OpenEuroLLM

Research Data Day & EOSC NTE Brno

Jan Hajič

hajic@ufal.mff.cuni.cz

OPEN
EURO
LLM

May 21st, 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

What is a Large Language Model

- Known to the public primarily as conversational LLM (e.g. ChatGPT, Llama-3.3-70B-Instruct)
- Technology
  - Deep Neural Networks
  - Trained from data (texts) – Machine Learning
  - Basic function: generate next word (segment, token) based on (long) sequence of previous words (tokens)
    - In interactive systems: start with a user „prompt“
    - Can be up to a million words (in some LLM systems)

OPEN
EURO
LLM

What is a (base, foundation(al)) LLM?

- Model trained on running text only
  - i.e., not interactive
  - Cannot answer questions
  - Can be monolingual, multilingual, include (source) code
  - Can be multimodal (w/suitably encoded images, video, etc.)
- It is a basis for applications
  - Interactive (chatbot, conversational) LLM) is created by
    - fine-tuning, continous pre-training
    - human interaction – anotated data, relevance rating, etc.

How large is a Large Language Model?

- Model size is specified as
    - Number of parameters (wieghts), in millions (M) or [U.S.] billions (B)
        - Weight is a „real“ number in certain precision (from 32 down to 1.58 bit)
            - From that, byte size can be calculated: (1B weigths à 8 bit = 1 GB)
- Known model sizes (open-weigth models)
    - Llama 3.1: 405 B parameters (META / Facebook)
    - Llama 3.3: 70 B parametrů
        - Quantized (smaller precision than original) e.g. to 6 bits: 53 GB size)
        - For inference („runtime“): 1 or more GPU cards
            - Context size matters: takes a large proportion of GPU card’s memory
- Model training:
    - Number of parameters fixed (in the standard setting)
    - Different data (text) sizes (in words/tokens: tokenization very important)
        - Llama 3.1: 15 T (trillion) tokens

OPEN EURO LLM

OpenEuroLLM PROJECT PARTNERS

Charles University

AMD
SILO AI

alt-edic
eL
Da

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

OpenEuroLLM PROJECT PARTNERS

M

Charles University

AMD
SILO AI

alt-edic
eL
Da

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

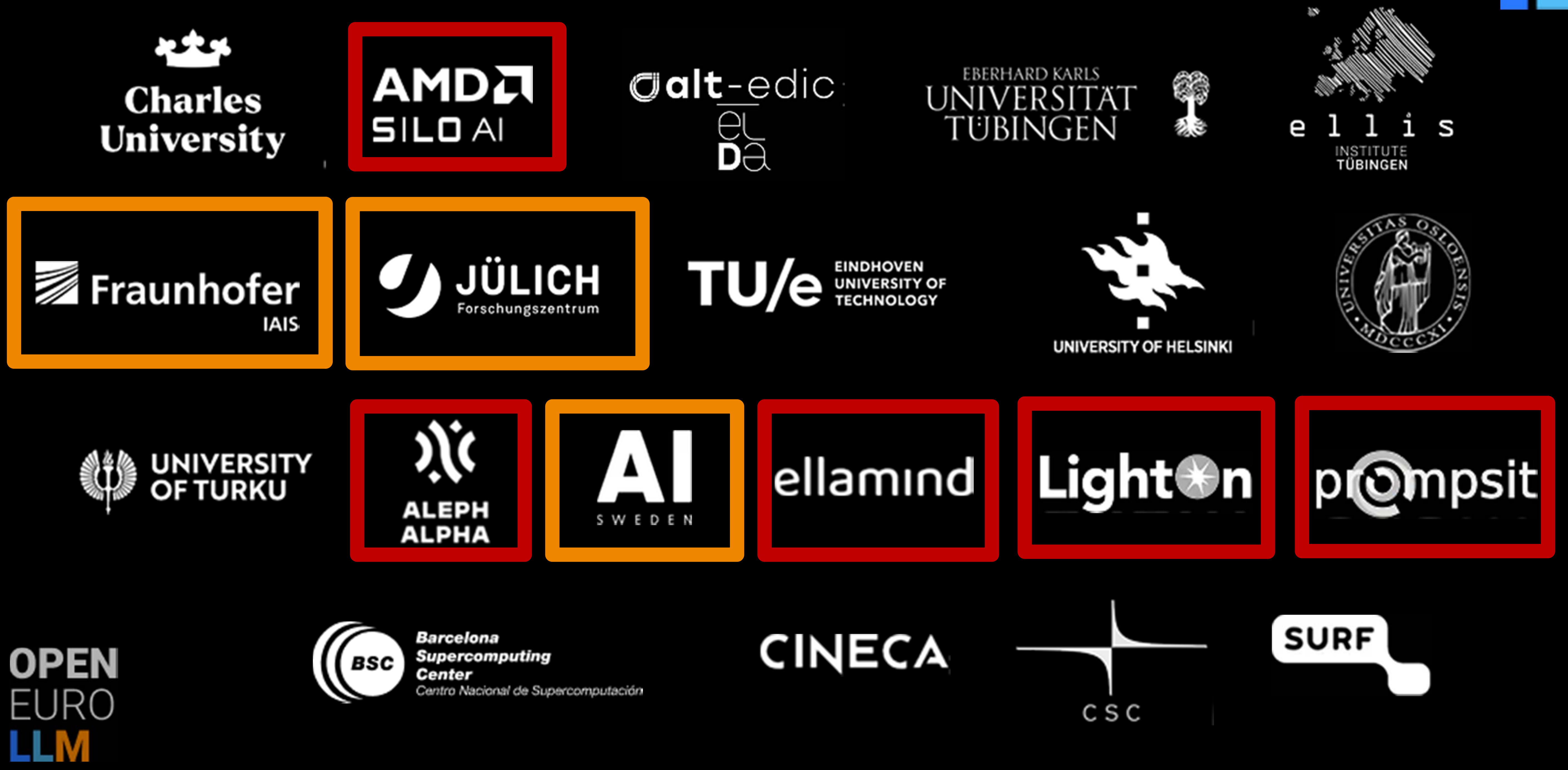
BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

CINECA

CSC

SURF

OpenEuroLLM PROJECT PARTNERS



OpenEuroLLM 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

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CINECA

CSC

SURF

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 egovernment and education
  - Contribute to EU’s digital sovereignty

OPEN
EURO
LLM

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, …

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)
        - Just received 3m GPU hours for May-Nov. 2025 on Leonardo (CINECA)

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

OPEN
EURO
LLM

Thank you!

- Questions?

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.

Co-funded by the European Union