Overview of Language Models in HPLT – High Performance Language Technologies

Gema Ramírez-Sánchez,
on behalf of the HPLT project consortium
https://hplt-project.eu
gramirez@prompsit.com
Valencia, 26th Oct. 2023

EUROPEAN BIG DATA VALUE FORUM
25-27 OCTOBER | VALENCIA - SPAIN

The HPLT project in a nutshell

A space combining petabytes of natural language data with large-scale model training:

- Consistently formatted and curated data sets
- Efficient and high-quality LLMS and MT models
- Sustainable and reusable workflows using HPC

7 petabytes of web data from the internet archive

5 petabytes of web data from commoncrawl

2.5 trillion words of monolingual text

600 unique corpora

~80 languages to cover

100s of efficient language and translation models

36 months to complete the project

8 consortium partners collaborating together

Funded by

UK Research and Innovation

Dates:
Sept. 2022-Aug. 2025

Visit our website:
https://hplt-project.org

Download HPLT datasets:
https://hplt-project.org/datasets/

Follow us:

EUROPEAN BIG DATA VALUE FORUM
25-27 OCTOBER | VALENCIA - SPAIN

The HPLT project partners

HPLT gathers together academic partners, an NLP company and HPC centers from all around Europe.

CHARLES UNIVERSITY

UNIVERSITY OF OSLO

UNIVERSITY OF EDINBURGH

UNIVERSITY OF TURKU

UNIVERSITY OF HELSINKI

PROMPSIT

CESNET

SIGMA2

Funded by

UK Research and Innovation

Dates:
Sept. 2022-Aug. 2025

Visit our website:
https://hplt-project.org

Download HPLT datasets:
https://hplt-project.org/datasets/

Follow us:

EUROPEAN BIG DATA VALUE FORUM
25-27 OCTOBER | VALENCIA - SPAIN

HPLT LLMs: one size does not fit all

HPLT will train not only GPT-like models:

- **Encoder-only (BERT-like):** still very popular for classification tasks (NER, sentiment, etc.)
- **Encoder-decoder (T5-like):** still very popular for taks where the output heavily depends on the input (summarization, translation, etc.)
- **Decoder-only (GPT-like):** just very popular and even more finetuned on instruction datasets (content creation, Q&A, etc.)
- **A Flagship model: TBD**

HPLT wants to bring as many languages as possible to the state-of-the-art of language modelling and beyond.

HPLT LLMs: (data and computing) size matters!

Compute to create compute-optimal (cf. Chinchilla) models scales **quadratically** with available data.

Words	# Langs	Languages	Comments
> 1.5T	1	EN	Too big / competitive?
> 100B	5	DE, ES, FR, RU, ZH	Sweet spot
> 10B	21	AR, CS, DA, EL, FA, FI, HE, HU, ID, IT, JA, KO, NB, NL, PL, PT, RO, SV, TR, UK, VI, NO	Sweet spot
> 1B	18	AZ, BE, BG, BN, CA, ET, HI, HR, LA, LT, LV, MS, SK, SL, SQ, SR, TA, TH	Too small?
> 100M	31	AF, CY, EO, EU, GA, GL, GU, HY, IS, KA, KK, KN, KY, MK, ML, MN, MR, MT, MY, NE, NN, PA, PS, SI, SO, SW, TE, TL, TT, UR, UZ	Toy models only

HPLT LLMs: sourcing data at scale

HPLT wants to source as much data as possible for as many languages as possible to democratize resources for language modelling.

1st release completed:

- derived from large crawls (Internet Archive & Common Crawl)
- covers 75 languages
- 22TB of raw data = 7.6TB of dedup data in compressed JSONL
- metadata (useful to filter data for quality)
- collection licensed CC0

HPLT LLMs: models and pipelines

HPLT wants to produce efficient models and pipelines to democratize language modelling training.

Some steps accomplished:

- Exploring data-efficiency in low-resource settings:
  - training successful BERT-and-T5-like models on just 100M words
  - evaluating training data repetition effects in GPT-like models
- Exploring compute-optimality, monolingual vs multilingual LLMs, etc.
- Released: FinGPT, ~30B tokens from scratch and from Bloom
- Released: LTG-BERT efficient architecture → NorBERT and NorT5

HPLT LLMs: computed in an EuroHPC environment

HPLT wants to work hand-in-hand with European HPC centers to make them NLP aware.

Working in CESNET and mainly LUMI:

- requires adapting software to each particular center
- requires huge anticipation of computing needs and timing
- requires coping with instability and scaling issues

But LLMs and HPLT scale heavily depend on their huge computational capacity

(currently available **14MGPUh** for model building)

HPLT wants to hear from you

At HPLT we are looking forward to being useful for your purposes and we also ask for your contribution:

- Datasets
- Benchmarks
- Use cases

Keep in touch: https://hplt-project.org/

Thanks!

EUROPEAN BIG DATA VALUE FORUM
25-27 OCTOBER | VALENCIA - SPAIN