OpenEuroLLM

UFAL seminar, March 2025

Jan Hajič

hajic@ufal.mff.cuni.cz

OPEN
EURO
LLM

March 24th, 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

OPEN EURO LLM

PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic ELDA

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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

PROJECT PARTNERS

Charles University

AMD SILO AI

alt-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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

PROJECT PARTNERS

Charles University

AMD SILO AI

alt-edic ELDA

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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

PROJECT PARTNERS

Charles University

AMD
SILO AI

alt-edic
el Dà

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

OPEN
EURO
LLM

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

CINECA

CSC

SURF

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 egovernment and education
  - Contribute to EU's digital sovereignty

OPEN EURO LLM

Open Source and Community

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

OPEN EURO LLM

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)

OPEN EURO LLM

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

OPEN EURO LLM

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

OPEN EURO LLM

Thank you!

- Questions?

OPEN
EURO
LLM