



OpenEuroLLM

e-IRG Poznań, Poland (EU Presidency)

Jan Hajič

hajic@ufal.mff.cuni.cz

OPEN
EURO
LLM

June 24, 2025



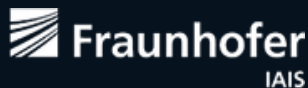
OpenEuroLLM

Our goal:

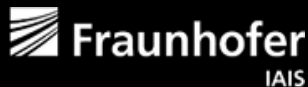
**Open
Multilingual
European
Generative
Foundational
LLM**

- Open Source (in full)
including fully inspectable data
 - 32+ languages
EU + associated (+ business)
 - High-quality
standard and native benchmarks
 - Compliant with EU regulations
-

OpenEuroLLM PROJECT PARTNERS



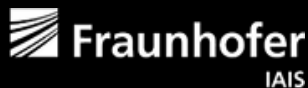
OpenEuroLLM PROJECT PARTNERS



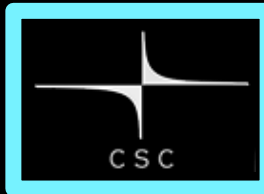
OpenEuroLLM PROJECT PARTNERS



OpenEuroLLM PROJECT PARTNERS



OPEN
EURO
LLM





Wider context

- Programme: Digital Europe (25/50% co-funding)
- Set of AI-06 calls (projects started Jan-Mar 2025):
 - Two large projects: **OpenEuroLLM** and LLMs4EU
 - Coordination (**ALT-EDIC4EU**), total **~80 mil. EUR + HPC**
 - Part of an ecosystem (Deploy AI, TAILOR, TrustLLM, HPLT, ...)
- Together we will
 - Develop **open**, high quality **foundation models**
 - **Adapt** them to applications in all areas, from commerce to government and education
 - Contribute to EU's **digital sovereignty**



Open Source and Community

- Open Strategic Partnership Board (Strategic advisory role)
 - **Open source community** members
 - Experts on LLMs (incl. from non-EU ones)
 - Former commercial and/or open source model developers
- Experts on **legal issues**
- Informal cooperations
 - Data side: CommonCrawl, Internet Archive EU, OpenWebSearch (TBC)
 - Open source models community
 - EuroLLM (Univ. of Edinburgh, UnBabel)
 - LAION, open-sci, ...



Computing facilities

- 5 EuroHPC centers on board (project partners)
 - Technical expertise
 - Jumps start using the respective facilities
- Some compute available from previous projects
- Participation in EuroHPC calls in 2025
 - In line with project plan for the rest of 2025
- Strategic allocations in the future
 - “STEP” seal awarded
 - Using current facilities & new in AI Factories (2026/2027)
 - Received 3m GPU hours for May-Nov. 2025 on Leonardo (CINECA, Italy)
 - Plus 1.5m GPU hours on LUMI (CSC, Finland) for initial LLM training



Data for 37+ languages

- Using available Open Source data
 - **HPLT** 2.0 (HPLT 3.0, July 25), Fineweb2, Cultura-X, ...
 - Mixtures to be experimentally determined
 - Ultimate (re)sources: **CommonCrawl**, Internet Archive, IA Europe
 - OpenWebSearch – negotiations ongoing
- Focus on **low-resource languages** for additional data
 - Incl. specific cases for very similar languages
- Additional data for
 - Fine-tuning, instruction-tuning, reasoning
 - ... if necessary for benchmarking



Evaluation and Benchmarking

- For initial experiments:
 - Standard benchmarks for base models
- Project longer-term goal
 - Benchmarks for **all languages in native form**
 - i.e., manually translated or inspected, incl. contents
- Continuous evaluation
- Tests for evaluation data purity
 - i.e., not used in training/SFT/...
- Models released based on evaluation results

Thank you!



<https://openeurollm.eu/>

- Questions?



<https://www.linkedin.com/company/open-euro-llm/>

**OPEN
EURO
LLM**

Supported by the project OpenEuroLLM, GA No. 101195233, a ALT-EDIC4EU, GA No. 101195344, Digital Europe Programme by European Commission and co-funded by the JU subprogramme of the MEYS CR and other partners's national agencies.



**Co-funded by
the European Union**

