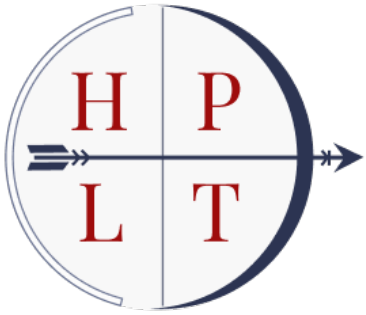
Co-funded by the European Union



HPLT: High Performance Language Technologies

Report on language model evaluation

Deliverable number: 4.2

Version 1.0

Horizon Europe

UK Research and Innovation

Funded by the European Union's Horizon Europe search and innovation programme under grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546] programme

Project details

Project Acronym:	HPLT
Project Full Title:	HPLT: High Performance Language Technologies
Year of the Call:	2021
Type of Action:	HORIZON-IA (Innovation Action)
Grant Number:	101070350
Project URL:	https://hplt-project.org

Report details

Report on language model evaluation	
Lead author:	Sampo Pyysalo (UTU)
Contributing authors:	Pinzhen Chen (UEdin) Mariia Fedorova (UiO) Barry Haddow (UEdin) Andrey Kutuzov (UiO) Vladislav Mikhailov (UiO) Stephan Oepen (UiO) Fedor Vitiugin (UTU)
Internal reviewers:	Jaume Zaragoza (Prompsit) Jan Hajič (CUNI)
Deliverable number:	4.2
Dissemination level:	Public (PU)
Contractual Delivery Date:	Nov 30, 2025 (extension granted)
Actual Delivery Date:	Nov 30, 2025
Number of pages:	33

Document history

1.0	Nov 30, 2025	Original Submission

Abstract

This report provides a description of deliverable D4.2 – the final report on language model evaluation in the HPLT project. The report presents the evaluation of decoder-only, encoder-only and encoder-decoder models, detailing newly introduced benchmarks and evaluation results as well as additional evaluation efforts led by project partners.

Contents

1 Executive summary

1.1 Introduction

This deliverable, *Report on language model evaluation*, describes the evaluation efforts in WP4 at the end of the HPLT project. The document is organized by model class: the evaluation of decoder-only (GPT-like) models is presented in Section 2, encoder-only (BERT-like) models in Section 3, and encoder-decoder (T5-like) models in Section 4. Section 5 presents other evaluation efforts and Section 6 provides additional technical details.

1.2 Brief summary of the HPLT project

The Horizon Europe initiative HPLT (High-Performance Language Technologies) applies high-performance computing to scale up and advance language technologies, with particular emphasis on languages beyond English. Taking advantage of recent advances in machine learning and combining massive storage and compute capabilities, the project creates monolingual data sets comprising trillions of tokens over hundreds of languages. Through automated and large-scale discovery of translational equivalences across languages, HPLT further produces very large parallel data sets for dozens of languages. These data resources are validated through the development and evaluation of different types of language and translation models for a broad and typologically diverse range of languages. HPLT capitalizes on transparency and replicability. All data sets, models, and software pipelines are made available to the general public under permissive licenses.

The project, coordinated by Charles University in Prague (CUNI), gathers partners from five different universities, two national HPC centers, and a private NLP company from all around Europe. In the first half of the project, the University of Edinburgh (UEDIN) served as a technical coordinator for the consortium, while the University of Oslo (UiO) took on this role for the remaining project duration.


CHARLES UNIVERSITY


UNIVERSITY OF OSLO


UNIVERSITY OF EDINBURGH


UNIVERSITY OF TURKU


UNIVERSITY OF HELSINKI


PROMPSIT


CESNET


SIGMA2

Partners from University of Turku, University of Oslo, and Charles University were involved in WP4, on which this report is focused.

2 Decoder-only language models

A broad range of decoder-only models were trained and evaluated to assess different corpora and data sampling strategies. This section presents the evaluation results for the following key experiments:

- **Deduplication Strategies**: Comparison of HPLT 2.0 (Burchell et al., 2025) and pre-HPLT 3.0 data deduplication strategies across nine selected languages (HPLT 3.0 pre-release; see §2.4.1).
- **Dataset Comparison**: Evaluation of HPLT 2.0, HPLT 3.0, FineWeb2.1.0 (Penedo et al., 2025), and MADLAD-400 1.0 (Kudugunta et al., 2023) on nine selected languages (HPLT 3.0 release; see §2.4.2).
- **WDS-Based Sampling**: Analysis of HPLT 3.0 corpora sampled using different Web Document Scorer thresholds, focusing on Spanish and French (HPLT 3.0 release; see §2.4.3).

All models, evaluation results, and the evaluation framework are publicly released as part of this deliverable. We note that additional evaluation results for previously released decoder-only models have been detailed in deliverable D4.1 (*First language models trained*); these results are not repeated here.

2.1 Evaluation design

For the evaluation of decoder-only models, we developed HPLT-E, a framework for automated large-scale multilingual evaluation designed to systematically assess models and compare and refine data preparation choices across nine selected languages: Basque, Catalan, Czech, Finnish, French, Galician, Norwegian (Bokmål and Nynorsk), Spanish, and Ukrainian. These languages are chosen to ensure both the availability of native speakers in our team and a minimum level of diversity in terms of language resources, families, and scripts. HPLT-E includes 124 language understanding and generation tasks, each supporting 3–7 human-written prompts. We aim to cover different task categories in all languages: entailment, commonsense reasoning, language-specific and world knowledge, paraphrase detection, reading comprehension, sentiment analysis, toxicity detection, and truthfulness. HPLT-E integrates with LM Evaluation Harness (Gao et al., 2024), for experimental flexibility and replicability. Our framework is available at `https://github.com/hplt-project/hplt-e`.

Benchmarks We combine open-source human-curated multi-task benchmarks: IberoBench (Baucells et al., 2025) (Catalan, Spanish, Basque, Galician), FrenchBench (Faysse et al., 2024) (French), NorEval (Mikhailov et al., 2025) (Norwegian Bokmål and Nynorsk), BenCzechMark (Fajcik et al., 2025) (Czech), and Finbench v2 built on Finbench (Luukkonen et al., 2023) (Finnish). In addition, we create a benchmark for Ukrainian (UkrainianBench), which comprises Global MMLU (Singh et al., 2025), INCLUDE (Romanou et al., 2025), UA-SQuAD (Ivanyuk-Skulskiy et al., 2021), ZNO (Romanyshyn et al., 2024), Belebele (Bandarkar et al., 2024), TextDetox (Dementieva et al., 2024), and WMT24++ (Deutsch et al., 2025).[1]

Prompt Collection HPLT-E enables evaluation across 500+ prompts that have diverse structure and answer formatting to mitigate prompt sensitivity, a model limitation where variations in prompt

[1]The full list of supported tasks can be found in our GitHub repository at `https://github.com/hplt-project/hplt-e`.

formulation can affect downstream performance (Pezeshkpour and Hruschka, 2024; Sclar et al., 2024). We adapt the single-prompt benchmarks (IberoBench, FrenchBench, and UkrainianBench) to the multi-prompt design through (i) manual translation of English prompts from PromptSource (Bach et al., 2022) and (ii) prompt creation by native speakers.

Task Selection We evaluate the models at regular pretraining intervals (every 1B tokens) in a 0-shot regime using the standard task-specific metrics. We report the maximum score across the prompts as the main performance aggregation method. We extend the FineTasks evaluation design of Penedo et al. (2025) and select tasks that provide pretraining evaluation signal based on the following criteria:

- **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 (Chen et al., 2023).
- **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.

Specific requirements for each evaluation series can be found in the corresponding subsection.

Performance Aggregation Following Penedo et al. (2025); Fourrier et al. (2024), we compute a *language score* as the average of min–max-normalized performance scores across selected tasks. In particular, we first rescale all scores between the random baseline and the maximum possible score. We then average the rescaled scores within each task category and take the average of per-category scores to compute the language score. To produce a *multilingual score*, we utilize several approaches:

- **Average language score**: We average min-max normalized language scores.
- **Average multilingual rank**: We rank the final checkpoints' language scores across all corpora configurations and average their ranks.
- **Borda count**: Borda count (Colombo et al., 2022; Rofin et al., 2023) assigns the lowest-ranked model zero points, the second-lowest-ranked one points, etc. This alternative to average-based aggregation allows for the aggregation of heterogeneous metrics by leveraging rank-based differences. First, we rank the final checkpoints for each language; second, we apply the Borda count on the language-wise rankings to compute the final ranking.

2.2 Training data

Each evaluation series involves pretraining individual 2.15B-parameter decoder-only language models for every language, following a fixed pretraining setup. For lower-resource languages with less than 30B/100B tokens of available data, datasets are uniformly upsampled (repeated) following Muennighoff et al. (2023).

2.3 Model description

All models employ the Gemma-3 tokenizer and follow the Llama architecture (Touvron et al., 2023) with 24 layers, 32 attention heads, and a sequence length of 2048. Pretraining is run using the Megatron-LM framework (Shoeybi et al., 2019) on the LUMI supercomputer, employing 16 AMD MI250x nodes. Additional details on the model and training configuration are provided as part of the model release.

2.4 Evaluation results

2.4.1 Deduplication Strategies

- **Models**: https://huggingface.co/collections/HPLT/2505-deduplication
- **Evals:** https://huggingface.co/datasets/HPLT/2505-deduplication-evals

This section describes the key results of our HPLT 3.0 pre-release evaluations comparing models trained with different data deduplication stategies for the pre-HPLT 3.0 corpora and the previous HPLT 2.0 version. We pretrain and evaluate decoder-only models on 30B tokens for each language as described in §2.3. We compare the following data deduplication strategies to guide our design choices, and guard against data quality regression compared to HPLT 2.0:

- **pre-HPLT 3.0 CD**: per-crawl deduplication.
- **pre-HPLT 3.0 GD**: global deduplication.
- **pre-HPLT 3.0 GDR**: global deduplication & rehydration.

Task Selection Table 6.1 (see §6.1) shows our task selection requirements for the 30B-tokens models.

Corpus	Avg. rank	Borda count
HPLT 2.0	[4] 3.37	[4] 5
pre-HPLT 3.0 CD	[3] 2.50	[3] 10
pre-HPLT 3.0 GD	[2] 2.25	[2] 11
pre-HPLT 3.0 GDR	[1] 1.87	[1] 18

Table 2.1: Average multilingual ranks and Borda counts for the 30B-tokens models in §2.4.1. Rank indicators are shown in square brackets.

Key Takeaways In this ablation study, we analyzed over 50,000 performance scores and report the results across 32 selected tasks. Table 2.1 shows the aggregation-based results, while Figure 2.1 depicts

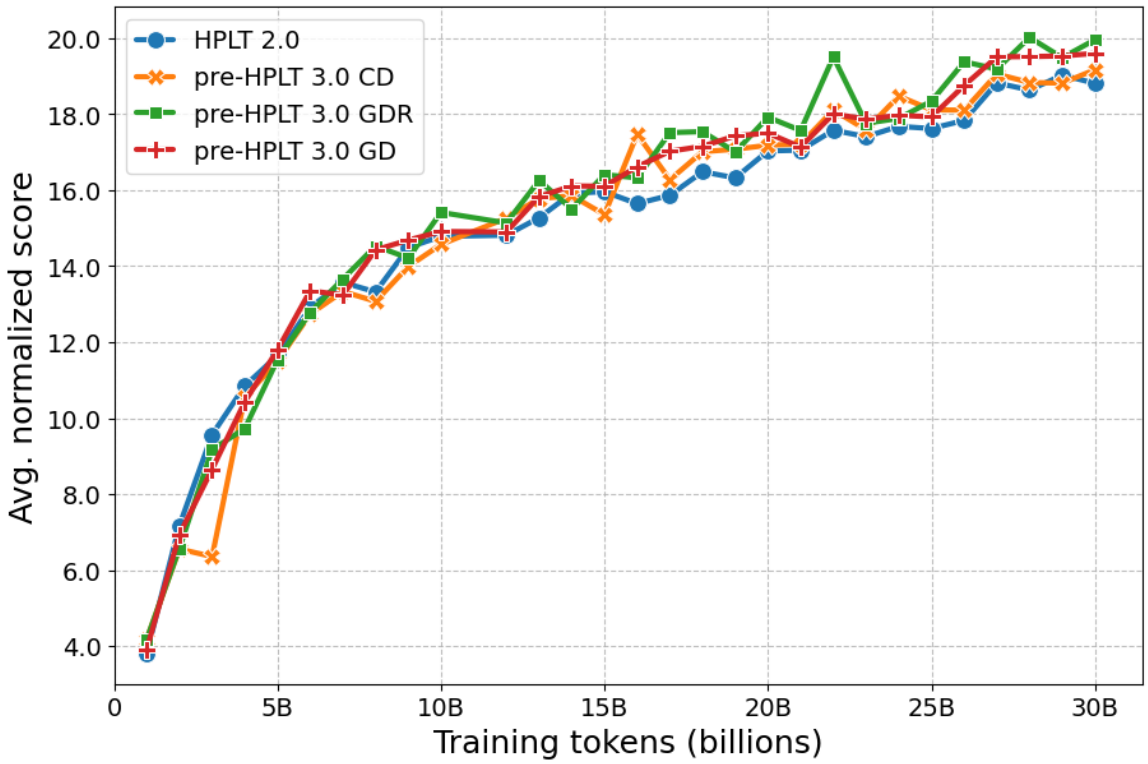


Figure 2.1: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

the pretraining trajectories based on the average normalized score. The results for each language can be found in §6.2.

Our pre-release pretraining corpus comparison shows that LLMs pretrained on HPLT 3.0 consistently outperform those pretrained on HPLT 2.0 across the HPLT-E languages. In particular, HPLT 3.0 models achieve stronger results for Ukrainian, Basque, Catalan, Finnish, and French, perform comparably for Czech, and show decreases for Spanish and Norwegian.

We also observe that no tasks for Galician meet the task selection criteria. For Ukrainian, Czech, and French, only a single SQuAD-style task is selected. These are included due to their generative nature, even though they slightly violate the **Stable pretraining** criterion, which reflects the smoothness of pretraining trajectories.

Overall, the results indicate that global deduplication (**pre-HPLT 3.0 GD**) and rehydration (**pre-HPLT 3.0 GDR**) provide the strongest performance gains, with language-specific variation. Based on these results, we adopt global deduplication as our default design decision, and provide the user with the opportunity to obtain the rehydrated version of our HPLT 3.0 corpora.

2.4.2 Dataset Comparison

- **Models**: https://huggingface.co/collections/HPLT/2508-datasets
- **Evals:** https://huggingface.co/datasets/HPLT/2508-datasets-evals

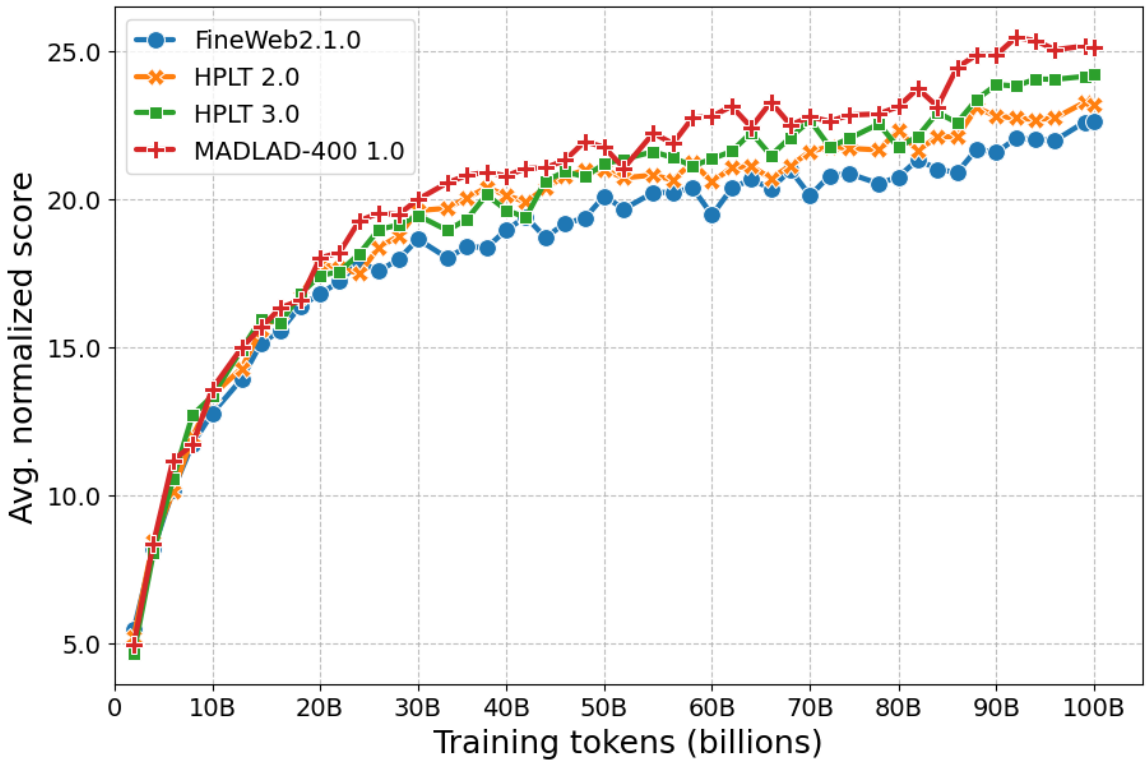


Figure 2.2: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb 2.1.0, and MADLAD-400 1.0.

This section presents the key results from our HPLT 3.0 release evaluations comparing models trained on the new HPLT 3.0 corpora with models trained on the previous HPLT 2.0 version, FineWeb 2.1.0, and MADLAD-400 1.0. We pretrain and evaluate decoder-only models on 100B tokens for each language as described in §2.3.

Task Selection Table 6.2 (see §6.1) shows our task selection requirements for the 100B-tokens models.

Corpus	Avg. rank	Borda count
HPLT 3.0	[2] 2.43	[3] 8
HPLT 2.0	[4] 3.28	[4] 3
MADLAD-400 1.0	[1] 1.71	[1] 15
FineWeb2.1.0	[3] 2.57	[2] 9

Table 2.2: Average multilingual ranks and Borda counts for the 100B-tokens models in §2.4.2. Rank indicators are shown in square brackets.

Key Takeaways In this ablation study, we analyze over 96,000 performance scores. Table 2.2 shows the aggregation-based results, while Figure 2.2 depicts the pretraining trajectories based on the average normalized score. The results for each language can be found in §6.3.

We find that tasks for lesser-resourced languages, notably Basque and Galician, are unsuitable for pretraining evaluation due to their relative difficulty, evaluation data quality, and the lack of monotonic performance progression during pretraining. We thus report our key findings on a final suite of 26

selected tasks across the seven remaining languages.

All models show monotonic performance improvement on our selected tasks as pretraining progresses. Models pretrained on MADLAD-400 1.0 achieve the highest multilingual score, followed by HPLT 3.0, while HPLT 2.0 and FineWeb perform on par. These results are consistent with rank-based aggregation. The models are ranked as (1) MADLAD-400 1.0; (2) HPLT 3.0; (3) FineWeb2.1.0; and (4) HPLT 2.0; by average multilingual ranks, HPLT 3.0 slightly outperforms FineWeb2.1.0, whereas Borda counts show the inverse ordering. Overall, our findings indicate that refined data preparation in HPLT 3.0 has improved average dataset quality, which translates into competitive performance gains for model pretraining.

2.4.3 WDS-Based Sampling

- **Models**: https://huggingface.co/collections/HPLT/2508-wds
- **Evals:** https://huggingface.co/datasets/HPLT/2508-wds-evals

This section describes the key results from our HPLT 3.0 release evaluations comparing models trained on the new HPLT 3.0 corpora sampled using different Web Document Scorer (WDS) thresholds, focusing on Spanish and French. We pretrain and evaluate decoder-only models on 100B tokens for both languages as described in §2.3.

Task Selection Table 6.2 (see §6.1) shows our task selection requirements for the 100B-tokens models.

Corpus	Avg. rank	Borda count
Top	[1] 1.5	[2] 2
Random	[1] 1.5	[1] 3
Bottom	[2] 3.0	[3] 0

Table 2.3: Average multilingual ranks and Borda counts for the 100B-token models in §2.4.3. Rank indicators appear in square brackets.

Key Takeaways In this ablation study, we analyze over 10,500 performance scores and report the results across 6 selected tasks across the two languages. Table 2.3 shows the aggregation-based results, while Figure 2.3 depicts the pretraining trajectories based on the average normalized score. The results for each language can be found in §6.4.

Here, **Random** sampling represents the default approach, drawing uniformly on the full corpus, while **Top** and **Bottom** take advantage of the sorting by WDS levels and sequentially draw 100B training tokens from either end of the corpus. Low WDS levels clearly lead to inferior model performance, while sampling from only the **Top** does not clearly improve over the full corpus, possibly owing to overly limited diversity.

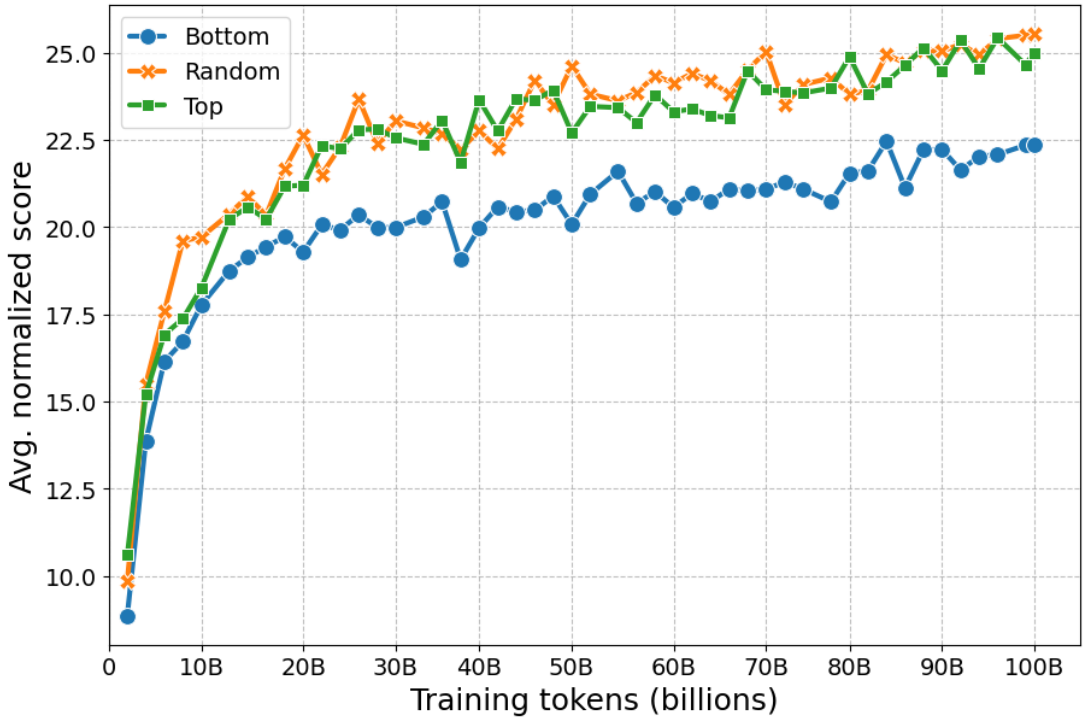


Figure 2.3: Comparison of different WDS-based sampling strategies on 100B tokens from HPLT 3.0.

3 Encoder-only language models

We trained encoder-only language models using the standard masked language modeling (MLM) objective on the monolingual HPLT 2.0 datasets following the LTG-BERT (Samuel et al., 2023a) architecture to allow comparison with models trained on the HPLT v1.2 release data. This section details the evaluation and comparison of these encoder-only models.

3.1 Evaluation design

We evaluate the trained LTG-BERT models on part-of-speech (POS) tagging, lemmatization and dependency parsing using the Universal Dependencies (UD) treebanks (de Marneffe et al., 2021), as well as named entity recognition (NER) using the WikiAnn datasets (Pan et al., 2017). We compare to mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) models as multilingual baselines, and to HPLT v1.2 BERT models[1] as monolingual baselines. The performance is measured using the official CoNLL 2018 evaluation code (Zeman et al., 2018) for the UD tasks, and `seqeval` (Nakayama, 2018) balanced F1 score for the NER task.

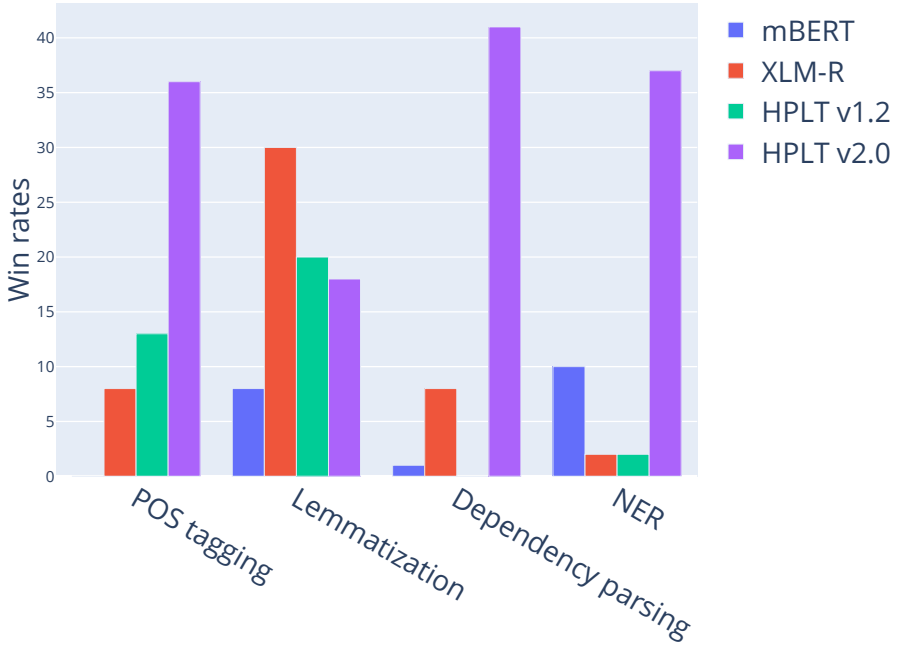


Figure 3.1: Win rates for MLMs at part-of-speech tagging, lemmatisation, dependency parsing, and named entity recognition.

3.2 Evaluation results

Figure 3.1 shows the win rates achieved by the models for the four tasks, where win rate is the number of languages on which a given model outperforms other models. Models trained on the HPLT 2.0 datasets

[1] https://huggingface.co/collections/HPLT/hplt-12-bert-models

show a considerably higher win rate compared to the baselines in all the tasks except lemmatization, where XLM-R and HPLT v1.2 yield competitive results. However, we note that the difference between XLM-R, HPLT v1.2 and HPLT 2.0 on the lemmatization task is less than 1% in accuracy, meaning that no model notably outperforms any other. Detailed scores by language and task are shown in Table 3.1. We make the HPLT 2.0 BERT models with intermediate checkpoints publicly available.[2]

	POS tags				Lemmas				Dependency parsing				NER			
Language	mBERT	XLM-R	HPLT v1.2	HPLT 2.0	mBERT	XLM-R	HPLT v1.2	HPLT 2.0	mBERT	XLM-R	HPLT v1.2	HPLT 2.0	mBERT	XLM-R	HPLT v1.2	HPLT 2.0
als_Latn	59.1	61.6	64.0	**64.5**	**78.2**	75.0	76.3	77.2	**33.1**	29.3	25.3	24.7	92.3	92.9	92.4	**93.9**
bel_Cyrl	94.1	94.6	95.5	**95.7**	93.2	93.8	93.8	**97.1**	88.1	89.9	91.1	**91.7**	91.7	90.3	90.1	**92.8**
bos_Latn	95.5	96.2	96.4	**96.6**	97.2	**97.4**	97.2	97.1	90.2	91.3	91.3	**91.7**	91.5	91.6	89.3	**92.8**
hrv_Latn	95.5	96.2	96.4	**96.8**	97.2	**97.4**	97.2	97.1	90.2	91.3	91.3	**91.6**	91.5	91.6	89.3	**92.5**
bul_Cyrl	97.0	97.5	97.8	**97.9**	97.5	**97.7**	97.3	97.3	92.7	94.4	94.0	**94.5**	92.2	92.2	91.5	**93.0**
cat_Latn	97.1	97.2	97.4	**97.5**	99.4	99.4	99.4	**97.5**	93.6	94.1	94.4	**99.4**	92.1	91.0	90.1	**94.5**
ces_Latn	97.8	98.0	98.3	**98.4**	99.3	99.3	99.4	99.4	93.5	94.2	94.4	**94.6**	91.2	91.2	89.0	**91.8**
cym_Latn	87.2	88.3	**89.2**	89.0	**94.6**	94.4	93.7	92.3	80.8	**82.8**	82.3	**82.8**	92.5	90.0	89.4	**93.4**
dan_Latn	96.7	97.8	97.8	**97.9**	97.2	**97.6**	97.1	97.1	86.7	89.1	88.8	**89.5**	91.2	91.6	90.3	92.0
deu_Latn	88.8	89.4	80.7	**89.9**	97.6	**97.7**	95.5	97.5	84.6	87.1	76.4	**87.6**	**89.4**	87.7	64.1	89.2
ell_Grek	94.6	95.7	96.1	**96.2**	94.6	**94.7**	94.1	94.1	91.7	**93.5**	92.2	93.2	90.2	90.7	90.2	**92.6**
eng_Latn	96.1	96.8	96.7	**97.0**	97.8	98.0	97.9	**98.1**	91.3	92.6	92.2	**93.0**	2.2	81.1	81.0	**82.7**
spa_Latn	95.7	95.9	96.0	**96.2**	**99.4**	**99.4**	**99.4**	99.4	92.3	93.0	93.1	**93.4**	**90.9**	89.9	89.6	90.8
est_Latn	96.0	96.6	**97.1**	**97.1**	94.8	95.0	**95.2**	**95.2**	88.1	89.7	90.8	**91.0**	91.8	90.4	89.6	**93.0**
eus_Latn	91.0	91.4	**92.3**	**92.3**	95.7	95.9	**96.0**	95.9	85.3	87.3	88.1	**88.2**	91.3	90.7	89.8	**92.9**
pes_Arab	95.9	96.3	**96.4**	96.3	99.1	99.4	99.4	**99.5**	92.7	93.8	93.9	**94.1**	92.0	92.9	91.8	**93.9**
fin_Latn	95.1	96.4	96.8	**97.0**	90.6	91.5	**91.6**	91.4	90.2	93.0	93.3	**94.0**	90.2	90.0	89.2	**91.6**
fra_Latn	97.8	**98.1**	**98.1**	98.0	98.6	**98.8**	93.8	98.6	93.8	94.4	94.5	**94.8**	**90.5**	88.7	87.2	90.0
gle_Latn	86.5	87.1	88.7	**89.3**	95.5	95.8	**96.1**	95.6	81.3	82.7	83.4	**84.3**	**80.8**	78.0	55.9	78.2
glg_Latn	96.9	**97.1**	**97.1**	97.0	**98.3**	**98.3**	98.2	98.0	82.3	**82.6**	82.3	82.2	92.5	93.3	91.1	**94.1**
heb_Hebr	95.6	96.1	96.5	**96.7**	97.0	**97.2**	97.1	**97.2**	89.8	91.6	91.0	**91.9**	2.6	84.2	88.4	**89.3**
hin_Deva	92.4	93.3	93.6	**93.7**	98.9	**99.0**	**99.0**	**99.0**	92.6	93.3	93.5	**93.6**	88.6	88.0	84.3	**89.5**
hrv_Latn	95.5	96.2	96.4	**96.7**	97.2	**97.4**	97.2	97.2	90.2	91.3	91.3	**91.8**	91.5	91.6	89.3	**92.0**
hun_Latn	93.0	**94.3**	93.0	94.1	93.0	**94.3**	93.0	92.3	84.3	**86.7**	82.4	86.1	92.2	91.9	92.8	**93.1**
hye_Armn	88.7	91.2	**92.7**	**92.7**	94.4	**94.9**	93.9	94.7	80.4	85.3	84.1	**86.8**	95.7	95.3	94.8	**95.9**
ind_Latn	89.5	**89.8**	89.6	89.1	98.2	**98.3**	98.0	97.5	82.4	**82.7**	81.7	81.8	91.3	91.6	89.1	**92.0**
isl_Latn	87.7	88.1	88.6	**88.7**	96.2	96.4	**96.5**	96.4	85.2	86.6	86.9	**87.4**	**81.7**	63.9	55.9	78.3
ita_Latn	98.0	98.0	98.1	**98.3**	98.6	98.7	**98.8**	98.7	94.1	94.4	94.6	**95.1**	90.5	89.7	87.8	**91.2**
jpn_Jpan	97.5	97.7	**97.8**	**97.8**	98.3	98.3	98.3	**98.4**	94.1	94.6	94.6	**94.8**	66.5	65.9	**67.4**	67.2
kat_Geor	91.3	**92.6**	92.4	92.4	**92.8**	93.7	92.5	92.5	79.5	80.9	80.8	**81.3**	87.2	4.7	89.6	**90.7**
kor_Hang	88.6	89.7	89.9	**90.1**	94.0	94.3	**94.4**	94.4	88.0	89.0	89.4	**89.7**	87.8	87.0	88.3	**89.3**
lvs_Latn	91.6	92.8	92.4	**93.6**	96.9	91.6	96.8	**97.7**	88.8	90.9	90.9	**92.1**	93.2	92.6	90.7	**93.9**
lit_Latn	87.7	91.9	92.0	**92.5**	90.2	**91.6**	91.5	91.2	79.3	85.7	84.9	**86.8**	89.1	89.3	87.0	**91.0**
ltz_Latn	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.2
mkd_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.6
mlt_Latn	94.7	94.5	97.0	**97.7**	**100.0**	**100.0**	**100.0**	**100.0**	78.2	78.5	83.2	**87.2**	-	-	-	-
nob_Latn	97.0	97.4	**97.6**	97.5	98.5	**98.8**	**98.8**	98.7	93.2	94.3	94.5	**94.7**	91.9	92.6	91.1	**93.2**
nld_Latn	96.2	96.9	97.1	**97.2**	94.1	**94.7**	94.4	94.1	91.6	92.9	93.8	**94.1**	**91.7**	90.4	88.6	91.0
nno_Latn	96.6	97.0	97.7	**97.8**	98.2	98.4	**98.5**	**98.5**	92.9	93.9	94.6	**95.0**	**95.8**	93.6	93.2	95.5
pol_Latn	95.6	95.5	96.9	**97.2**	97.8	**98.2**	**98.2**	**98.2**	93.7	95.2	95.3	**95.6**	12.9	88.8	**89.7**	89.6
por_Latn	93.6	94.0	**94.1**	**94.1**	98.1	**98.3**	**98.3**	98.2	83.4	84.5	84.9	**85.3**	91.2	90.3	88.0	**91.5**
ron_Latn	97.3	97.6	97.7	**97.9**	97.7	**97.9**	97.8	97.8	89.5	91.0	90.6	**91.6**	**94.5**	93.6	91.2	93.6
rus_Cyrl	93.8	94.4	94.5	**94.7**	98.3	98.5	**98.6**	**98.6**	92.6	93.4	93.6	**93.8**	88.0	86.9	85.6	**89.0**
slk_Latn	89.1	97.6	**98.1**	91.9	95.7	**96.1**	95.6	95.5	92.9	94.4	93.8	**95.0**	93.2	92.9	91.2	**93.3**
slv_Latn	96.7	97.6	98.1	**98.2**	98.5	**98.7**	98.6	**98.7**	93.4	94.7	94.8	**95.3**	93.4	93.1	93.6	**94.2**
srp_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	91.6	92.4	-	**93.4**
swe_Latn	96.5	**97.4**	**97.4**	97.3	97.3	**97.6**	97.1	97.0	89.4	**92.1**	90.8	91.7	94.3	**94.5**	93.5	94.4
tat_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	**89.7**	80.6	82.9	84.0
tur_Latn	90.4	91.0	**91.5**	91.4	91.1	91.3	**91.9**	91.4	70.9	73.0	73.6	**74.6**	92.2	92.0	90.8	**92.5**
ukr_Cyrl	93.1	94.7	72.9	**95.3**	87.0	**97.2**	87.0	97.0	89.4	91.8	61.3	**92.1**	92.0	91.7	77.5	**92.8**
vie_Latn	89.8	**92.1**	91.8	**92.1**	**99.9**	**99.9**	**99.9**	**99.9**	66.5	**70.3**	68.0	**70.3**	**91.9**	90.6	89.2	90.3
zho_Hans	96.2	**96.3**	96.0	96.0	**99.9**	**99.9**	**99.9**	**99.9**	86.1	**86.9**	84.6	85.6	0.1	**76.5**	75.5	74.5

Table 3.1: Results of monolingual MLMs trained on the HPLT 2.0 datasets compared to the baselines on POS tagging, lemmatization, dependency parsing and NER. For POS tagging, we evaluate the AllTags performance, which is the exact match accuracy of the UPOS, XPOS and UFeats UDtags. For dependency parsing, we report LAS, and for lemmatization accuracy.

[2] https://huggingface.co/collections/HPLT/hplt-20-bert-models

4 Encoder-decoder language models

We trained 57 language-specific monolingual encoder–decoder language models following the `T5-base` architecture (Raffel et al., 2020) on the HPLT 3.0 data. Despite the popularity of decoder-only LLMs in recent years, encoder–decoder models are still widely used in real-world applications, showing strength in both generative and discriminative tasks (Zhang et al., 2025).

Our encoder-decoder models serve two primary purposes:

1. to evaluate HPLT 3.0 quality as training data across a large number of languages; and
2. to provide a family of comparable monolingual encoder–decoders trained on current data.

The models follow the setup described in Samuel et al. (2023b). They are $\approx 275M$ parameters in size each.

4.1 Language choice

The choice of languages to train the models on was motivated by typological diversity. This means we aimed to cover as many different language families as possible. At the same time, we did not train models on extremely under-represented languages (less than $\approx$ 0.25M documents in our datasets). As a result, our set of 57 models covers the following language families: Indo-European, Sino-Tibetan, Japanese, Austronesian, Austro-Asiatic, Uralic, Altaic, Afro-Asiatic, Korean, Tai-Kadai, Dravidian, Kartvelian, Niger-Kongo, and Basque.

4.2 Evaluation design

We evaluate our encoder-decoder models on two tasks:

1. named entity recognition (NER) using the WikiAnn benchmark (Rahimi et al., 2019),
2. linguistic competence using the MultiBLiMP benchmark (Jumelet et al., 2025).

Their performance was compared to the performance of the original `mT5-base` model[1] and the language-specific BERT models[2] described in §3 (for NER only). For reference, we also demonstrate the performance of a much larger `mT5-xxl` model[3] on the MultiBLiMP benchmark.

MultiBLiMP is a dataset of minimal linguistic pairs: each sample consists of a pair of sentences. The first sentence is grammatical ("correct"). The second sentence is the same as the first one, but its syntactic head is modified in such a way, that the resulting sentence is ungrammatical ("incorrect"). We re-use MultiBLiMP's official evaluation code, adding support for encoder–decoder models. Regardless of the particular model architecture, evaluation works as follows: the probabilities of both "correct" and "incorrect" sentences are produced using the language model under evaluation. The resulting accuracy is the share of samples where the probability of the "correct" sentence is higher than that of the "incorrect" one. The difference between decoder-only and encoder–decoder models is that with the

[1] https://huggingface.co/google/mt5-base
[2] https://huggingface.co/collections/HPLT/hplt-20-bert-models
[3] https://huggingface.co/google/mt5-xxl

former, each token is only conditioned on the previous tokens and a sentence may be evaluated as-is, while with the latter, each token is conditioned both on the encoder's output and the previous tokens; thus, we replicate the T5 training procedure and mask the syntactic head in the encoder input, and the rest of tokens in the decoder input.

Since Norwegian is missing from MultiBLiMP, we used the grammatical error correction benchmark `NoCola` (Jentoft and Samuel, 2023) for Norwegian Bokmål. Unlike MultiBLiMP, it may contain more than one ungrammatical sequence (derived from real mistakes of Norwegian as a second language learners), manifesting not only in words, but also punctuation marks, e.g. a single comma. One might expect that a model trained on web data would perform poorly on such a dataset, and it turned out to be true: Bokmål is the only language for which both `mT5` models achieve higher accuracy than our models.

4.3 Evaluation results

Table 4.1 shows the evaluation results for the languages with the benchmarks available. It demonstrates that the HPLT 3.0 monolingual encoder–decoders offer a competitive alternative to the multilingual `mT5` models. On average, our models achieve the same performance as HPLT 2.0 BERTs on the NER task, while at the same time offering all the advantages of encoder–decoder models in comparison to encoder-only architectures; they also outperform both `mT5` models on the MultiBLiMP task.

We conducted additional evaluations on the English MultiBLiMP: The original monolingual `T5-base` performed better (93.5% accuracy) than the multilingual one; instruction-tuned derivatives `mt0-xl` (90.5% accuracy) and `aya-101` (86.6% accuracy) further supported the finding by Jumelet et al. (2025) that fine-tuning worsens a model's BLiMP performance; the most recent English T5-base model `t5gemma-b-b-ul2` showed surprisingly low result (66.5% accuracy); however, its training objective was more sophisticated (Tay et al., 2023) than in the original T5.

We make this novel family of HPLT encoder-decored models publicly available, including also intermediate checkpoints.[4]

[4] https://huggingface.co/collections/HPLT/hplt-30-t5-models

Language	Named entity recognition (WikiAnn, F1)				Linguistic competence (MultiBLiMP, accuracy)			
	size	BERT HPLT 2.0	T5 HPLT 3.0	mT5-base	size	T5 HPLT 3.0	mT5-base	mT5-xxl
Tosk Albanian (**als_Latn**)	100	93.9	93.2	86.7	243	95.5	90.5	88.9
Arabic (**ara_Arab**)	10000	-	91.7	80.8	1215	92.4	87.7	95.1
Asturian (**ast_Latn**)	1000	-	89.4	60.2	-	-	-	-
Belarusian (**bel_Cyrl**)	1000	92.8	91.5	86	2570	97.2	84.5	90.3
Bosnian (**bos_Latn**)	1000	-	94.2	88.4	3286	92.2	78.6	92
Bulgarian (**bul_Cyrl**)	10000	93	93.3	78.6	2458	93	87.7	91.6
Catalan (**cat_Latn**)	10000	94.5	92.7	87.4	2284	95.6	91.6	93.0
Czech (**ces_Latn**)	10000	91.8	91.6	85.2	4256	95.9	88.8	93.4
Mandarin Chinese (**cmn_Hans**)	10000	74.5	80.5	70.6	-	-	-	-
Welsh (**cym_Latn**)	1000	93.4	93.6	81.4	1120	89.3	78.1	86.1
Danish (**dan_Latn**)	10000	92	91.6	87.5	50	100	98	96
German (**deu_Latn**)	10000	89.2	88.6	83.4	2298	96	94	97
English (**eng_Latn**)	10000	82.7	82.1	77.6	770	94.2	90.6	95.3
Standard Estonian (**ekk_Latn**)	10000	93	92	81.1	2575	97.3	82.6	85.7
Modern Greek (1453-) (**ell_Grek**)	10000	92.6	92.5	86.1	1096	98.5	96.4	98.3
Basque (**eus_Latn**)	10000	92.9	92.0	82.8	273	97.4	94.9	96.0
Faroese (**fao_Latn**)	100	-	-	-	232	95.7	71.6	85.3
Finnish (**fin_Latn**)	10000	91.6	90.3	1.8	2570	95.6	81.4	86.1
French (**fra_Latn**)	10000	90.0	88.9	83.3	2548	93.6	91.7	94.8
Japanese (**jpn_Jpan**)	10000	67.2	73.6	54.3	-	-	-	-
Irish (**gle_Latn**)	1000	78.2	82.1	60.1	28	89.3	53.6	78.6
Galician (**glg_Latn**)	10000	94.1	93.4	89.2	753	96.0	90.7	95.4
Hebrew (**heb_Hebr**)	10000	89.3	88.9	77.1	2330	82.4	79.6	90.6
Croatian (**hrv_Latn**)	10000	92	91.4	86.8	3286	92.8	78.6	92
Hungarian (**hun_Latn**)	10000	93.1	91.9	84.9	845	99.1	92.8	95.9
Armenian (**hye_Armn**)	1000	95.2	96.2	89.5	1415	90.2	89.5	92.2
Indonesian (**ind_Latn**)	10000	92	92.4	85.9	-	-	-	-
Icelandic (**isl_Latn**)	1000	78.3	83.8	71	2801	94	87.3	91.1
Italian (**ita_Latn**)	10000	91.2	90.9	85.4	2999	93.9	88.5	94.7
Georgian (**kat_Geor**)	10000	90.7	90.4	80.4	204	96.6	93.6	90.7
Korean (**kor_Hang**)	10000	89.3	85.9	79.5	-	-	-	-
Northern Kurdish (**kmr_Latn**)	100	-	-	-	544	94.7	77	84
Lithuanian (**lit_Latn**)	10000	91	90	84.5	1180	98	92.2	87.7
Luxembourgish (**ltz_Latn**)	1000	89.2	88.6	4	-	-	-	-
Standard Latvian (**lvs_Latn**)	10000	93.9	92.9	86.1	3032	96.4	84	87.3
Macedonian (**mkd_Cyrl**)	1000	94.6	93.8	78.3	39	100	94.9	92.3
Dutch (**nld_Latn**)	10000	91	90.7	85.6	2331	92.1	89.3	94.1
Bokmål (**nob_Latn**)	10000	93.2	91.8	87.0	*3463	40.6	68.0	71.8
Nynorsk (**nno_Latn**)	1000	95.5	94.0	88.2	-	-	-	-
Polish (**pol_Latn**)	10000	89.6	89.6	87.8	3272	94.9	86.6	89.3
Portuguese (**por_Latn**)	10000	91.5	91.3	89.9	3048	93.5	92	95
Romanian (**ron_Latn**)	10000	93.6	93.6	86.4	2056	91.3	86.9	91.8
Russian (**rus_Cyrl**)	10000	89	88.2	82.9	3832	96.3	93	96.7
Slovak (**slk_Latn**)	10000	93.3	92.9	88.8	4145	92.8	80.2	86.6
Slovenian (**slv_Latn**)	10000	94.2	92.5	86.4	4483	92.6	83.6	90
Spanish (**spa_Latn**)	10000	90.8	90.7	84.0	2541	95.2	93.8	96.3
Serbian (**srp_Cyrl**)	10000	93.4	92.6	83.5	-	-	-	-
Swedish (**swe_Latn**)	10000	94.4	94.5	91.5	201	99.5	100	100
Swahili (**swh_Latn**)	1000	-	89.2	79.8	-	-	-	-
Tamil (**tam_Taml**)	1000	-	90	81	382	98.7	95.5	96.3
Thai (**tha_Thai**)	10000	-	80	32.1	-	-	-	-
Turkish (**tur_Latn**)	10000	92.5	92.3	87.9	1742	96.4	85.2	89.7
Ukrainian (**ukr_Cyrl**)	10000	92.8	92.5	82.1	2744	95.7	89.4	94.8
Vietnamese (**vie_Latn**)	10000	90.3	91.5	58.1	-	-	-	-
Average	-	**90.5**	**90.5**	78.8	-	**93.5**	86.8	91.4

Table 4.1: Evaluation results of HPLT 3.0 monolingual encoder–decoders (Bokmål and Nynorsk are two varieties of Norwegian), along with the test set sizes for each language. Average results are computed over all the 57 languages we have trained models for, where benchmarks are available. *Bokmål competence benchmarks are not part of MultiBLiMP.

5 Other LLM Evaluation Efforts

In addition to the previously described work that directly evaluated models trained on HPLT data releases, we also conducted research and organised an international shared task on multilingual and trustworthy evaluation of LLMs. We believe the findings from HPLT are important for our community. We report these efforts below.

5.1 Using Translation in Multilingual Evaluation

In this work, we consider the interlinked questions of whether multilingual LLMs can be instruction-tuned and/or evaluated using translated data.

In the training of LLMs, instruction tuning is an important step that turns the pre-trained or "base" model into a model that can follow instructions in the prompt rather than completing them as a sentence. Instruction tuning requires many thousands of instruction-response pairs, which are largely human-written, and for multilingual LLMs, we need instruction data in all supported languages. There are some freely available instruction-tuning data (e.g. Aya by Singh et al., 2024), but such data is time-consuming to produce. So a natural question to ask is whether translating English instruction data is a viable route to producing multilingual instruction data in an affordable way.

To answer this question, we compared models fine-tuned on native instruction data to models fine-tuned on instruction data translated from English. We found that using translated instruction data could be effective, but it depends on how the resulting model is evaluated. Evaluating the model on translated benchmarks can suggest that there is little difference between native and translated instruction data, but evaluating on natively created or text generation benchmarks shows a larger advantage for native instruction data. In further experiments, we investigated whether translation errors were the cause of the disparity, by comparing translated instruction data with native data that had been subjected to round-trip translation (via English, back to the original language of the native data). We found that the native data was still superior, indicating the quality of the translation was not the problem.

Full details of this work can be found in our published paper (Chen et al., 2024).

5.2 WMT25 Multilingual Instruction Task

We found that current multilingual benchmarks lack coverage (languages, tasks, etc), scientific practices, or consistent adoption across research labs, undermining their value in guiding multilingual LLM development. Among common problems, a few critical ones are: benchmark contamination, where training data inadvertently includes some part of test data (Ahuja et al., 2024); quality issues and noise in benchmarks (Chalamalasetti et al., 2025), as well as the over-reliance on translation, as we reported earlier.

Therefore, HPLT was involved in creating a brand new multilingual benchmark covering five important tasks across 30 languages: machine translation, linguistic reasoning, open-ended generation, cross-lingual summarization, and LLM-as-a-judge. We took care to ensure that the questions are language

and locale-specific and of high quality with human inspection.

We tested a wide range of open- and closed-weight systems, providing a multi-faceted evaluation framework that highlights the strengths and limitations of current LLMs across diverse linguistic phenomena. One highlight is that, in addition to automated metrics, we run principled human evaluations and LLM-human correlation tests inspired by the best practices proposed by the MT evaluation community.

All test sets, system outputs, and human judgments are released with a permissive license.[1] Full details about this effort can be seen in our published work (Kocmi et al., 2025).

5.3 A Dynamic Benchmark for Mathematics

The aim of this work is to address the problems of contamination and over-fitting in mathematical benchmarks. We address these by proposing a dynamic, counterfactual benchmark called MatheMagic. In this benchmark, we dynamically transform the interpretation of mathematical symbols and use these transformed symbols to generate new test instances that can challenge a model's mathematical reasoning. This process preserves the verifiability of mathematics benchmarks, but ensures that the model cannot obtain the correct answer through memorisation.

Our MatheMagic benchmark allows us to test both the models' deductive and inductive reasoning. We require the model to predict the results of mathematical expressions using the transformed mathematical rules and either provide a description of the new rules (deductive), or some examples of using the new rules (inductive). We can also provide both. As an example of a transformation, the `reverse_place_values` transformation reverses the digits both in the query and the response, so that the number `123` occurring in the query must be transformed into `321` before executing the calculation.

We test a variety of models with our benchmark, and show that whilst stronger models can perform reasonably well on the deductive version of the task, all models struggle with the inductive version of the task. We also show that some types of transformation, especially those that require multi-step reasoning, prove harder for models than others.

Full details of this work can be found in our paper currently under review (O'Brien et al., 2025).

[1] https://github.com/wmt-conference/wmt-mist

6 Appendix

6.1 HPLT-E: Details on Task Selection

Criterion	Pretraining Window	Description	Requirement
Monotonicity	All checkpoints (1B-30B)	Spearman correlation between step and performance score	≥ 0.5
Stable pretraining	Mid–late (15B–30B)	Trajectory-level coefficient of variation	≤ 15
Ranking consistency	Mid–late (15B–30B)	Kendall's τ correlation between rankings at consecutive pretraining intervals	No strict threshold
Prompt sensitivity	Mid–late (15B–30B)	Median absolute deviation across prompts	≤ 5
Prompt-switch rate	Mid–late (15B–30B)	Best-performing prompt consistency across checkpoints (prompt lottery)	No strict threshold
Signal-to-noise ratio (SNR)	Final checkpoint (30B)	Noise from prompt variation	≥ 3
Non-randomness	Final checkpoint (30B)	The absolute difference between the maximum score across final checkpoints and random baseline	Must be positive and satisfactory

Table 6.1: Pretraining evaluation signal criteria and requirements used in §2.4.1.

Criterion	Pretraining Window	Description	Requirement
Monotonicity	Mid–late (15B–80B)	Spearman correlation between step and performance score	≥ 0.5
Stable pretraining	Mid–late (15B–80B)	Trajectory-level coefficient of variation	≤ 15
Ranking consistency	Mid–late (15B–80B)	Kendall's τ correlation between rankings at consecutive pretraining intervals	No strict threshold
Prompt sensitivity	Mid–late (15B–80B)	Median absolute deviation across prompts	≤ 5
Prompt-switch rate	Late (40B–80B)	Best-performing prompt consistency across checkpoints (prompt lottery)	No strict threshold
Signal-to-noise ratio (SNR)	Final (80B–100B)	Noise from prompt variation	≥ 3
Non-randomness	Final (80B–100B)	The absolute difference between the maximum score across final checkpoints and random baseline	Must be positive and satisfactory

Table 6.2: Pretraining evaluation signal criteria and requirements used in §2.4.2 and §2.4.3.

6.2 HPLT-E: Details on Comparison of Deduplication Strategies

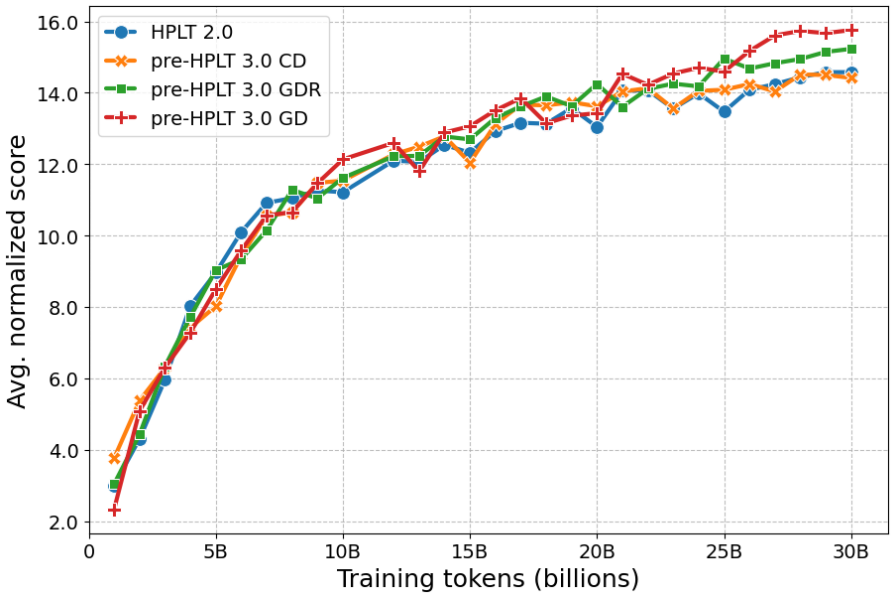


Figure 6.1: Catalan: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

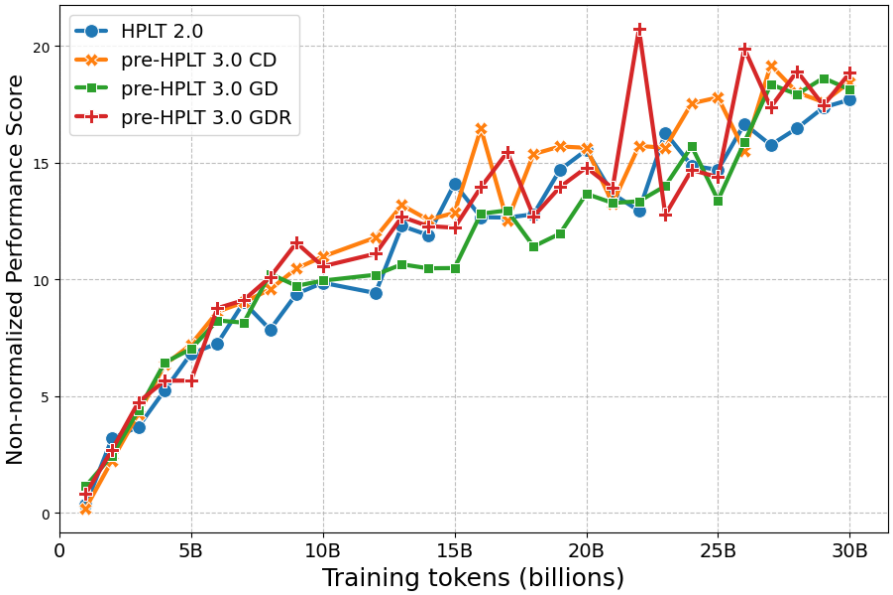


Figure 6.2: Czech: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

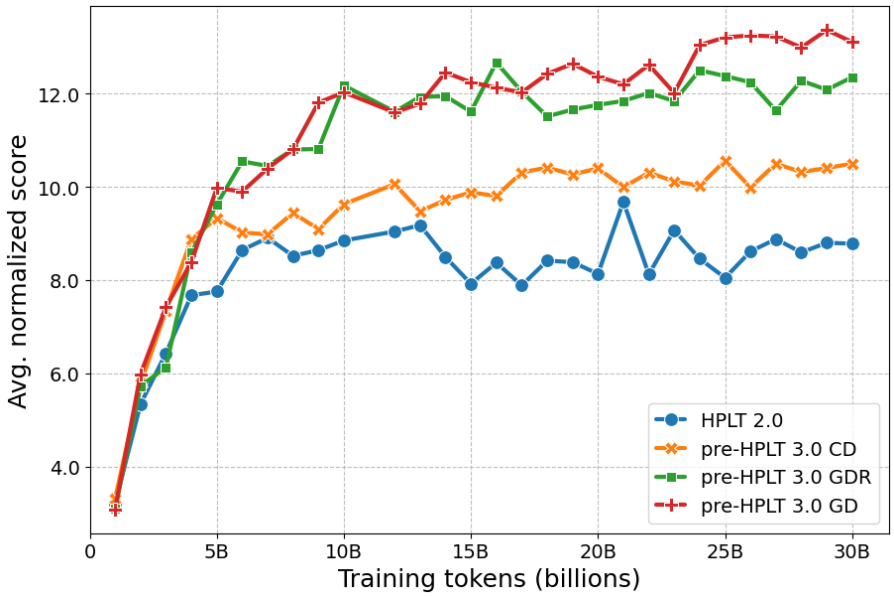


Figure 6.3: Basque: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

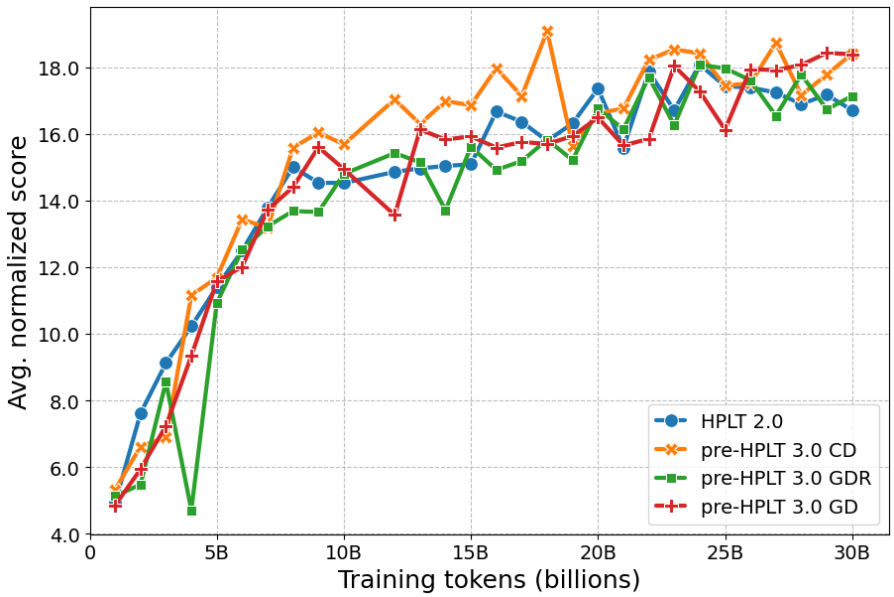


Figure 6.4: Finnish: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

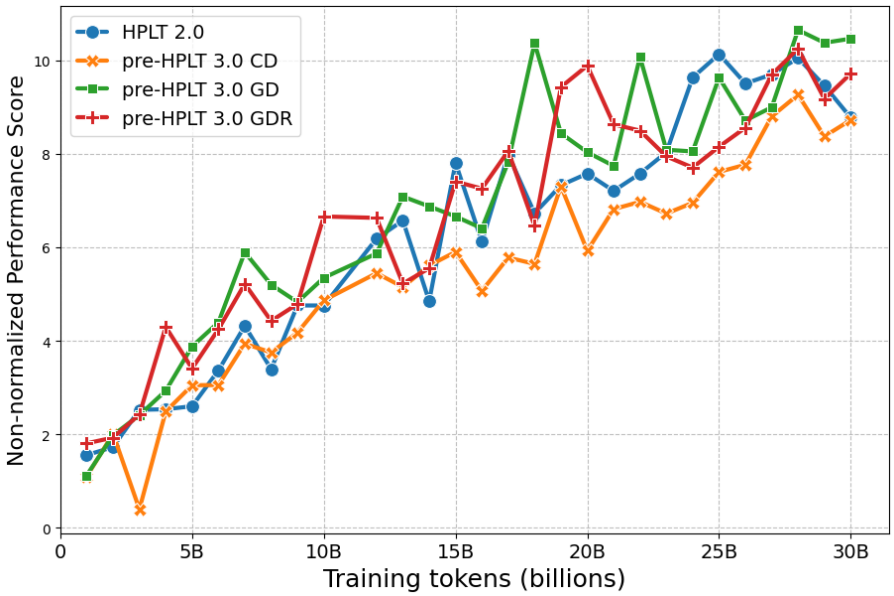


Figure 6.5: French: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

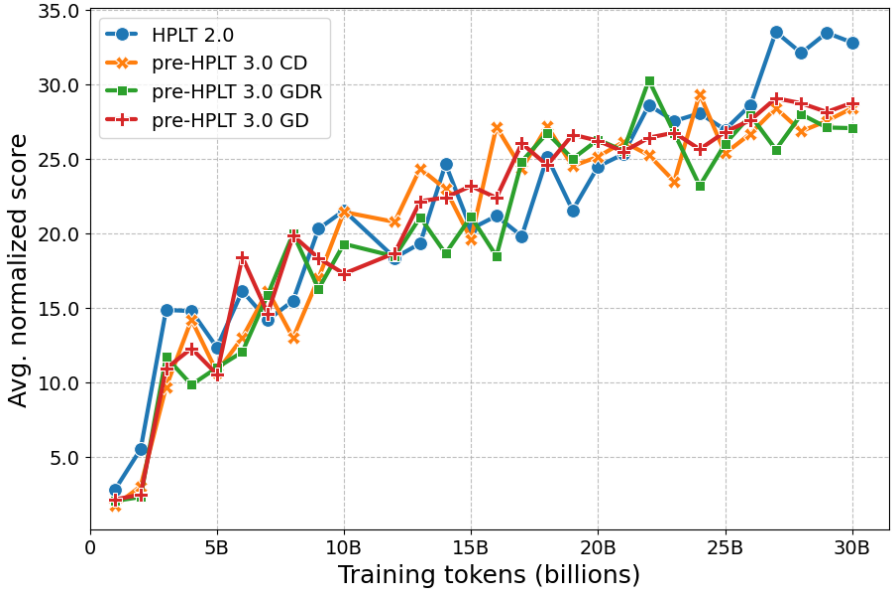


Figure 6.6: Norwegian: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

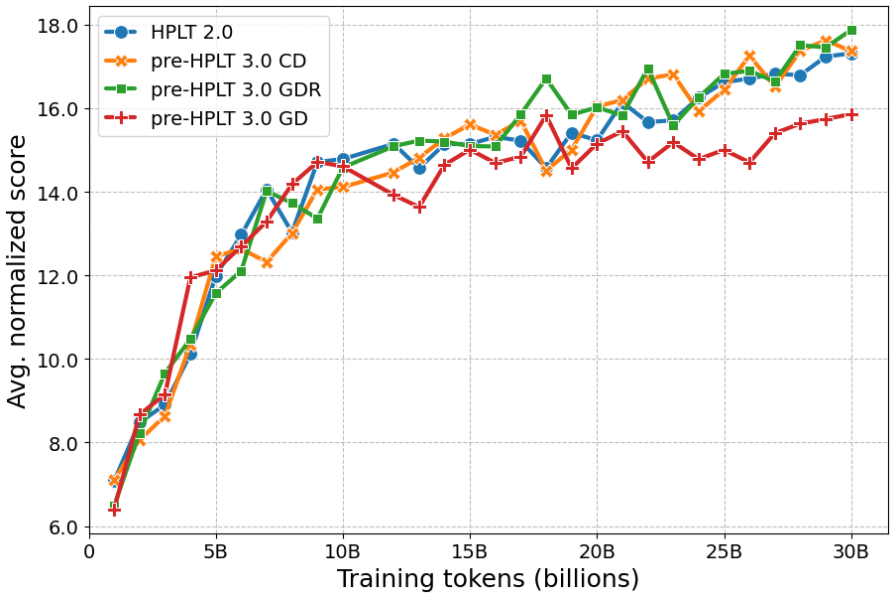


Figure 6.7: Spanish: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

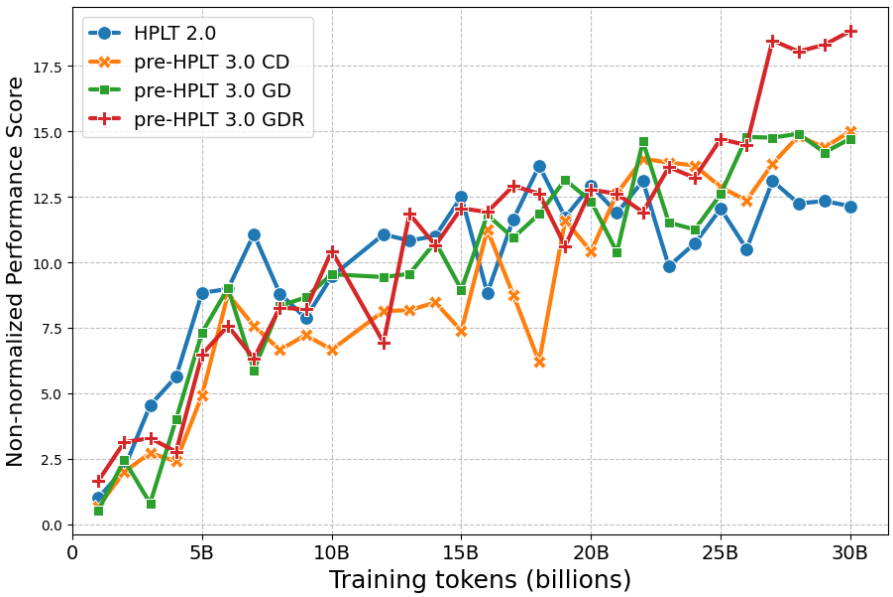


Figure 6.8: Ukrainian: Comparison of models pretrained on 30B tokens from HPLT 2.0, pre-HPLT 3.0 CD, pre-HPLT 3.0 GD, and pre-HPLT 3.0 GDR.

6.3 HPLT-E: Details on Comparison of Corpora

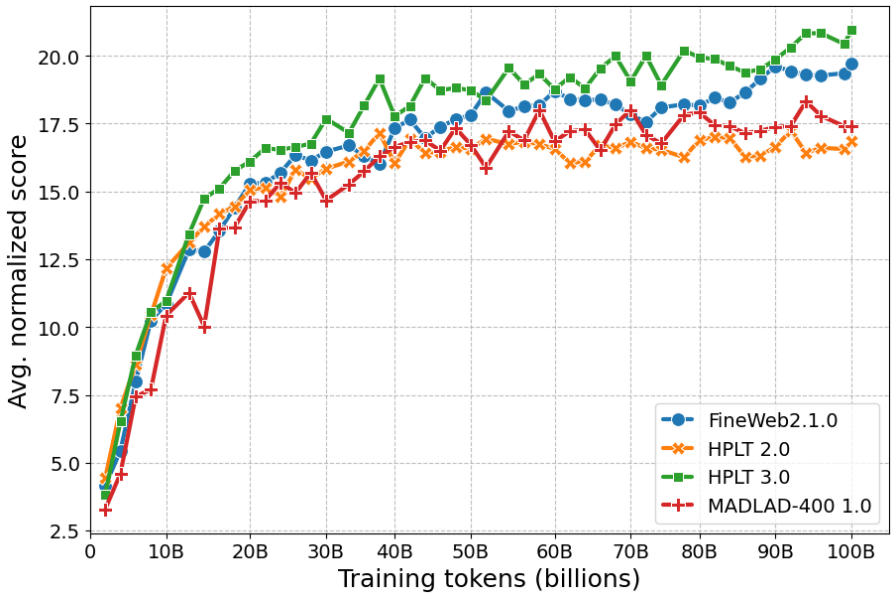


Figure 6.9: Catalan: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

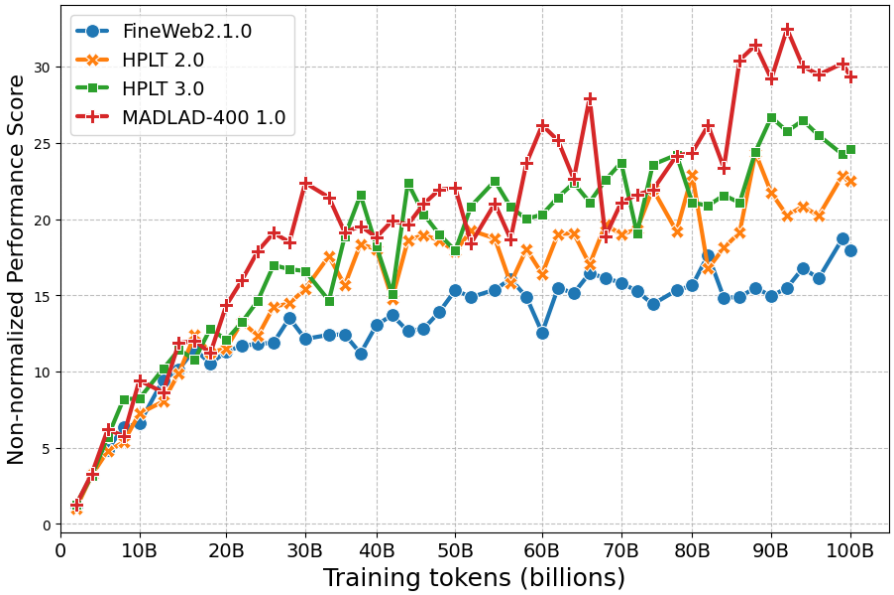


Figure 6.10: Czech: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

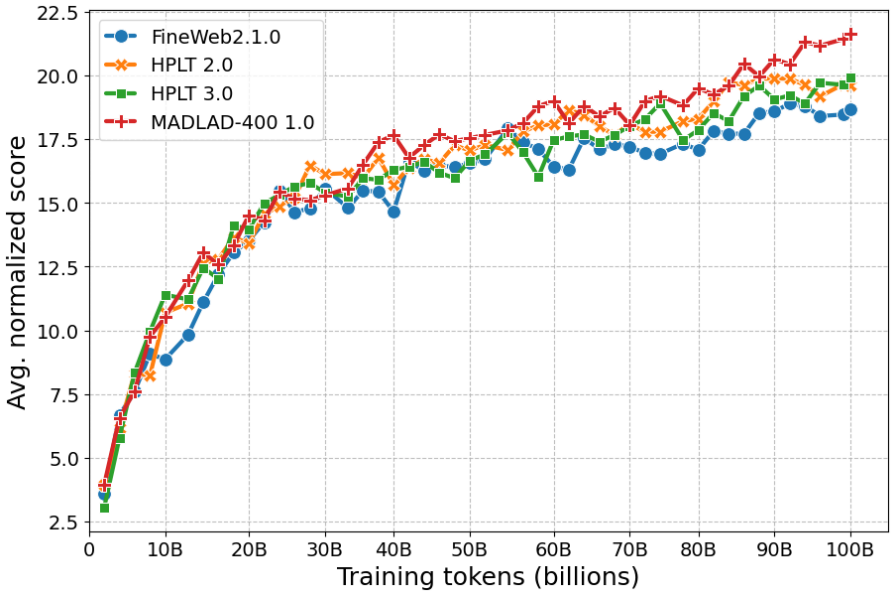


Figure 6.11: Finnish: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

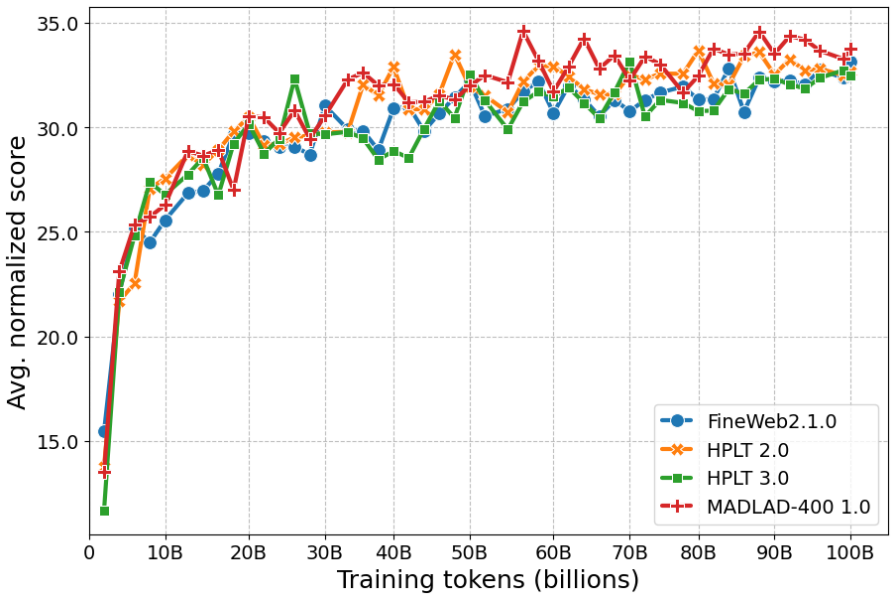


Figure 6.12: French: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

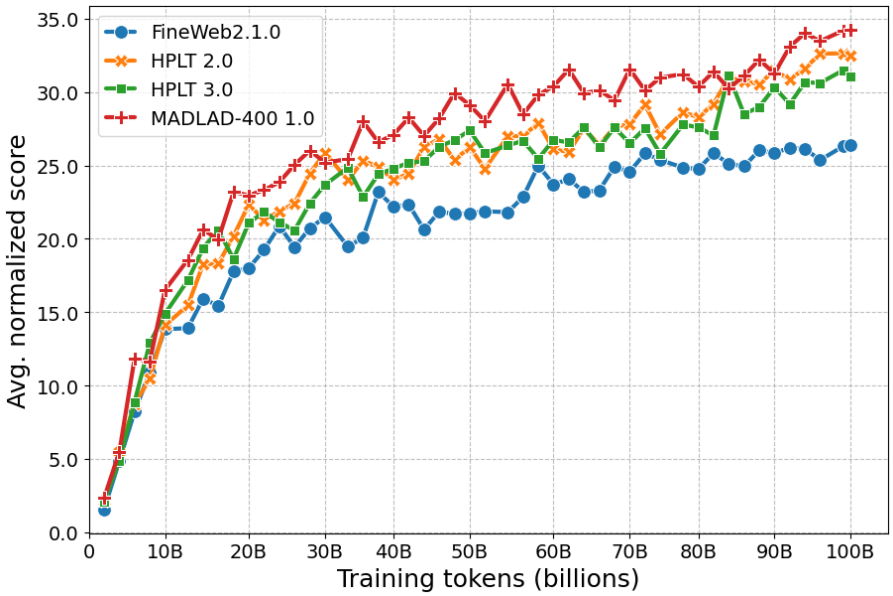


Figure 6.13: Norwegian: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

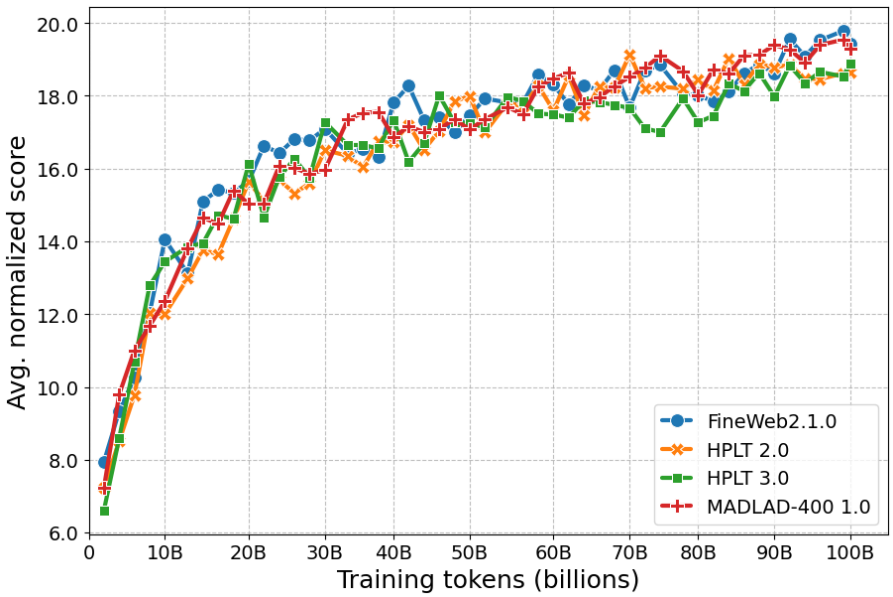


Figure 6.14: Spanish: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

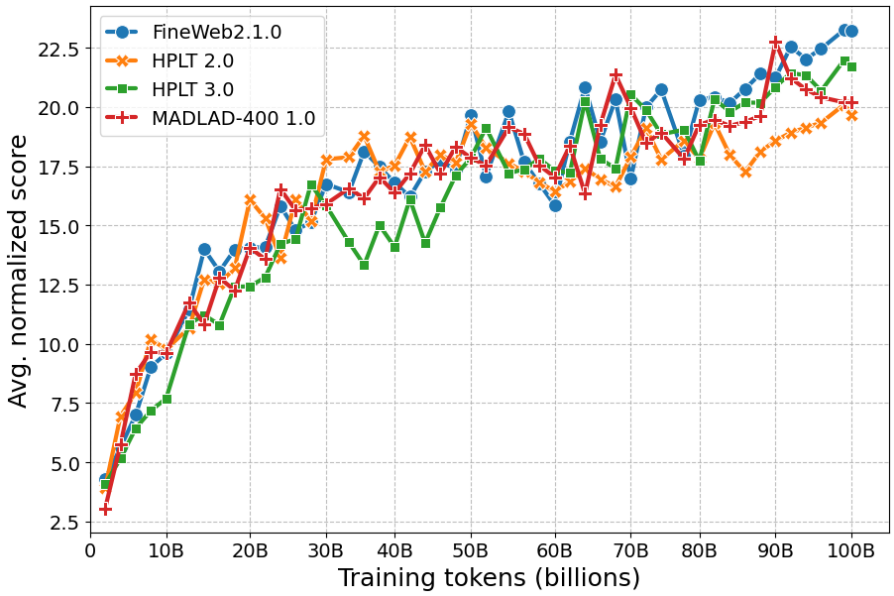


Figure 6.15: Ukrainian: Comparison of models pretrained on 100B tokens from HPLT 2.0, HPLT 3.0, FineWeb2.1.0, and MADLAD-400 1.0.

6.4 HPLT-E: Details on WDS-Based Sampling Analysis

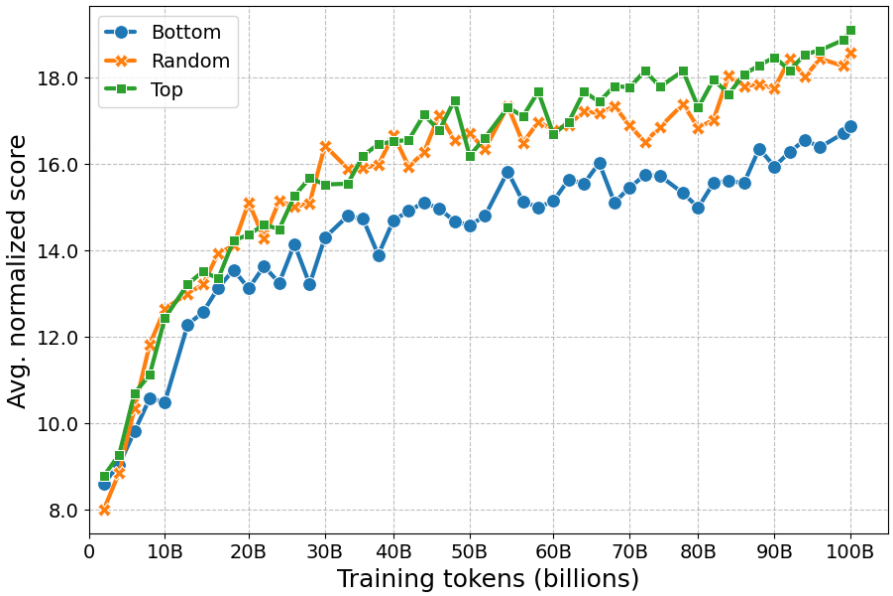


Figure 6.16: Spanish: Comparison of different WDS-based sampling strategies on 100B tokens from HPLT 3.0.

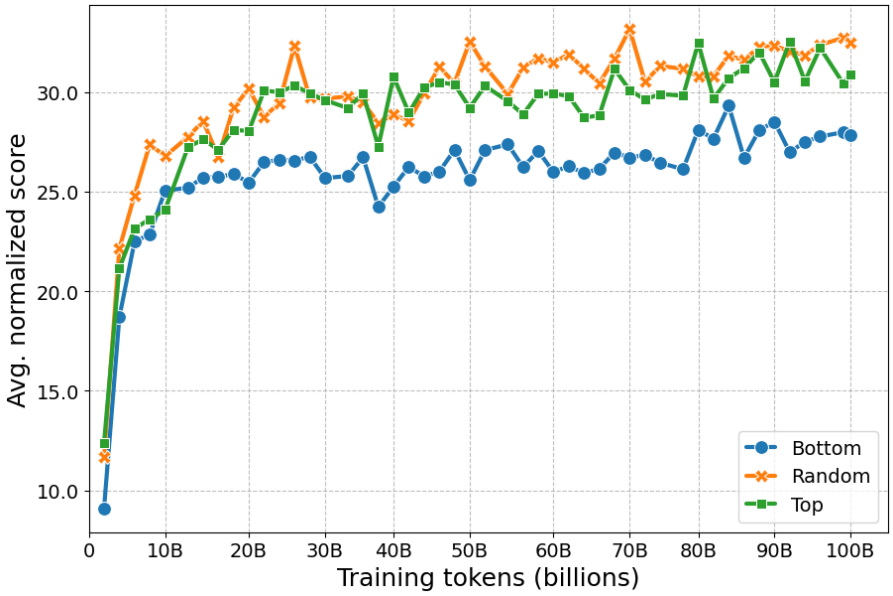


Figure 6.17: French: Comparison of different WDS-based sampling strategies on 100B tokens from HPLT 3.0.

Bibliography

Laurie Burchell, Ona De Gibert Bonet, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.854/.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025. URL https://arxiv.org/abs/2506.20920.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. MADLAD-400: A multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. IberoBench: A benchmark for LLM evaluation in Iberian languages. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.699/.

Manuel Faysse, Patrick Fernandes, Nuno M Guerreiro, António Loison, Duarte Miguel Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Henrique Martins, et al. CroissantLLM: A Truly Bilingual French-English Language Model. *Transactions on Machine Learning Research*, 2024.

Vladislav Mikhailov, Tita Enstad, David Samuel, Hans Christian Farsethås, Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. NorEval: A Norwegian language understanding and generation evaluation

benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3495–3541, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.181/.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, et al. BenCzechMark: A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism. *Transactions of the Association for Computational Linguistics*, 13:1068–1095, 2025.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. FinGPT: Large generative models for a small language. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.164. URL https://aclanthology.org/2023.emnlp-main.164/.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.919/.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. ua_datasets: a collection of ukrainian language datasets, October 2021. URL https://github.com/fido-ai/ua-datasets.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In Mariana Romanyshyn, Nataliia Romanyshyn, Andrii Hlybovets, and Oleksii Ignatenko, editors, *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.unlp-1.9/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. Toxicity classification in Ukrainian. In Yi-Ling Chung, Zeerak Talat, Debora Nozza, Flor Miriam Plaza-del Arco, Paul Röttger, Aida Mostafazadeh Davani, and Agostina Calabrese, editors, *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.woah-1.19. URL https://aclanthology.org/2024.woah-1.19/.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. URL https://aclanthology.org/2025.findings-acl.634/.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli,

Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. PromptSource: An integrated development environment and repository for natural language prompts. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.9. URL https://aclanthology.org/2022.acl-demo.9/.

Yulin Chen, Ning Ding, Xiaobin Wang, Shengding Hu, Haitao Zheng, Zhiyuan Liu, and Pengjun Xie. Exploring lottery prompts for pre-trained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15428–15444, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.860. URL https://aclanthology.org/2023.acl-long.860/.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2, 2024.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. What are the best systems? new perspectives on nlp benchmarking. *Advances in neural information processing systems*, 35: 26915–26932, 2022.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 670–686, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.48. URL https://aclanthology.org/2023.eacl-main.48/.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling Data-Constrained Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. Trained on 100 million words and still in shape: BERT meets British National Corpus. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.146. URL https://aclanthology.org/2023.findings-eacl.146/.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/coli_a_00402. URL https://aclanthology.org/2021.cl-2.11/.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://aclanthology.org/P17-1178/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In Daniel Zeman and Jan Hajič, editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2001. URL https://aclanthology.org/K18-2001/.

Hiroki Nakayama. seqeval: A python framework for sequence labeling evaluation, 2018. URL https://github.com/chakki-works/seqeval. Software available from https://github.com/chakki-works/seqeval.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation, 2025. URL https://arxiv.org/abs/2504.06225.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. NorBench – a benchmark for Norwegian language models. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational*

Linguistics (NoDaLiDa), pages 618–633, Tórshavn, Faroe Islands, May 2023b. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.61/.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015. URL https://aclanthology.org/P19-1015/.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs, 2025. URL https://arxiv.org/abs/2504.02768.

Matias Jentoft and David Samuel. NoCoLA: The Norwegian corpus of linguistic acceptability. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.60/.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6ruVLB727MC.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2402.06619.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.542. URL https://aclanthology.org/2024.emnlp-main.542/.

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. Contamination report for multilingual benchmarks. *arXiv preprint arXiv:2410.16186*, 2024.

Kranti Chalamalasetti, Gabriel Bernier-Colborne, Yvan Gauthier, and Sowmya Vajjala. Test set quality in multilingual llm evaluation. *arXiv preprint arXiv:2508.02635*, 2025.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.23. URL https://aclanthology.org/2025.wmt-1.23/.

Dayyán O'Brien, Barry Haddow, Emily Allaway, and Pinzhen Chen. Mathemagic: Generating dynamic mathematics benchmarks robust to memorization, 2025. URL https://arxiv.org/abs/2510.05962.