LLMs in Europe:
Data Preparation and Training in the HPLT & OpenEuroLLM projects

Jan Hajič

hajic@ufal.mff.cuni.cz

Letní škola Czech AI Factory
Ostrava, IT4Innovations

Sep. 10, 2026

Outline

- Generative Large Language Models
- The OpenEuroLLM and HPLT projects
  - Model Training
  - Data Preparation
- Legal Aspects
- European Digital Sovereignty
- Conclusions

Generative LLMs

What is a (Generative) Large Language Model

- Known to the public primarily as conversational LLM (e.g. ChatGPT, Llama-3.3-70B-Instruct & many more)
- 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 systems)

OPEN EURO LLM

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

- Model trained on running text only
  - i.e., not interactive
    (i.e., cannot answer questions [well])
  - Can be monolingual, multilingual, include (source) code, math
  - Can be multimodal (w/suitably encoded images, video, etc.)
- It is a basis for applications
  - Interactive (chatbot, conversational) & adapted LLMs created
    - By fine-tuning, continuous pre-training
    - By human interaction – annotated data, relevance rating, etc. (RLHF, RL in general, PPO, DPO, …)

OPEN EURO LLM

How large is a Large Language Model?

- Model size is specified as
  - **Number of parameters** (weights), 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 weights à 8 bit = 1 GB)
- Known model sizes (open-weight models)
  - Llama 3.1 (META): 405 B parameters (dense architecture)
  - Qwen 3.8 Max (Alibaba): 2.4 T / Active: 95B parameters
    - Sparse model, Mixture of Experts architecture
  - Llama 3.3 (META): 70 B parameters
    - **Quantized** (lower 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

OPEN EURO LLM

Model training: how much compute do we need?

- Model training:
  - Number of parameters fixed (in the standard setting)
  - Different **data (text) sizes** (in tokens: **tokenization** very important)
    - Llama 3.1: 15 T (trillion, U.S.) tokens
- Compute needed depends on architecture (dense vs. sparse (MoE))
  - Standard formula for determining compute need (in TFLOPs):
    - 6 x N x D (N – [active] parameters, D – data size): https://arxiv.org/abs/2001.08361
  - Computing hours (and calendar days in near ideal conditions w/1024 cards) needed:
    - Depends on hardware and model type: here, Nvidia H100:

Model (D/S)	(A) Params	Data	FLOPs	GPU speed	Secs (1.2x)	mHours	Days (ca.)
32B/15T (D)	3.2E10	1.5E13	2.88E24	3.2E14	1.08E10	3.0	85
120B-A10B/15T (S)	1.0E10	1.5E13	9.00E23	2.2E14	3.38E09	0.9	27
2.4T-A95/36T (S)	9.5E10	3.6E13	2.05E25	2.2E14	1.12E11	31.1	881

OPEN EURO LLM

OpenEuroLLM and other LLM-related projects in EU

OpenEuroLLM

The goal:

Open Multilingual European Generative Foundational LLM

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

OPEN EURO LLM

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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

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

OPEN EURO LLM

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CINECA

CSC

SURF

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 Team Leaders

Jan Hajic
CUNI (WP1)

Jenia Jitsev
FZJ (WP4)

Sampo Pyysalo
UTU (WP4)

Joaquin Vanschoren
TU/e (WP5)

Stephan Oepen
UIO (WP3)

Edouard Geoffrois
ALT-EDIC (WP2)

Gema Ramirez
Prompsit (WP3/6)

Jussi Karlgren
AMD Silo (WP6)

David Salinas
ELLIS (WP5)

Matthias Bethge
UT (WP2)

Wider context

- Programme: Digital Europe (25/50% co-funding), 3 years 2025-28
- Set of AI-06 calls (projects started Jan-Mar 2025):
  - Two large projects: OpenEuroLLM and LLMs4EU
  - Coordination (ALT-EDIC4EU), total **~80 mil. EUR + HPC**
  - Strong cooperation (Deploy AI, TAILOR, TrustLLM, HPLT, …)
- Goals
  - Develop open LLMs, including conversational
  - Adapt them to applications in all areas, from commerce to e-government 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
  - Open source models community
    - EuroLLM (Univ. of Edinburgh - UK, UnBabel - Portugal)
    - Apertus (Switzerland – ETH Zürich, EPFL (Lausanne))
    - LAION, open-sci, …

OPEN EURO LLM

Computing facilities

- 5 EuroHPC centers on board (project partners)
  - Technical expertise – jump-starts using the respective facilities
- Participation in EuroHPC calls in 2025 and 2026
  - In line with project plan for the rest of 2025
- „Strategic“ allocation since Dec. 2025 (1.25% capacity of 4 HPCs)
  - Approx. 10 mil. GPUh over 2026 and 2027 (but less in practice, due to outages etc.)
  - Using current facilities & new in AI Factories (2027-)
  - Also received 3m GPU hours for May-Nov. 2025 on Leonardo (CINECA), 3M on Jupiter BOOSTER from 2026
  - … and 1.5m GPU hours at CSC, on LUMI-G
    - Testing data staging, multilingual training, MoE (~scaling laws)

OPEN EURO LLM

Data for 37+ languages

- Which…? - EU, EEA, candidate members (Albania, Ukraine, …)
- Using available data
  - **HPLT** 2.0 (HPLT 3.0/4.0 pool), Fineweb Edu, DCLM, Nemotron CC…
  - 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, synthetic data generated
- Additional data for
  - Fine-tuning, instruction-tuning, reasoning and benchmarking

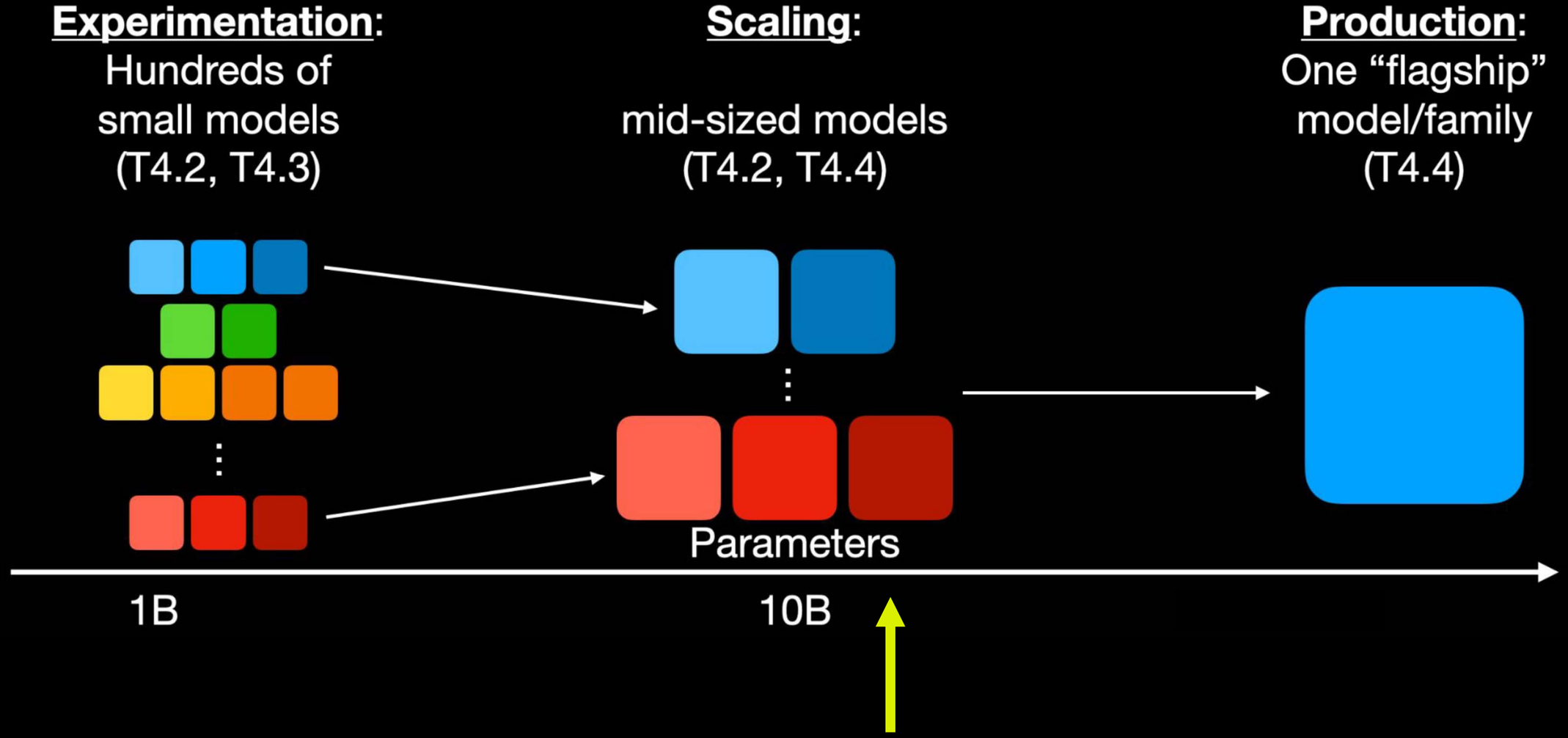
OPEN EURO LLM

Evaluation and Benchmarking

- For initial experiments:
  - Standard benchmarks for base models
  - Small reference models already published (HF)
- 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/…

OPEN EURO LLM

Training plan



OPEN EURO LLM

Training models in OpenEuroLLM

- Training in progress
  - 9B/10T „baby“ dense model, Llama-setup, Megatron-LM
    - Nemotron-CC, DCLM, FinePDFs(-edu), HPLT 3.0, MultiSynt, code, math
    - Cineca HPC / Leonardo GPU cluster, started June 2026, to finish soon
- Next
  - 1st Flagship 32B/15TT dense model, starting now (Qwen-branch)
    - Simialr data (larger), also HPLT 4.0, own quality filtering, parallel data
    - Jupiter Booster at JSC (largest available public HPC in Europe)
    - Expected to use 3.0 mil. GPU hours (GH200/H100)
    - To finish in Dec. 2026

OPEN EURO LLM

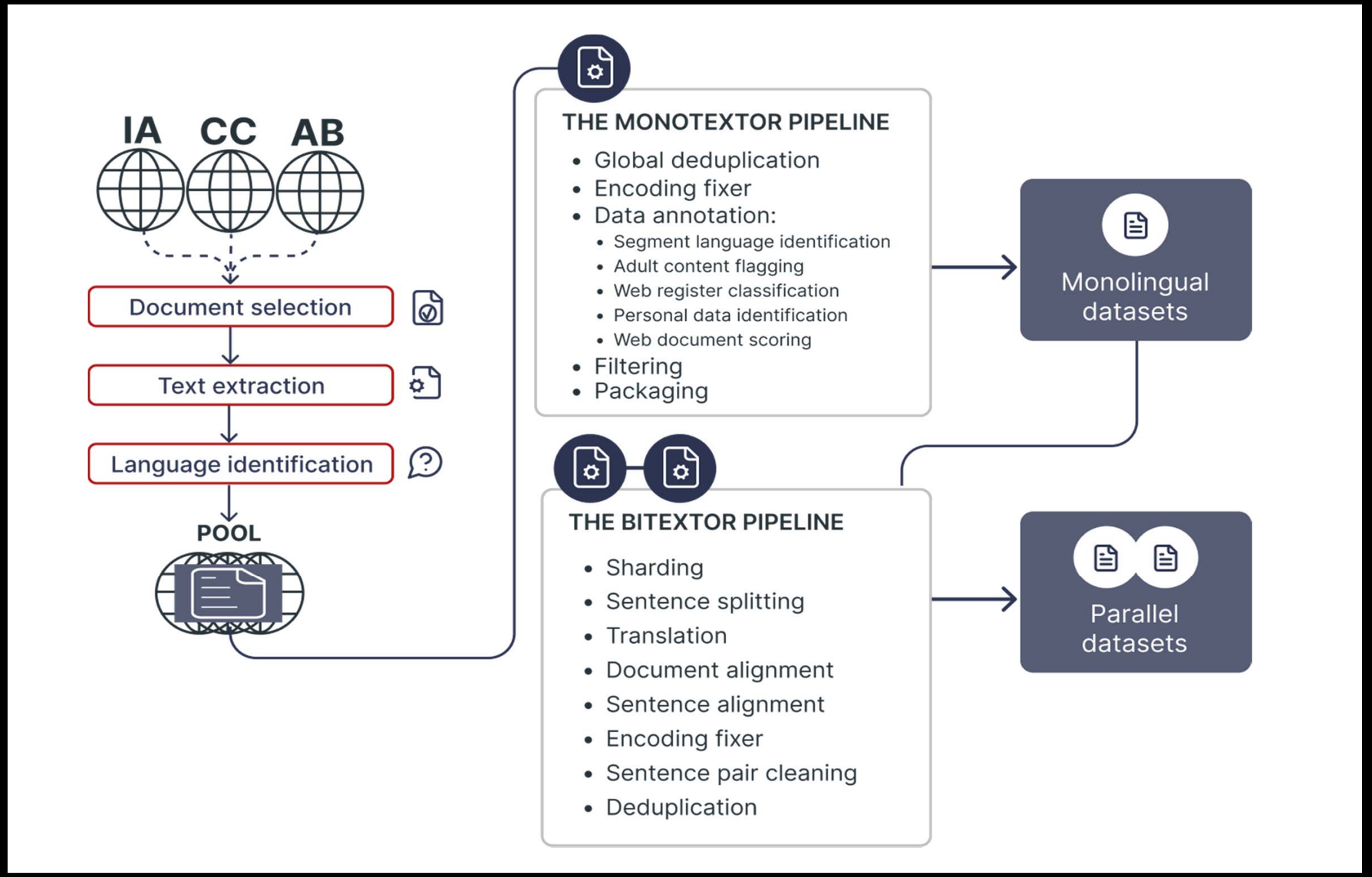
Pre-training Data Preparation

Data preprocessing

- For LLMs, data comes mostly from the internet (not much choice…)
  - Closed sources unavailable (Google, Microsoft, Seznam, …)
  - Public crawlers
    - CommonCrawl, Internet Archive (and various smaller crawlers)
      - OpenWebSearch – European project
- Data from internet
  - HTML (and all its extensions, like css, JavaScript, other scripting langauges, etc.)
    - Plus additional referenced files: images, video, audio, pdfs, spreadsheets, etc.
  - Preprocessing needed (next slides) – HPLT and OpenEuroLLM projects
  - Output: plain text (+metadata)

OPEN EURO LLM

Data preprocessing (in the HPLT project)



Data sources with original sizes

- All web data, different approaches to crawling by source

Release (HPLT)	Internet Archive	Common Crawl	ArchiveBot	Total
	PB	PB	PB	PB
2.0	3.4	0.7	0.3	4.5
3.0	3.5	3.8	-	7.2
4.0	3.4	4.6	2.3	10.2

https://hplt-project.org/datasets/vX.0

Yield: from 1 mil. characters (ArchiveBot) to 33 mil. characters (cc_n) per compressed GB

- e.g., for HPLT 3.0: 30T clean tokens; HPLT 4.0 (pool): close to 40T (yield: 4.3%)

OPEN EURO LLM

Data (monolingual) pipeline

- From HTML to annotated texts for LLM pre-training
- Rich metadata

Text extraction (from WARC files)

- Near-deduplication
- MinHash + UnionFind for id
- Keeps 1 record for near-duplicates
- Enhanced performance

- Robots.txt
- Web registers
- Adult content flags
- PII Identification
- WDS Scores (quality estimation)

- Character fixing (Monofixer)
- Mojibake
- HTML entities
- Wring alphabet

- LangID (OpenLID 2.0/3.0)
- segment-level LID

- Metadata formatting, packaging into shards, and publishing
- Parallel (en-xx) data extracted and available as well

Metadata, distribution

- Crawling, language, cleaning, scoring and enrichment information
  - Includes register labels, flags, PII annotation
- Overall: JSONL format
  - XML for rendering info
- Distribution
  - Zstandard-compressed
  - HPLT web (Sigma2), HF

OPEN EURO LLM


```
{"f": "./segments/1652663048462.97/warc/CC-MAIN-20220529072915-....warc.gz",
 "o": 424865140, "s": 9226, "rs": 51579,
 "u": "https://lynghaug.no/news/oppdatering-angaende-rehabilitering?tm=",
 "c": "text/html", "ts": "2022-05-29T08:06:01Z", "de": "utf-8",
 "crawl_id": "CC-MAIN-2022-21",
 "lang": ["nob_Latn", "nno_Latn", "dan_Latn"], "prob": [0.9963, 0.0026, 0.0011],
 "text": "UKE 2 2022
    Styret har registrert en del misnøye i blant beboere  [...]
    Vi har regelmessig kontakt/møter med Markhus og har drøftet  [...]
    ...",
 "xml":"<doc fingerprint="e589be41f0c5445f">
  <main>
    <p>
      <hi rend="#b">UKE 2 2022</hi>
    </p>
    <p><hi rend="#b"/>Styret har registrert en del misnøye i blant beboere  ...</p>
    <p><lb/>Vi har regelmessig kontakt/møter med Markhus og har drøftet  ... </p>
    ...
  </main>
  <comments/>
  </doc>",
 "cluster_size": 8,
 "seg_langs": ["ssw_Latn", "nob_Latn", "nob_Latn", "nob_Latn", "nob_Latn", "nob_Latn"],
 "id": "bf4cf56d1c47d62db151874c8fa9d53f",
 "filter": "keep", "pii": [[1428,1457]],
 "doc_scores": [8.9, 10, 10, 10, 10, 10, 10, 4, 5.4, 10],
 "web-register": {"MT":0.028, "LY": 0.059, "SP": 0.073, "ID": 0.158, "NA": 0.655,
                  "HI": 0.17, "IN": 0.286, "OP": 0.194, "IP": 0.299,
                  "it": 0.069, "ne": 0.164, "sr": 0.059, "nb": 0.613, "re": 0.066,
                  "en": 0.036, "ra": 0.048, "dtp": 0.114,
                  "fi": 0.098, "lt": 0.077, "rv": 0.055, "ob": 0.097, "rs": 0.098,
                  "av": 0.142, "ds": 0.107, "ed": 0.072}}
```

Legal aspects of LLMs

Legal questions (introduction)

- Input data
  - Copyright
  - Licensing of existing datasets
- Models
  - AI Act (refers to copyright, too)
  - GDPR
- Systems
  - Inference (or „running“ the models)
    - Might include all other aspects (e.g., search/RAG, using databases etc.)
- Research vs. commercial use, software licenses

OPEN EURO LLM

Legal questions (copyright)

- Copyright
  - TDM in Art. 3 and 4 of the CD 790/2019, as implemented in MSs
    - Czechia: “Autorský zákon“ 121/2000 Sb., as amended
    - Part 1, Hl. 1, par. 39c and 39d (also par. 31.1(c) – research use)
  - Art. 3 – use of copyrighte materiál for research
  - Art. 4 – use also for commercial purposes
    - More strict than Art. 3, most notably: must respect “opt-out” mechanism
      - Machine readable – currently only robots.txt (“deny” claues)
  - In practice, certain crawlers now denied access even if lawful
- Would need in fact more liberal approach
  - at least for model bulding per se
    - Even currently not on par with U.S. (“Fair use”) and some other regions

OPEN EURO LLM

Legal questions (models and systems)

- Models
  - Specific provisions for GPAI models in AI Act
    - Art. 53 – transparency
      - Not necessarily every source, but documentation
        - Less obligations for OpenSource
      - Stricter rules for high-end models (>10^25 FLOPs for training)
        - Called “models with systemic risk” (Article 51(1) AI Act)
          - Regardless if OpenSource(!)
    - Reference to Copyright
- Systems
  - I.e., models deployed in AI systems (“inference”)
  - Difference: high-risk vs. other systems

OPEN EURO LLM

Where are we (in Europe)?

LLMs (AI): what counts as “EU(+) sovereignty”

- 1 Data
  - Open Crawling procedure(s), other sources
  - Preparation pipelines (cleaning, filtering, annotation)
  - Evaluation data, posttraining data, multilinguality
- 2 Software
  - Data crawling, acquisition, preparation
- 3 Models – reproducible, on par with commercial ones
  - used as basis for CPT, SFT, instruction tuning, reasoning, etc.
- 4 Compute, Data and Networking Infrastructure in the EU
  - Compute – training, inference
  - Repositories - for all of the above: data, software, models
- 5 Legal Environment
  - Data, software and model use

OPEN EURO LLM

Reality check 1 (grading: 1: Excellent, 5: Insufficient)

Data (2-)

- Pre-training (1-)
  - Results of previous projects, e.g. HPLT, or Open Data available, multilingual (text only), depend on non-EU data (CommonCrawl)
  - Some high-quality data not originating in Europe (Fineweb, MADLAD)
  - Fully open data (3-) – attempts running, but so far, not large enough
- Instruction tuning, reasoning, benchmarking (3-)
  - Close to non-existent, or not physically in Europe
  - Benchmarks (multilingual) being developed, national exist (uncoordinated)
- Domain specific, multimodal (3)
  - Mostly not in Europe

Reality check 2 (grading: 1: Excellent, 5: Insufficient)

- Software (1-)
  - Mostly available and/or (some) created in Europe
    - Crawling (OpenWebSearch – but not linked to the data pipelines yet)
    - Data preparation (text)
      - HPLT, OpenEuroLLM, Apertus, EuroLLM, OpusMT
      - But: used non-EU sources -> legal uncertainty
    - Model training software
      - Not developed in Europe but available
    - Training recipes and scripts, HPC setups, …
      - In the works, available from the world (open source)

OPEN
EURO
LLM

Reality check 3 (grading: 1: Excellent, 5: Insufficient)

- Models (3-)
  - Closed: Mistral
  - Open weights (min.)
    - PORO, Viking, EuroLLM, Apertus, some national
      - Good for some tasks (e.g. EuroLLM for MT)
  - Not on par with closed models
    - closer to Open models (OLMO – Allen AI Institute, U.S.)
  - Distribution usually through HuggingFace (non-European org.)

OPEN
EURO
LLM

Reality check 4 (grading: 1: Excellent, 5: Insufficient)

- Compute and networking infrastructure (4-)
  - Main source of (public, EU) compute: EuroHPC, small national nodes
    - Absolutely insufficient in current state
    - Steps now being taken: AI Factories, AI Gigafactories, Data
      - But most will be operational late 2027 and later
      - Used in training/SFT/…
- Networking infrastructure (2)
  - Relatively fast
- Data capacity (2)
  - Typically tens of PB at each Data Center of HPC site; for LLMs alone, OK

Reality check 5 (grading: 1: Excellent, 5: Insufficient)

- Legal situation (3-, but possibly deteriorating)
  - Copyright
    - TDM in Art. 3 and 4 of the CD 790/2019, as implemented in MSs
    - Currently under attack (EP)
    - Would need the opposite – more liberal approach!
      - Even currently not on par with U.S. and some other regions
  - Software – mostly OK, especially for Open Models
  - Models
    - Mostly OK, confusion if models are copyrightable (vs. licensable)
      - Software-like licenses used (works in practice)

OPEN EURO LLM

Thank you!

Questions?

https://openeurollm.eu/

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

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

Co-funded by the European Union