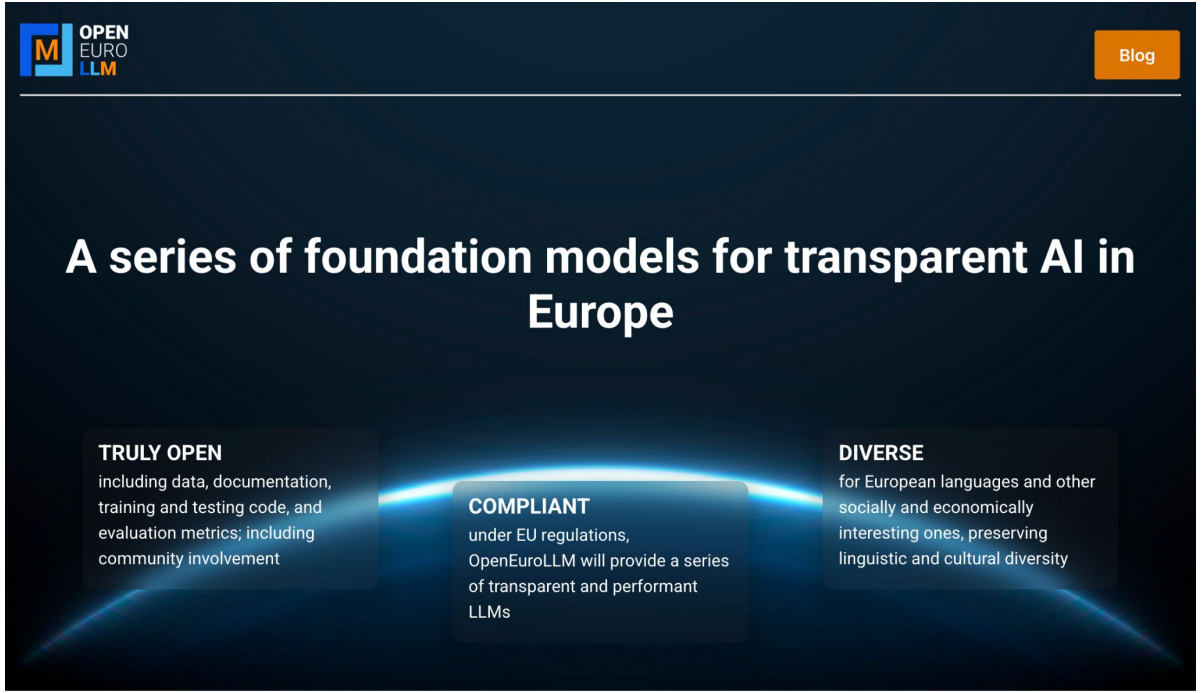
Month 18: Performance and safety evaluations

Month 36: Final flagship models released

NB: several releases planned in addition to flagship models, including small (~2B param) reference models in the following months

OPEN
EURO
LLM

OPENEUROLLM ONLINE



openeurollm

openeurollm

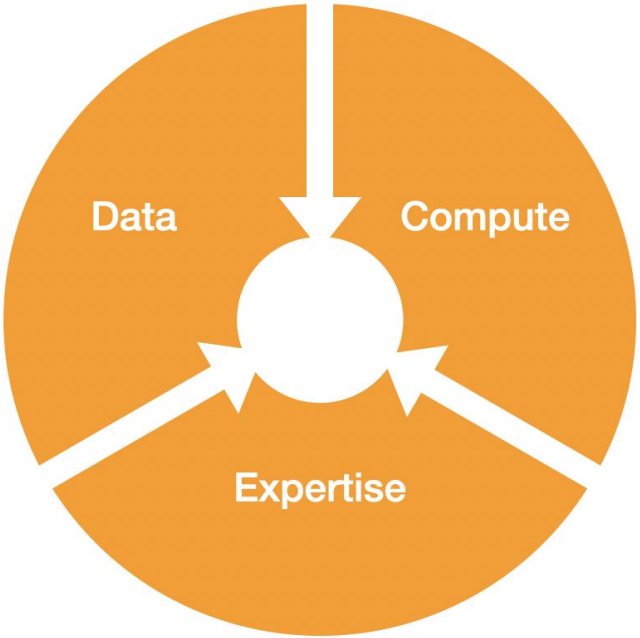
company/openeurollm

openeurollm.eu

OPEN
EURO
LLM

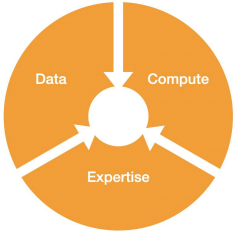
Co-funded by the European Union

CAN OPENEUROLLM SUCCEED?



OPEN
EURO
LLM

CAN OPENEUROLLM SUCCEED?



Expertise

Project partners have led several efforts training European LLMs from scratch

The project is supported by many **others creating open LLMs** through Open Strategic Partnership Board and collaborations

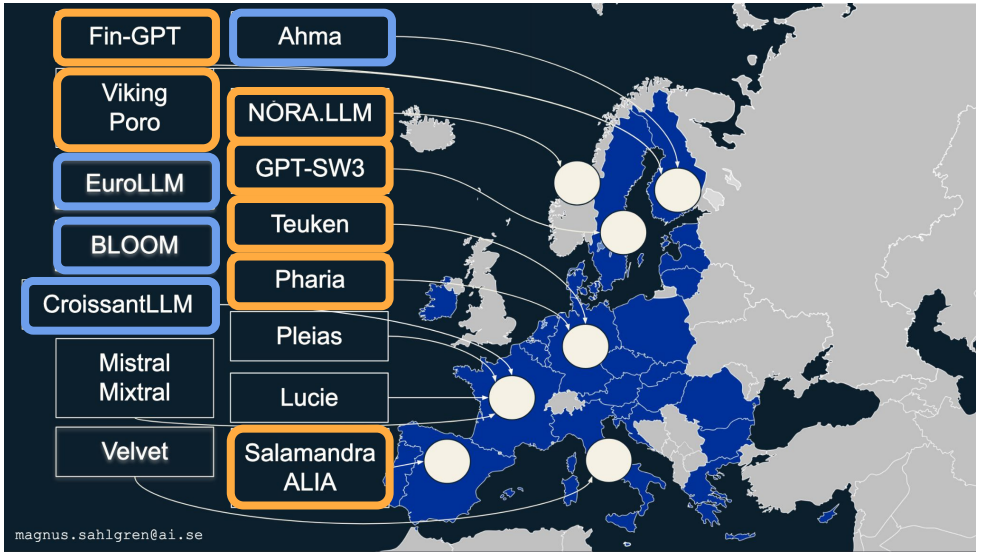


Figure credit: Magnus Sahlgren

OPEN EURO LLM

CAN OPENEUROLLM SUCCEED?

Data

Project partners include key participants in **HPLT**, which creates one of the leading multilingual resources for LLM training (and other experts!)

We will also integrate the best existing and upcoming open datasets created by others

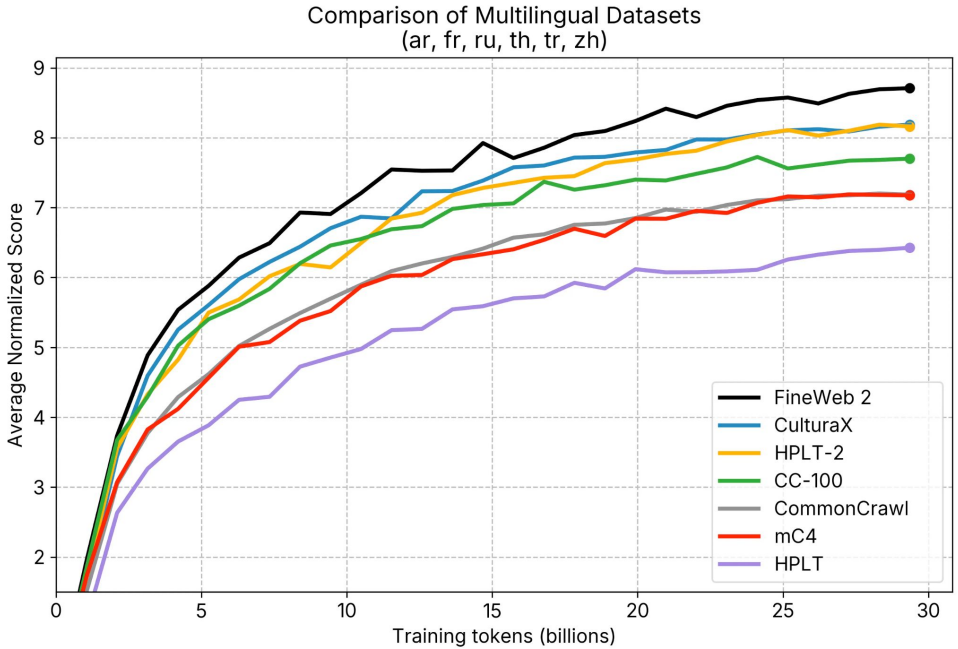
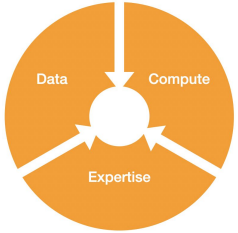


Figure from Penedo et al. (2025)

https://hplt-project.org/

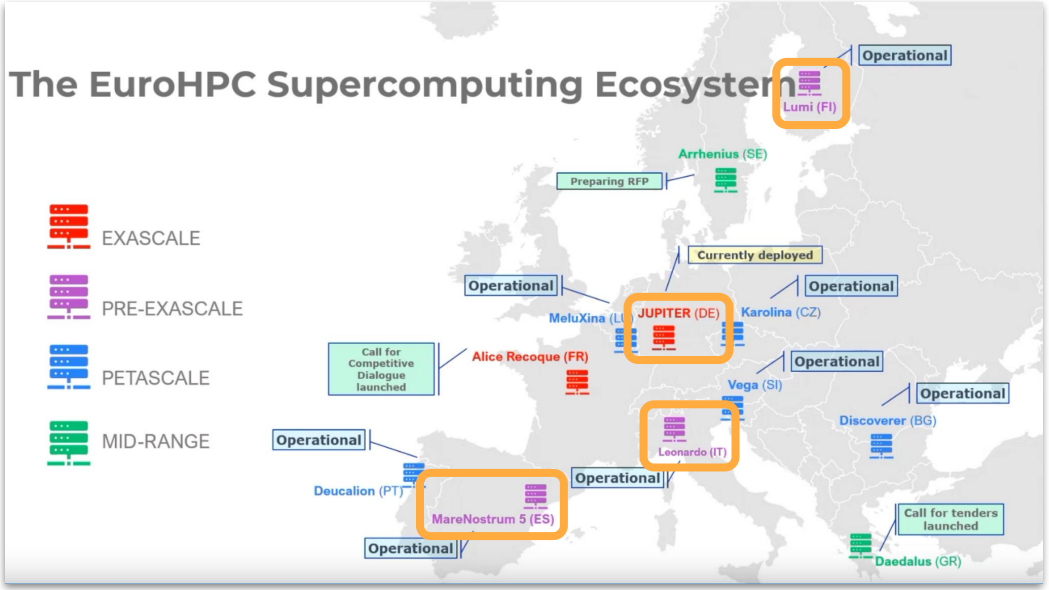
CAN OPENEUROLLM SUCCEED?



Compute

Project partners have been granted **tens of millions of GPUh** in previous projects

Approx. 30M GPUh have been applied for from national, EuroHPC extreme scale, and AI Factories calls

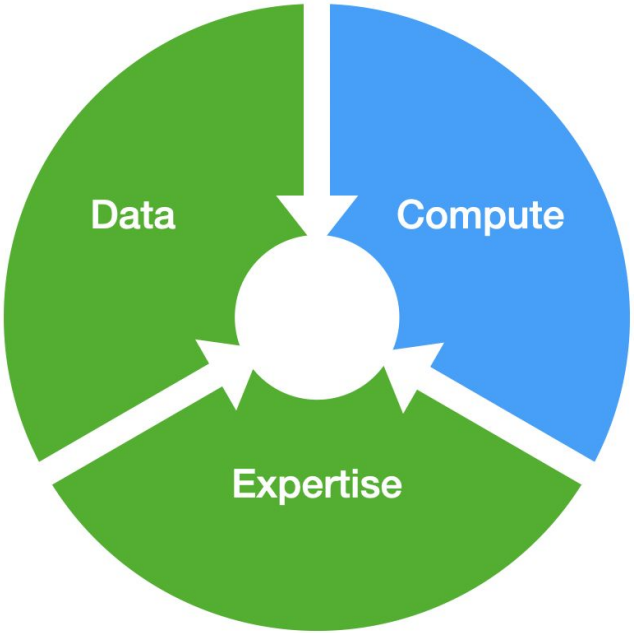


OPEN EURO LLM

CAN OPENEUROLLM SUCCEED?

We won’t suddenly jump from far behind to overtaking OpenAI

… but **if the project is given enough compute**, we have the rest of the parts in place to deliver the next generation (or two) of open LLMs for all European languages



OPEN
EURO
LLM

THANKS & ACKNOWLEDGEMENTS

The contents of this presentation are the sole responsibility of the author and do not necessarily reflect the opinion of the European Union.

This project has received funding from the Digital Europe Programme under grant agreement No 101195233

Co-funded by the European Union