HPLT

HIGH PERFORMANCE LANGUAGE TECHNOLOGIES

A space that combines petabytes of natural language data with large-scale model training

- Lots of monolingual and multilingual data consistently formatted and curated
- Efficient and high-quality language and translation models
- Sustainable and reusable workflows using high-perfomance computing

More about HPLT

https://hplt-project.org/ @hplt_eu

Funded by: Horizon Europe, UK Research and Innovation

WELCOME TO BOARDS

We will help users find what is in language data and models, how they compare to others, and how they were built through interactive boards.

Our Focus

We will retool how language data is generated, shared, and transformed into efficient large language and translation models making HPC centres ready to large scale NLP across Europe.

7 petabytes of web data from the internet archive

5 petabytes of web data from commoncrawl

2.5 trillion words of monolingual text

~300 unique corpora

~80 languages to cover

100s of efficient language and translation models

36 months to complete the project

8 consortium partners collaborating together

Our Partners

CHARLES UNIVERSITY

UNIVERSITY OF OSLO

UNIVERSITY OF EDINBURGH

UNIVERSITY OF TURKU

UNIVERSITY OF HELSINKI

PROMPSIT L.E.

CESNET

SIGMA2