# OpenEuroLLM

*Large Language Models for Europe*

Jan Hajič and Sampo Pyysalo

OPEN
EURO
LLM

# OpenEuroLLM

**Our goal:**

**Open Multilingual European Generative Foundational LLM**

- Open Source (in full)
  including fully inspectable data
- 36+ languages (42 with dialects)
  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
eld

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 eld 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

Charles University

AMD SILO AI

alt-edic eld 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

# 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, …)
  - Contribution to EU+ 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
- Informal cooperations
  - Data side: CommonCrawl, Internet Archive EU, OpenWebSearch
  - Open source models community
    - EuroLLM (Univ. of Edinburgh - UK, UnBabel - Portugal)
    - LAION, open-sci…; Switzerland / Apertus

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 line with project plan for the rest of 2025
- “Strategic“ allocations since January 2026
  - Using current facilities & new in AI Factories (2026/2027)

OPEN
EURO
LLM

# Data for 36+ (42) languages

- Using available data
  - HPLT 3.0(+), Fineweb 2, FineWeb edu, …
  - Mixtures experimentally determined
    - Ultimate (re)sources: CommonCrawl, Internet Archive, IA Europe
    - OpenWebSearch – negotiations ongoing
- Focus on low-resource languages for additional data (also synth.)
  - Incl. specific cases for very similar languages
- Additional data
  - Fine-tuning, instruction-tuning, reasoning

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
- Tests for evaluation data purity
  - I.e., not used in training/SFT/…
- Models released based on evaluation results

OPEN
EURO
LLM

# OPENEUROLLM

## Completed and upcoming model and data releases

| | |
|---|---|
| **Completed:** | **Reference models for EU languages +** (link) |
| **Completed:** | **Multilingual reference models** (link) |
| **Completed:** | **English reference models** (link) |
| **Ongoing:** | **Multilingual synthetic data and models** (link) |
| **Spring 2026:** | **First production models** (~10B/10TT) |
| **Autumn 2026:** | **First flagship models** (70B+/10TT+) |

\+ additional releases as part of open process

OPEN
EURO
LLM

# HPLT v2 REFERENCE MODELS

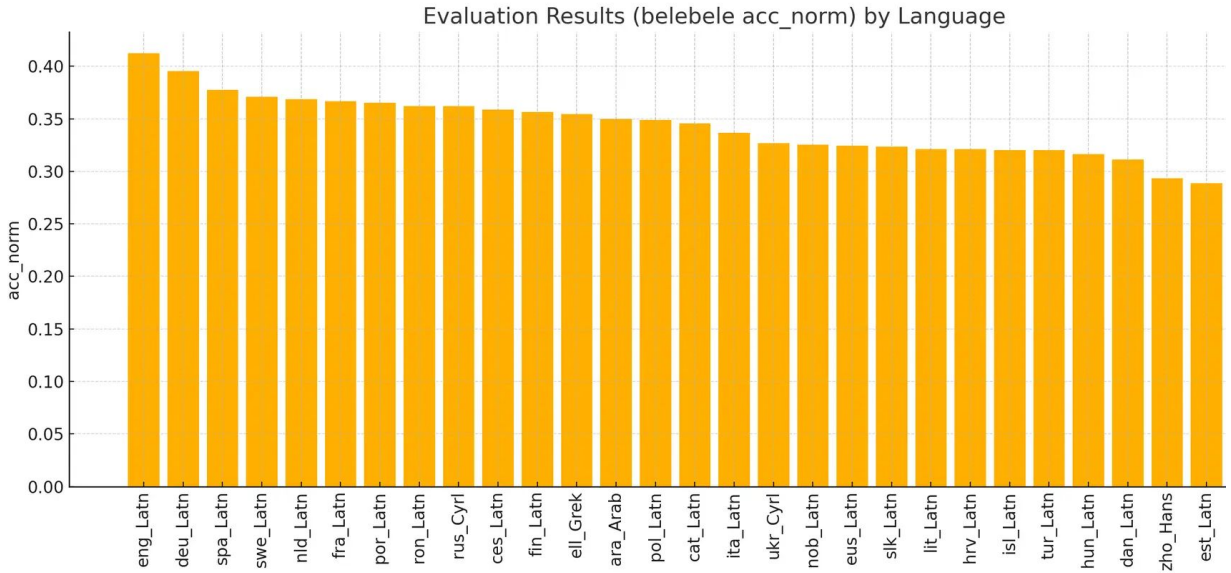
## Reference models for broad range of European languages (+more)

**Collab with HPLT**

https://hplt-project.org

~2B parameters, 100B tokens, 38 languages

huggingface.co/HPLT




OPEN EURO LLM

Release of 38 Monolingual 2.15B LLMs Trained on HPLT v2

# MULTILINGUAL REFERENCE MODELS

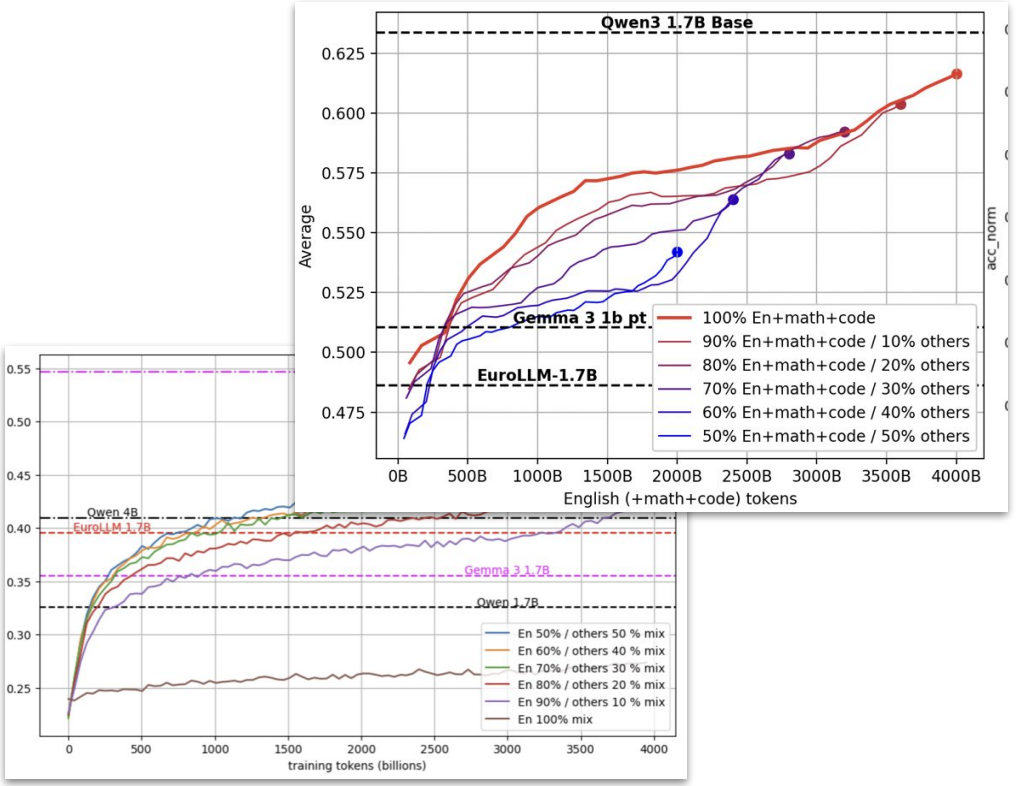
## Models for European languages

2B/4TT models, 38 languages

Six increasingly multilingual language mixtures

Competitive with comparable models on target languages

(9B/4TT mix models coming)




https://huggingface.co/openeurollm

OPEN EURO LLM

# OPEN-SCI REFERENCE MODELS

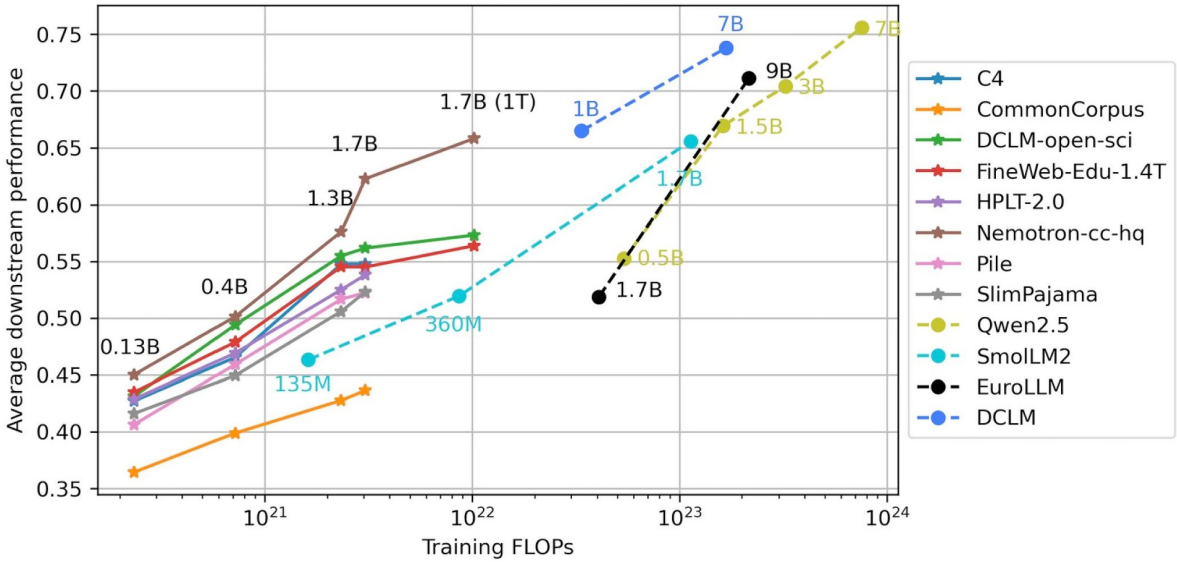
## Baselines for model and dataset comparison

**Collab with Open-Sci**

huggingface.co/open-sci

Up to 1.7B/1TT models, various English datasets

Competitive results, good scaling trends




OPEN EURO LLM

Open-sci and OpenEuroLLM release of reference models

# MULTISYNT DATA AND MODELS

## Open multilingual synthetic data for LLM pre-training

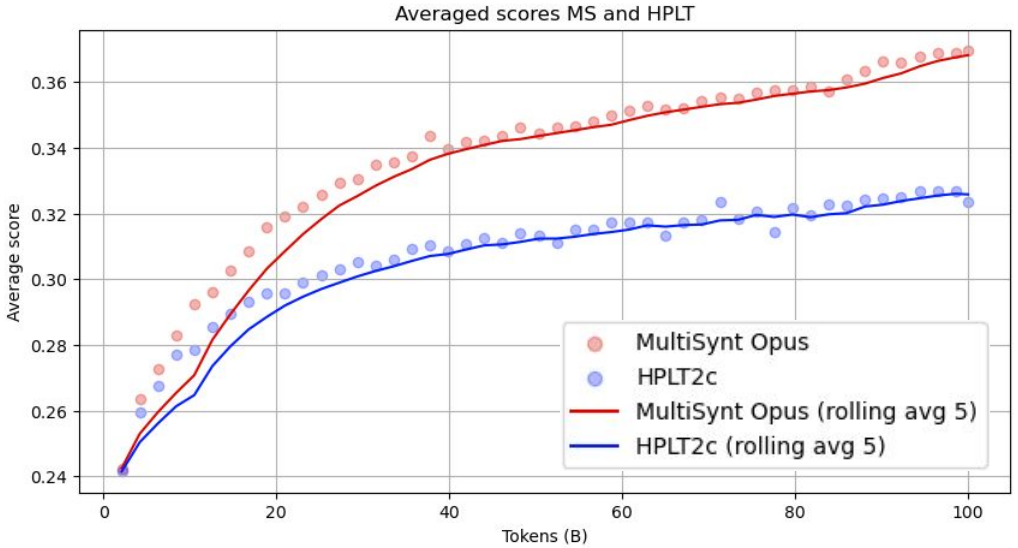
**Collaboration with EuroLLM**

https://eurollm.io/

3M GPUh AI Factories grant

Over 1T tokens of data and 20 2B/100BT models available:

https://huggingface.co/MultiSynt




Scores averaged for [eus, dan, nld, ita, por, swe] Multisynt Opus and HPLT2c models.
Tasks used [belebele, hellaswag, arc:challenge, mmlu]

OPEN EURO LLM

MultiSynt: an open multilingual synthetic dataset for LLM pre-training

# OPENEUROLLM

## Scope: will OpenEuroLLM …

Train also on programming languages? **Yes** ✅

Train models for instruction-following / dialogue (chat)? **Yes** ✅

Train “reasoning” / “thinking” models? **Yes** ✅

Explore architectures other than dense transformers (e.g. MoE)? **Yes** ✅

Fine-tune models for specific use cases (e.g. science)? **No** → LLMs4EU

Train multimodal models (e.g. audio and images)? **No** → ELLIOT

OPEN
EURO
LLM

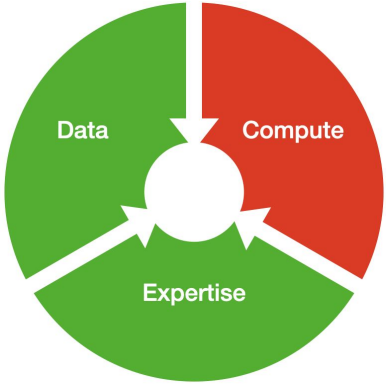
# CAN OPENEUROLLM SUCCEED?

**Goal**: leading fully open foundation models for EU languages (+more)

One of many efforts with similar goals

**Strengths**: expertise in LLM training through **partners** and **collaborations**, multilingual data curation (HPLT, collaborations)

**Challenge**: compute; applied for approx. 30M GPUh in various calls, got ~10M



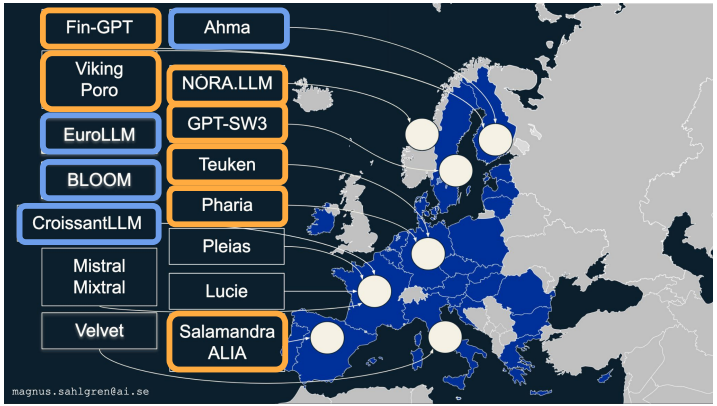




Figure credit: Magnus Sahlgren

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, 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

# MultiSynt: an open multilingual synthetic dataset for LLM pre-training

- A compute project funded by EuroHPC AI Factory
- Initially on Leonardo (6 months)
- Now on MareNostrum 5 (12 months)

Primary goal: Address multilingual data scarcity (non-engl. EU langs)


x

# MultiSynt: an open multilingual synthetic dataset for LLM pre-training

Initial focus: Machine Translation at Scale

- Source: 100BT sample from Nemotron-CC HQ (English)
- Target: 36 languages
- MT Models: Tower+72b, Tower+9b, OPUS-MT

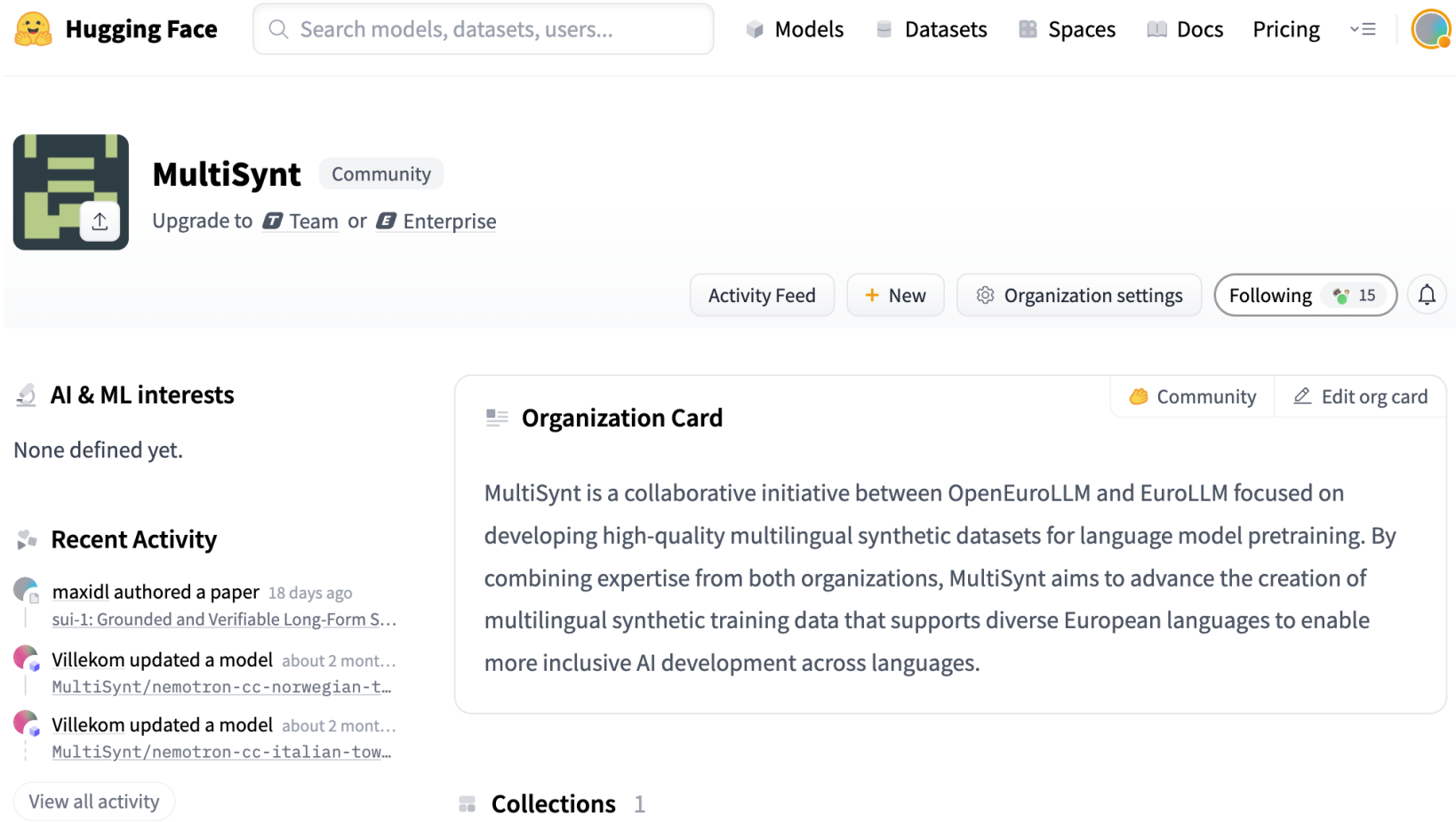
So far:

~5.8T tokens (hf.co/datasets/MultiSynt/MT-Nemotron-CC)

>30 ablation models trained

# MultiSynt: an open multilingual synthetic dataset for LLM pre-training

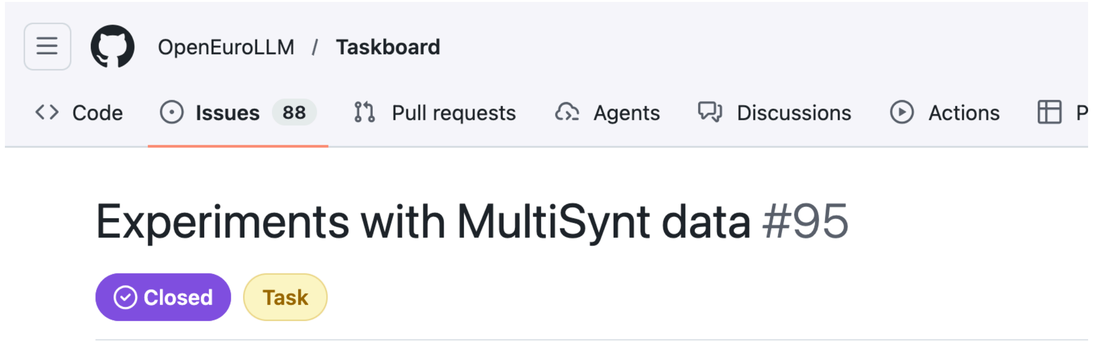
Find us on HuggingFace: hf.co/MultiSynt



# Multisynt Evaluation

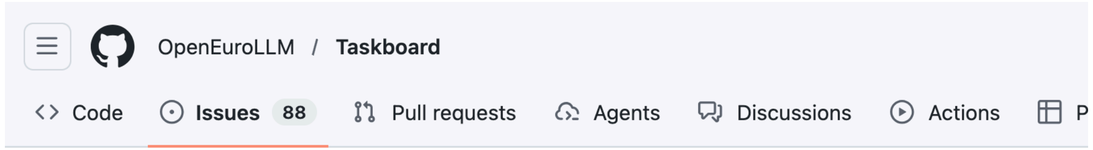
# Evaluations

# Multisynt evaluations




- https://github.com/OpenEuroLLM/Taskboard/issues/95

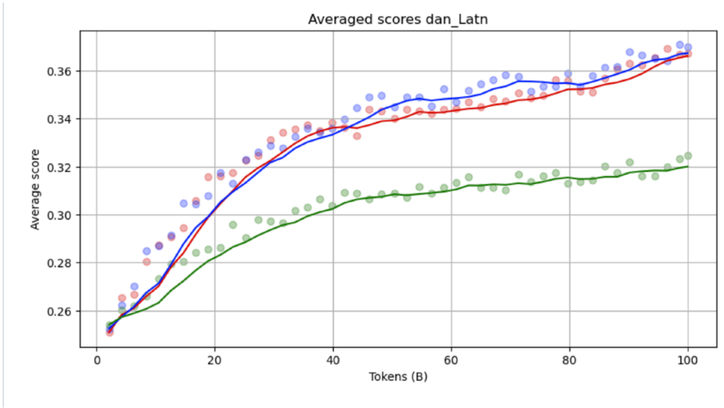
# Multisynt evaluations



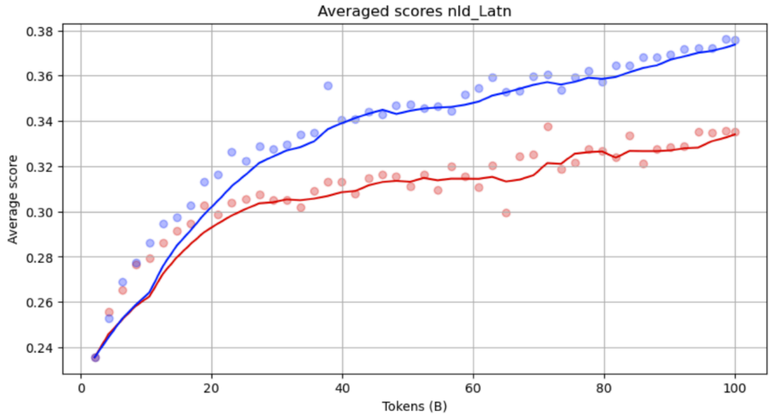

Experiments with MultiSynt data #95

Closed Task

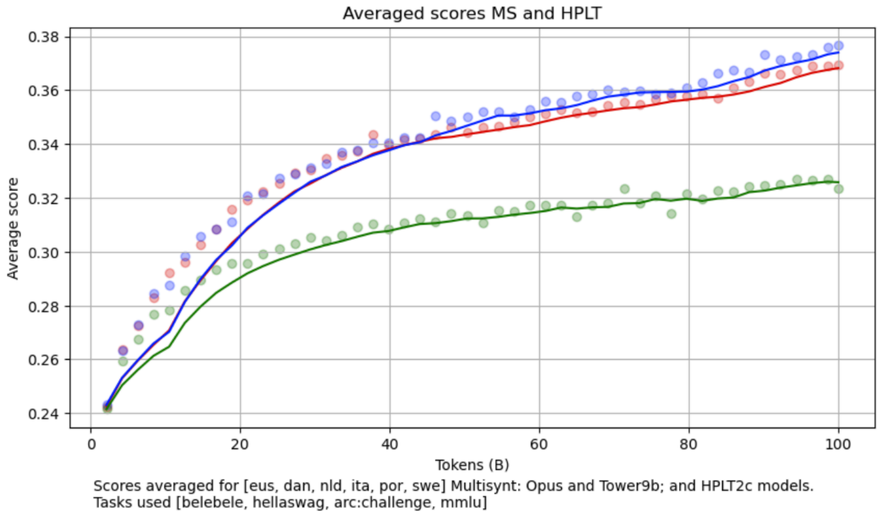
- https://github.com/OpenEuroLLM/Taskboard/issues/95




Danish




Deutsch




Average over eus, dan, nld, ita, por, swe

# Multisynt evaluations

- Overwhelming performance gain over native data for downstream evaluations
- Are the models good for downstream task
- … but have very bad fluency? (Eg translationese)

# Fluency

# Fluency

- The models seems to be much better for downstream performance
- But how much impacted are they by translationese?
- Do they fluency suffer? How much?

# Measuring fluency

## … for pretrained models

- Are the models trained on translated data worse for fluency?
- Perplexity only measure next token prediction quality
- How do we measure fluency for *pretrained* models?
- LLM judges to the rescue

# Measuring fluency

## LLM judges

```
system_prompt = """You are a highly efficient assistant,
who evaluates and selects the best large language model
based on the quality of completion of a sentence. You will
see a sentence to be completed and two completions from
Assistant A and Assistant B and will have to decide which
one is best. Make sure to not over-confidently prefer one
assistant or the other and also make sure to not bias your
preference based on the ordering or on the length of the
answers."""
```

- 1. Generate 100 completions from cut-out sentences
- 2. Compare completions with an LLM judge
- 3. Compare win rate with a baseline
- Studied languages: Finnish, French, German, Spanish and Swedish

Model 1: **Of-course,** she is knowing the answer.

Model 2: **Of-course,** she knows the answer.

→ Model 2

LLM judge

| | |
|---|---|
| general sentences | Die Sonne scheint warm über den Feldern und |
| general sentences | Das Kind lachte laut während es mit dem |
| history | Im Jahr 1914 begann ein Konflikt, der die |
| history | Die Entdeckung neuer Seewege veränderte den Handel und |
| economics | Das Angebot sank plötzlich, was den Preis auf |

| | |
|---|---|
| general sentences | Le chat noir traverse la rue vite |
| history | Napoléon a traversé le pays |
| economics | La politique monétaire influence les taux d'intérêt et l'investissement |
| math | La matrice invertible a un déterminant |
| programming | La boucle for itère sur chaque élément du tableau en |

# Measuring fluency

## LLM judges

```
system_prompt = """You are a highly efficient assistant,
who evaluates and selects the best large language model
based on the quality of completion of a sentence. You will
see a sentence to be completed and two completions from
Assistant A and Assistant B and will have to decide which
one is best. Make sure to not over-confidently prefer one
assistant or the other and also make sure to not bias your
preference based on the ordering or on the length of the
answers."""
```

- 1. Generate 100 completions from cut-out sentences
- 2. Compare completions with an LLM judge
- 3. Compare win rate with a baseline
- Studied languages: Finnish, French, German, Spanish and Swedish

Ok but how do we evaluate the judge? 🤔

Model 1: **Of-course,** she is knowing the answer.

Model 2: **Of-course,** she knows the answer.

→ Model 2

LLM judge

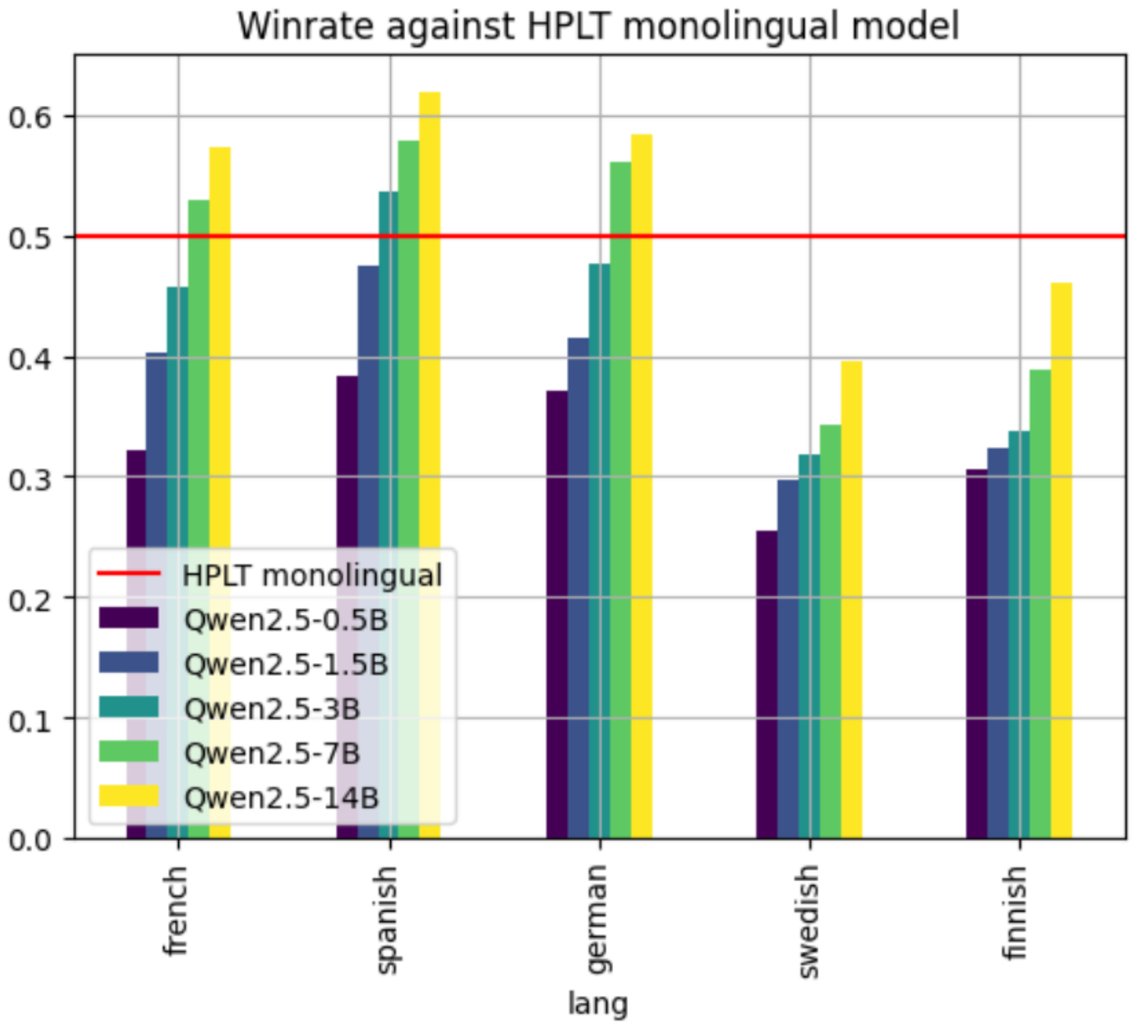
| | |
|---|---|
| general sentences | Die Sonne scheint warm über den Feldern und |
| general sentences | Das Kind lachte laut während es mit dem |
| history | Im Jahr 1914 begann ein Konflikt, der die |
| history | Die Entdeckung neuer Seewege veränderte den Handel und |
| economics | Das Angebot sank plötzlich, was den Preis auf |

| | |
|---|---|
| general sentences | Le chat noir traverse la rue vite |
| history | Napoléon a traversé le pays |
| economics | La politique monétaire influence les taux d'intérêt et l'investissement |
| math | La matrice invertible a un déterminant |
| programming | La boucle for itère sur chaque élément du tableau en |

# Measuring fluency

## Meta-evaluation

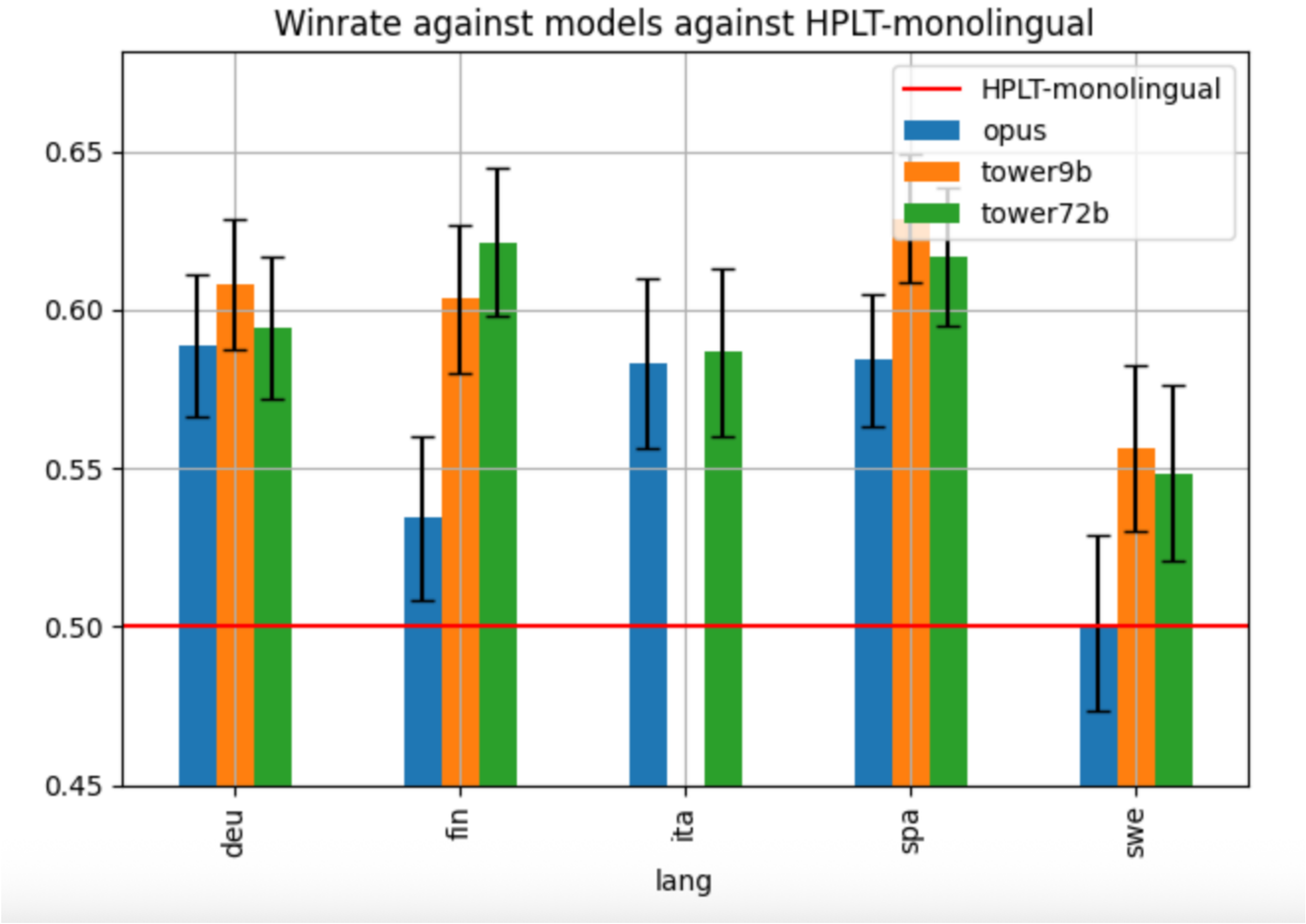
- Ok but how do we evaluate?
- Qualititative analysis: check that LLM judge can identify non-idiomatic text in French and English
- Monotonic analysis:
  - LLM judge: Deepseek-v3.1
  - Baseline: HPLT-1.7B trained on native monolingual datasets
  - Models: Qwen-2.5 series
- Bigger models are better



# Measuring fluency

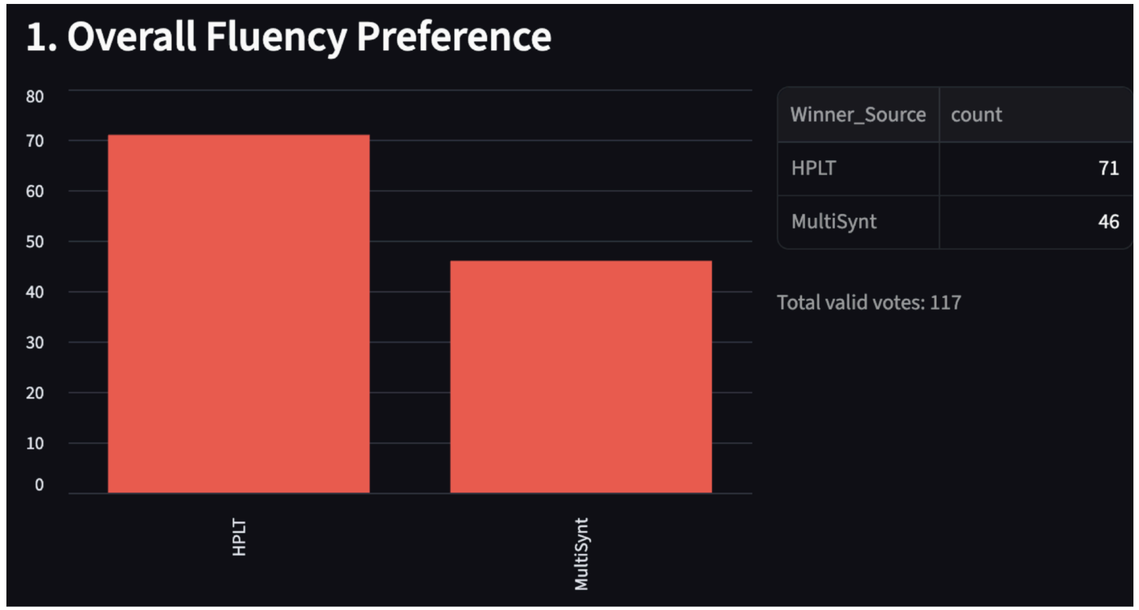
## Multisynt versus native

- Judge: Deepseek v3.1
- Winrate of models trained on multisynt data against monolingual HPLT model
- All models have 1.7B parameters and same architecture
- Roughly gets the same ordering of MT systems seen in our human evaluations: Opus < Tower-9B ~ Tower-72B



# Fluency

- We also did human evaluations
- And got a 64% win rate for native models…
- But models can be easily recognized and annotators may be biased
- Still ongoing debate and discussions about best ways to measure fluency



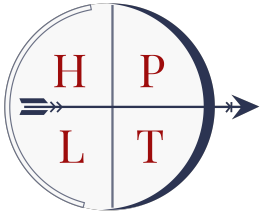
# Resources

- Dataset:
  - https://huggingface.co/datasets/geoalgo/multilingual-contexts-to-be-completed
- Report:
  - https://docs.google.com/document/d/1giDB4NnTzvZe2RfOxnYwnmCPgBpNtBjEBCLcHdNMp1Y/edit?usp=sharing
- Rerunning fluency on your model:
  - https://github.com/OpenEuroLLM/OpenJury

```
python openjury/generate_and_evaluate.py \
  --dataset fluency-german \
  --model_A gpt4_1106_preview \
  --model_B VLLM/utter-project/EuroLLM-9B \
  --judge_model OpenRouter/deepseek/deepseek-chat-v3.1 \
  --n_instructions 10
```

# Conclusion

- Overwhelming performance gain over native data for downstream evaluations
- Whether fluency is worse is still a debated question
- It is not saying that “native data is worse”
  - High quality data is filtered among a much larger set of tokens than other languages
  - Recall Pedro & Laurie’s talk, English is 44% of common-crawl, French is 4%
  - If the probability of generating a document with a quality score = 5 is equal between languages, there is much less French data with a quality score = 5
- There are other reasons that could explain the huge performance gain:
  - Contamination, Effect of translated benchmarks, Effect of translationese to boost performance, …





# HPLT-e & MultiSynt: A Norwegian Evaluation Deep-Dive

Vladislav Mikhailov, Stephan Oepen, Shenbin Qian; Language Technology Group

Circle U, NLPL, & OpenEuroLLM Winter School – February 3, 2026

# 1

# Project Background

# HPLT and OpenEuroLLM in a Nutshell


HEU & UKRI

2022-2025

8 Partners

1 Company

2 National HPC

Around 6 M€


DEP

2025-2028

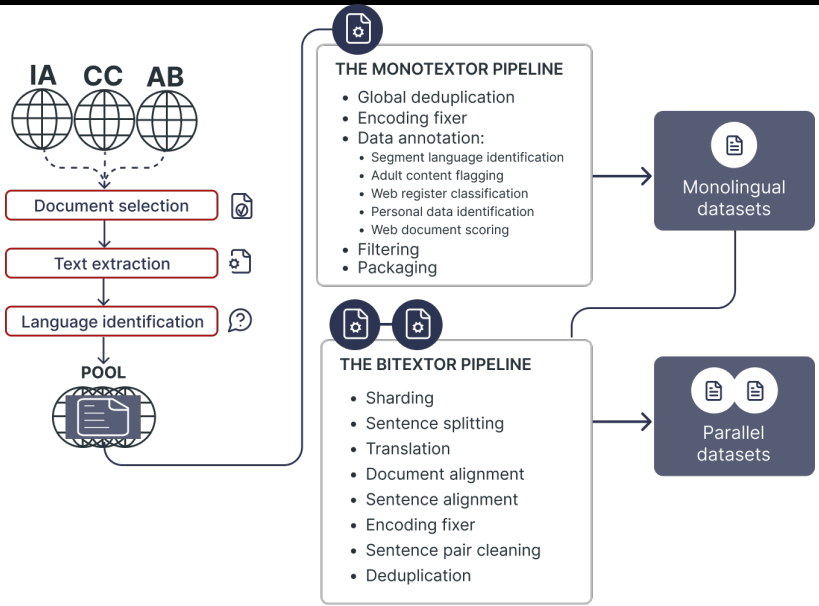
22 Partners

5 Companies

5 National HPC

Around 37 M€

# Three Years of Web-Scale Data Refinement



# HPLT 3.0: Some Contrastive Statistics

| Language | HPLT 3.0 T | HPLT 3.0 \|\| | HPLT 3.0 % | FineWeb 1 & 2 T | FineWeb 1 & 2 \|\| | FineWeb 1 & 2 % | HPLT 2.0 T | HPLT 2.0 \|\| | HPLT 2.0 % |
|---|---|---|---|---|---|---|---|---|---|
| **English** | 16T | 901 | 55 | 17T | 695 | 78 | 3.9T | 892 | 35 |
| **Multilingual** | 13T | 1187 | 45 | 4.9T | 976 | 22 | 7.2T | 1178 | 65 |
| **Basque** | 3.2B | 991 | 0.02 | 1.5B | 951 | 0.03 | 2.0B | 1030 | 0.03 |
| **Catalan** | 22B | 853 | 0.17 | 12B | 715 | 0.25 | 18B | 976 | 0.25 |
| **Czech** | 126B | 1171 | 0.93 | 67B | 1015 | 1.37 | 95B | 1266 | 1.32 |
| **Finnish** | 73B | 1491 | 0.55 | 48B | 1324 | 0.99 | 53B | 1538 | 0.74 |
| **French** | 584B | 968 | 4.32 | 292B | 811 | 5.95 | 379B | 943 | 5.24 |
| **Galician** | 3.1B | 772 | 0.02 | 1.8B | 695 | 0.04 | 2.7B | 906 | 0.04 |
| **Norwegian** | 52B | 1388 | 0.39 | 53B | 1318 | 1.09 | 42B | 1477 | 0.58 |
| **Spanish** | 658B | 908 | 4.86 | 329B | 746 | 6.71 | 471B | 936 | 6.51 |
| **Ukrainian** | 81B | 1014 | 0.60 | 49B | 938 | 1.02 | 60B | 1280 | 0.84 |

https://hplt-project.org/datasets/v3.0

# HPLT 3.0: Some Contrastive Statistics

| Language | HPLT 3.0 T | HPLT 3.0 \|\| | HPLT 3.0 % | FineWeb 1 & 2 T | FineWeb 1 & 2 \|\| | FineWeb 1 & 2 % | HPLT 2.0 T | HPLT 2.0 \|\| | HPLT 2.0 % |
|---|---|---|---|---|---|---|---|---|---|
| **English** | 16T | 901 | 55 | 17T | 695 | 78 | 3.9T | 892 | 35 |
| **Multilingual** | 13T | 1187 | 45 | 4.9T | 976 | 22 | 7.2T | 1178 | 65 |
| **Basque** | 3.2B | 991 | 0.02 | 1.5B | 951 | 0.03 | 2.0B | 1030 | 0.03 |
| **Catalan** | 22B | 853 | 0.17 | 12B | 715 | 0.25 | 18B | 976 | 0.25 |
| **Czech** | 126B | 1171 | 0.93 | 67B | 1015 | 1.37 | 95B | 1266 | 1.32 |
| **Finnish** | 73B | 1491 | 0.55 | 48B | 1324 | 0.99 | 53B | 1538 | 0.74 |
| **French** | 584B | 968 | 4.32 | 292B | 811 | 5.95 | 379B | 943 | 5.24 |
| **Galician** | 3.1B | 772 | 0.02 | 1.8B | 695 | 0.04 | 2.7B | 906 | 0.04 |
| **Norwegian** | 52B | 1388 | 0.39 | 53B | 1318 | 1.09 | 42B | 1477 | 0.58 |
| **Spanish** | 658B | 908 | 4.86 | 329B | 746 | 6.71 | 471B | 936 | 6.51 |
| **Ukrainian** | 81B | 1014 | 0.60 | 49B | 938 | 1.02 | 60B | 1280 | 0.84 |

https://hplt-project.org/datasets/v3.0

# HPLT 3.0: Some Contrastive Statistics

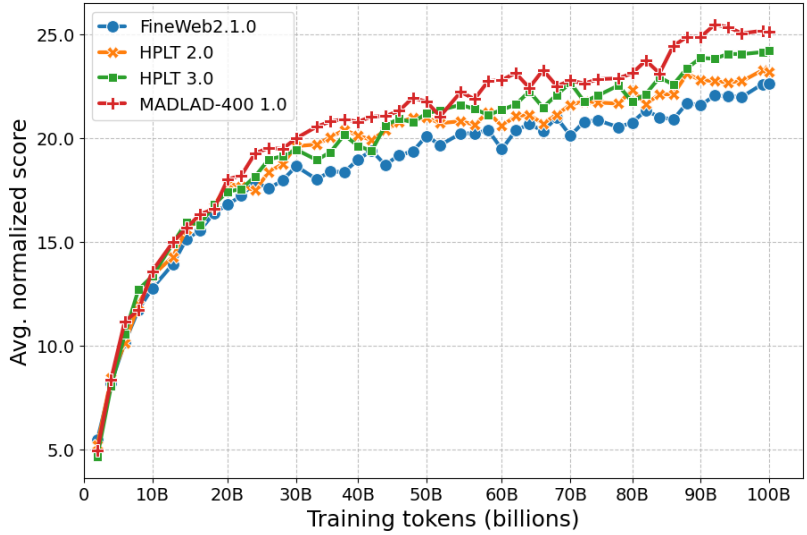
| Language | HPLT 3.0 T | HPLT 3.0 \|\| | HPLT 3.0 % | FineWeb 1 & 2 T | FineWeb 1 & 2 \|\| | FineWeb 1 & 2 % | HPLT 2.0 T | HPLT 2.0 \|\| | HPLT 2.0 % |
|---|---|---|---|---|---|---|---|---|---|
| **English** | 16T | 901 | 55 | 17T | 695 | 78 | 3.9T | 892 | 35 |
| **Multilingual** | 13T | 1187 | 45 | 4.9T | 976 | 22 | 7.2T | 1178 | 65 |
| **Basque** | 3.2B | 991 | 0.02 | 1.5B | 951 | 0.03 | 2.0B | 1030 | 0.03 |
| **Catalan** | 22B | 853 | 0.17 | 12B | 715 | 0.25 | 18B | 976 | 0.25 |
| **Czech** | 126B | 1171 | 0.93 | 67B | 1015 | 1.37 | 95B | 1266 | 1.32 |
| **Finnish** | 73B | 1491 | 0.55 | 48B | 1324 | 0.99 | 53B | 1538 | 0.74 |
| **French** | 584B | 968 | 4.32 | 292B | 811 | 5.95 | 379B | 943 | 5.24 |
| **Galician** | 3.1B | 772 | 0.02 | 1.8B | 695 | 0.04 | 2.7B | 906 | 0.04 |
| **Norwegian** | 52B | 1388 | 0.39 | 53B | 1318 | 1.09 | 42B | 1477 | 0.58 |
| **Spanish** | 658B | 908 | 4.86 | 329B | 746 | 6.71 | 471B | 936 | 6.51 |
| **Ukrainian** | 81B | 1014 | 0.60 | 49B | 938 | 1.02 | 60B | 1280 | 0.84 |

https://hplt-project.org/datasets/v3.0

# 2

# Multilingual Evaluation

# For Example: Dataset Comprison Across Languages




https://github.com/hplt-project/hplt-e/blob/main/results/2508-datasets/

# The Role Model: NorEval (Mikhailov, et al., 2025)

## NorEval: A Norwegian Language Understanding and Generation Evaluation Benchmark

**Vladislav Mikhailov**[1] **Tita Enstad**[2] **David Samuel**[1]
**Hans Christian Farsethås**[1] **Andrey Kutuzov**[1] **Erik Velldal**[1] **Lilja Øvrelid**[1]
[1]University of Oslo
[2]National Library of Norway
**Correspondence:** vladism@ifi.uio.no

https://aclanthology.org/2025.findings-acl.181/

# A Menagerie of Different LLM Capabilities

**NorEval** | BM, NN | 6 task types | 9 task categories | 24 datasets | 100+ prompts | Human-created

**Text classification**

- Sentence-level sentiment analysis
  - NoReC Sentence (BM)
- Document-level sentiment analysis
  - NoReC Document (BM)

**Sentence ranking**

- Norwegian language knowledge
  - NCB (BM)

**Sequence-to-sequence generation**

- Norwegian language knowledge
  - ASK-GEC (BM)
- Machine translation
  - Tatoeba (EN↔BM, EN↔NN)
- Text summarization
  - NorSumm (BM/NN)
- Instruction following
  - NorRewrite-IT (BM)
  - NorSummarize-IT (BM)

**Multiple-choice question answering**

- Commonsense reasoning
  - NorCommonsenseQA (BM/NN)
- Norwegian-specific & world knowledge
  - NorOpenBookQA (BM/NN)
  - NRK-Quiz-QA (BM/NN)
- Machine reading comprehension
  - Belebele (BM)
- Truthfulness
  - NorTruthfulQA MC (BM/NN)

**Generative question answering**

- Machine reading comprehension
  - NorQuAD (BM)
- Truthfulness
  - NorTruthfulQA Gen (BM/NN)

**Sentence completion**

- Norwegian language knowledge
  - NorIdiom (BM/NN)

# HPLT-e: Key Properties and Methodological Innovations

## Basic Approach

- "Our" nine European languages: CAT, CES, EUS, FIN, FRA, GLG, NOR, SPA, UKR;
- avoid (automatically) translated benchmarks: build on existing "native" collections: IberoBench, BenCzechMark, FinBench, FrenchBench, NorEval, UkrainianBench
- standardize on LM Evaluation Harness; aim to push all revisions back upstream;
- suitable for "early pre-training" application; thorough data-driven task selection.

# HPLT-e: Key Properties and Methodological Innovations

## Basic Approach

- "Our" nine European languages: CAT, CES, EUS, FIN, FRA, GLG, NOR, SPA, UKR;
- avoid (automatically) translated benchmarks: build on existing "native" collections: IberoBench, BenCzechMark, FinBench, FrenchBench, NorEval, UkrainianBench
- standardize on LM Evaluation Harness; aim to push all revisions back upstream;
- suitable for "early pre-training" application; thorough data-driven task selection.

## Multi-Prompt Design

- Prompt sensitivity serious methodological concern, e.g. Pezeshkpour, et al. (2025);
- HPLT-e: equip existing benchmarks with 3–7 human-created and diverse prompts;
- different options for score aggregation across prompts, e.g. average or maximum;
- average still "stricter" (or "arbitrary"), often leads to more narrow task selection.

https://github.com/hplt-project/hplt-e/

# Looking for Reliable Signals: Task Selection

▼ 🎯 **Task selection**

We use the standard task-specific metrics and report the maximum score across the prompts as the main performance aggregation method. We extend the FineWeb 2.1.0 evaluation design to examine the signal HPLT-e tasks provide based on the criteria and statistics summarized below.

- **Monotonicity**: performance should improve as pretraining progresses, even if the improvement differs across pretraining corpora. Tasks with fluctuating scores promote limited reliability.
- **Stable pretraining**: relative variability of performance across checkpoints should be low, reflecting smooth pretraining dynamics.
- **Ranking consistency**: relative ranking of models should remain consistent across consecutive pretraining intervals.
- **Prompt sensitivity**: performance should be consistent across various prompt formulations.
- **Prompt-switch rate**: frequent switches in best-performing prompt further reflects low evaluation reliability due to potential prompt lottery.
- **Signal-to-Noise ratio**: differences in task performance should primarily reflect differences in corpora quality, not random variation due to prompt choice.
- **Non-randomness**: final checkpoints should achieve performance above a random guessing baseline. Tasks where all models perform near random provide low discriminative power.

# 3

# MultiSynt Evaluations

# HPLT-e Active Norwegian Tasks Selected from NorEval

- NorCommonSenseQA (Norwegian Bokmål)
  - Multiple-choice question answering dataset for zero-shot evaluation of commonsense reasoning abilities with 1093 examples.
  - Evaluation metric: **accuracy**.
- NorIdiom (Nynorsk)
  - 1707 Norwegian idioms and phrases that appear more than 100 times in the online library of the National Library of Norway.
  - Evaluation metric: **exact match**.
- NorQuAD (Bokmål)
  - The first Norwegian question answering dataset for machine reading comprehension, created from scratch in Norwegian, consisting of 4,752 manually created question-answer pairs.
  - Evaluation metric: **F1 score**.
- NRK-Quiz-QA (Bokmål and Nynorsk)
  - Multiple-choice question answering dataset for zero-shot evaluation of Norwegian-specific and (some) world knowledge. It comprises 4.9k examples from over 500 quizzes on Norwegian language and culture
  - Evaluation metric: **accuracy**.

# Normalization and Aggregation across Tasks & Languages

- Step 1: Cross-prompt aggregation – pick the highest observed score (max aggregation).
- Step 2: Score normalization for each task following the Open LLM Leaderboard on Hugging Face.

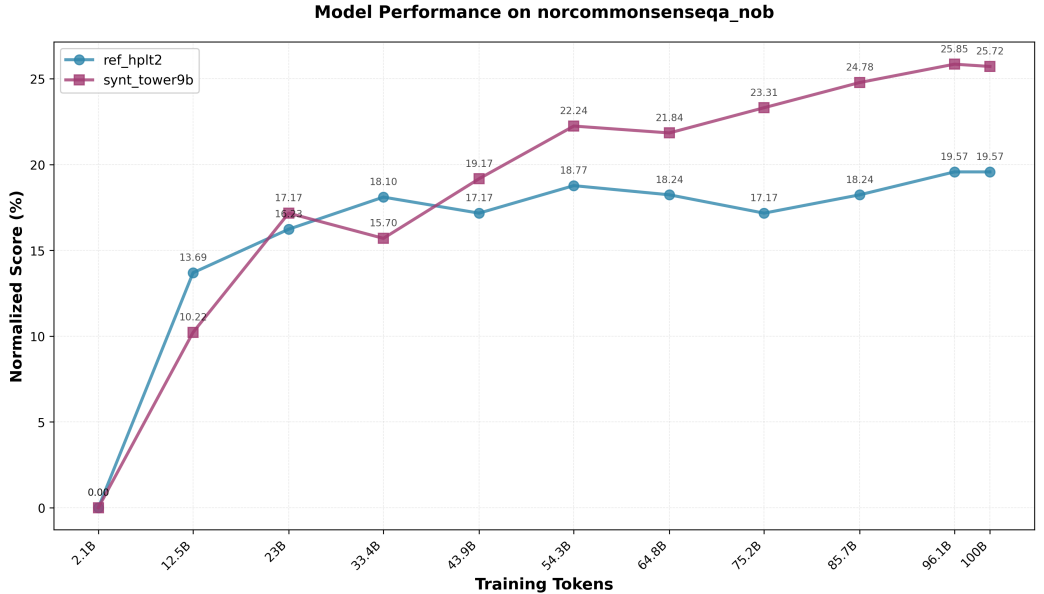
$$\text{normalized\_score} = \begin{cases} 0 & \text{if } x < L \\ \dfrac{x - L}{H - L} \times 100 & \text{if } x \geq L \end{cases} \quad \text{where} \quad \text{x} = \text{raw\_score}, \quad \text{L} = \text{lower\_bound=random\_baseline}, \quad \text{H} = \text{higher\_bound=100}$$

- Step 3: Per-category averaging – equal weights for different tasks under the same category such as NorCommonSenseQA under common sense reasoning.
- Step 4: Cross-category averaging.

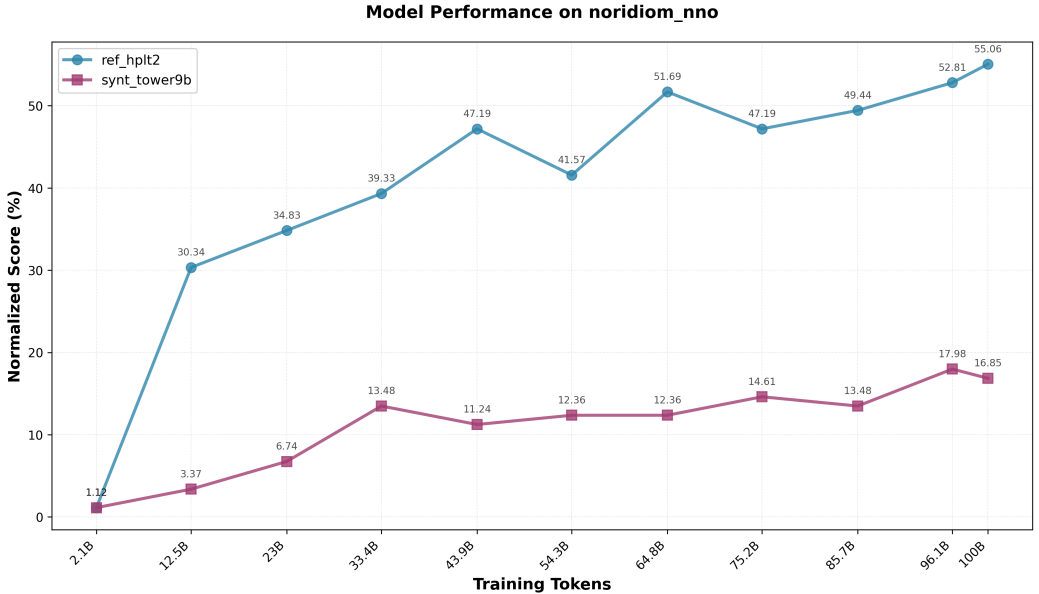
# Models and evaluation framework

- Reference model: trained on HPLT v2.0 Cleaned version for Norwegian
- Synthetic data (MultiSynt) model trained on synthetic data translated by Tower+9B
- Size: 2.15B
- Evaluation framework: lm-evaluation-harness

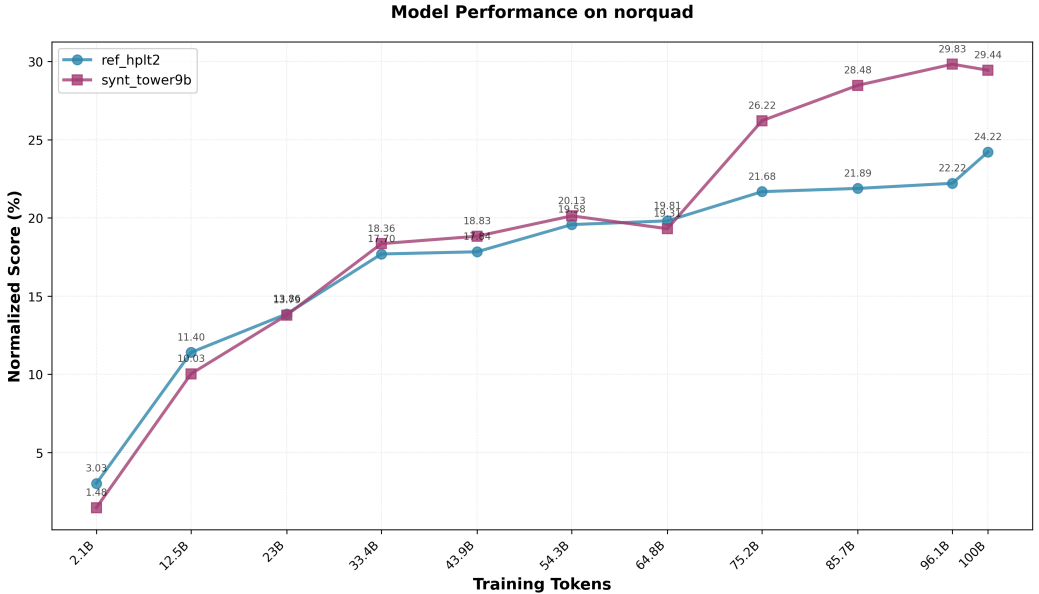
# For Example: Common Sense MCQA



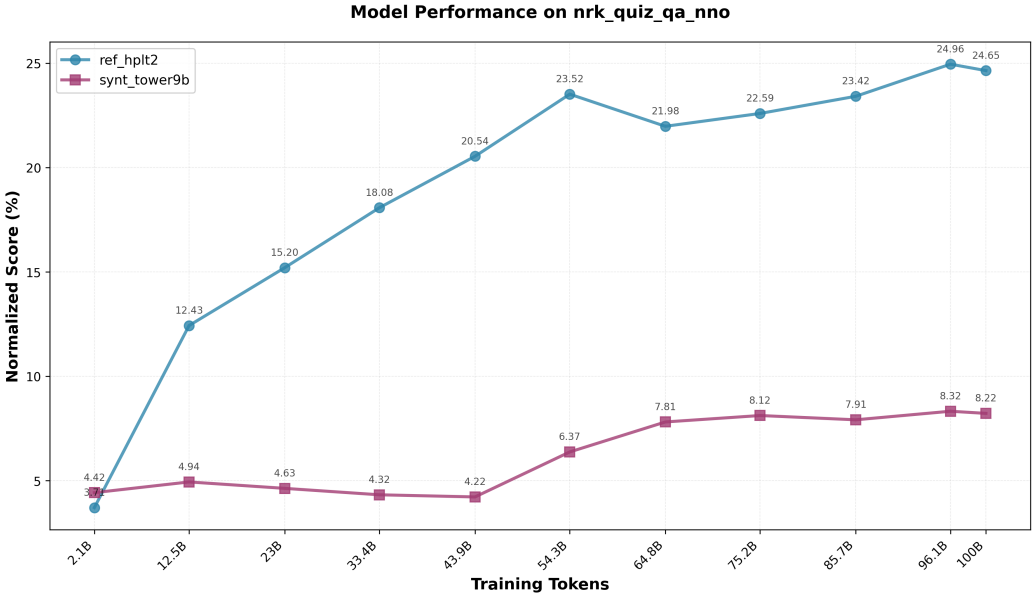
# For Example: Idiom Completion



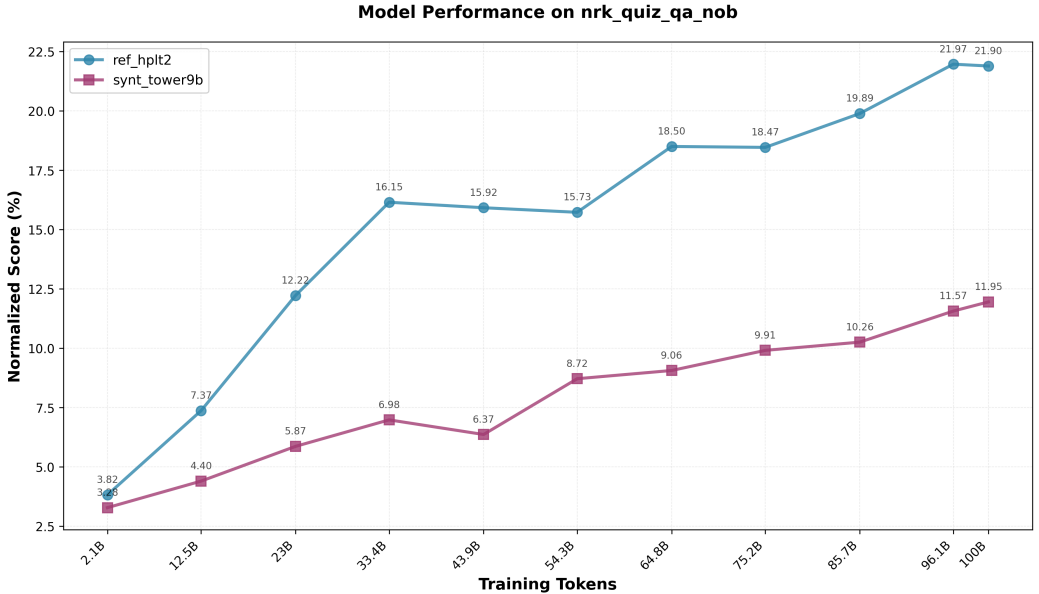
# For Example: Reading Comprehension



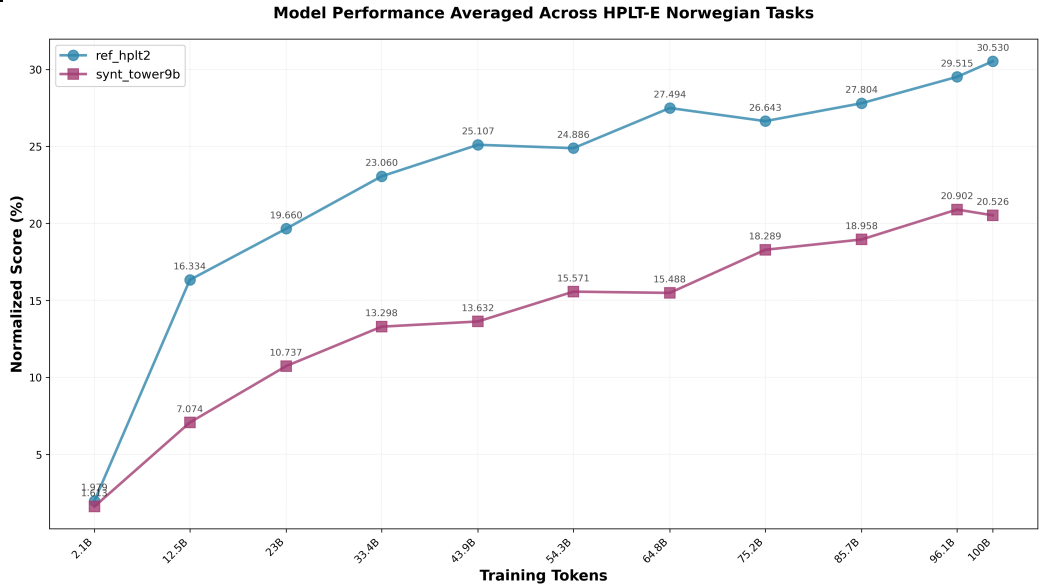
# For Example: Norwegian Knowledge MCQA (NNO)



# For Example: Norwegian Knowledge MCQA (NOB)



# Aggregate Language Score: Norwegian (Five Active Tasks)



# More Background in HPLT 3.0 Technical Report

## HPLT 3.0: Very Large-Scale Multilingual Resources for LLM and MT

### Mono- and Bi-lingual Data, Multilingual Evaluation, and Pre-Trained Models


**Stephan Oepen[♣], Nikolay Arefev[♣], Mikko Aulamo[♠], Marta Bañón[♡], Maja Buljan[♣], Laurie Burchell[♢], Lucas Charpentier[♣], Pinzhen Chen[•], Mariia Fedorova[♣], Ona de Gibert[♠], Barry Haddow[•], Jan Hajič[◦], Jindřich Helcl[♣], Andrey Kutuzov[♣], Veronika Laippala[⋆], Zihao Li[♠], Risto Luukkonen[⋆], Bhavitvya Malik[•], Vladislav Mikhailov[♣], Amanda Myntti[⋆], Dayyán O'Brien[•], Lucie Poláková[◦], Sampo Pyysalo[⋆], Gema Ramírez Sánchez[♡], Janine Siewert[♠], Pavel Stepachev[•], Jörg Tiedemann[♠], Teemu Vahtola[♠], Dušan Variš[◦], Fedor Vitiugin[⋆], Tea Vojtěchová[◦], Jaume Zaragoza[♡]**

[♣] University of Oslo, Department of Informatics
[♠] University of Helsinki, Department of Digital Humanities
[♡] Prompsit Language Engineering
[♢] The Common Crawl Foundation
[•] Edinburgh University, School of Informatics
[◦] Charles University, Prague, Institute of Formal and Applied Linguistics
[⋆] TurkuNLP, University of Turku, Department of Computing
oe@ifi.uio.no


Nov 2025

https://arxiv.org/abs/2511.01066

# Contact

**Shenbin Qian**

Language Technology Group

shenbinq@ifi.uio.no