TildeOpen LLM: Leveraging Curriculum Learning to Achieve Equitable Language Representation

Toms Bergmanis*, **Martins Kronis***, **Ingus Jānis Pretkalniņš***, **Dāvis Nicmanis, Jeļizaveta Jelinska, Roberts Rozis, Rinalds Vīksna, Mārcis Pinnis**
Tilde, Latvia
toms.bergmanis@tilde.ai, marcis.pinnis@tilde.ai

Abstract

Large language models often underperform in many European languages due to the dominance of English and a few high-resource languages in training data. This paper presents TildeOpen LLM, a 30-billion-parameter open-weight foundational model trained for 34 European languages to promote linguistic equity and improve performance for low-resource languages. To address the data imbalance, we combine dataset upsampling with a curriculum-based training schedule that alternates between uniform and natural language distributions. The resulting model performs favorably compared to other multilingual LLMs despite being trained with significantly fewer computing resources. Evaluation across multiple multilingual benchmarks shows that TildeOpen surpasses existing open-weight models in text generation and comprehension, particularly for Baltic, Finno-Ugric, and Slavic languages. Human evaluations confirm an up to tenfold reduction in linguistic errors relative to leading baselines. The model and associated resources are fully open-weight and publicly available at HuggingFace. These outcomes demonstrate that careful data curation and balanced training strategies can substantially enhance multilingual model quality without increasing model size or training volume.

Keywords: large language model training, European languages, language equity

1. Introduction

Large language models (LLMs) are trained on increasingly vast amounts of Web data, reaching trillions of tokens of written text. However, most of this data is in English, resulting in a growing imbalance between English and other languages. As models scale, the relative share of non-English data continues to decline, which risks further eroding both linguistic and cultural diversity in LLM development. Consequently, simply adding more Web data is unlikely to improve model quality for most European languages in the future.

Recent evaluations of open-weight LLMs have confirmed this imbalance in practice, revealing a substantial performance gap between English and other European languages. Notably, models perform considerably better on languages predominantly used in Western Europe—such as those in the Romance and Germanic families—than on languages primarily used in Central and Eastern Europe, belonging to the Balto–Slavic family (Thellmann et al., 2024). Further evidence suggests that disparities persist even within the Balto–Slavic family itself, primarily reflecting differences in the amount and quality of training data available for each language (Kapočiūtė-Dzikienė et al., 2025). Moreover, the same report finds that when producing free-form text, mainstream models such as Llama 3 exhibit linguistic errors in approximately one out of every six words, underscoring the limitations of current multilingual modeling approaches.

Taken together, these findings illustrate persistent disparities in both data availability and model performance across languages. Existing publicly known methods have yet to adequately address these imbalances. As a result, many European languages remain insufficiently supported for practical or user-facing applications—particularly outside of commercial LLMs hosted by large corporations overseas. This situation raises important questions about Europe's AI sovereignty, as nearly 170 million Europeans have their first language only partially or poorly represented in existing foundation models.

While several initiatives have aimed to develop highly multilingual European LLMs, such as EuroLLM (Martins et al., 2025), Tueken (Ali et al., 2024), and Salamandra (Gonzalez-Agirre et al., 2025), they largely address the problem by limiting the share of English rather than rebalancing the representation of smaller languages. For example, EuroLLM—a model focused on European languages—still allocates roughly 50% of its training data to English, 27% to major Western European languages (German, French, Spanish, Italian, and Portuguese), 14% to high-resource global languages (Chinese, Russian, Japanese, Korean, Turkish, and Hindi), and only the remaining 9% to all other European languages combined. Consequently, the linguistic diversity of Europe remains underrepresented in even the most multilingual

* Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10634–10652
11–16 May 2026.

open-weight models available today.

In this work, we train a 30B-parameter foundational LLM supporting 34 European languages. To address the challenges outlined above, we build upon previous research on LLM training in data-constrained settings and propose new approaches of our own. Specifically, we upsample data for low-resource languages by up to 2.5 times. While this helps to mitigate data imbalance, the overall distribution remains skewed toward high-resource languages. Therefore, we propose a *curriculum learning* approach that alternates between sampling from *uniformly* and *more naturally* distributed language data. This strategy encourages balanced language exposure during the initial and final phases of training, while maximizing the diversity and quantity of available resources for high-resource languages during the intermediate phase.

Despite being trained on just 2 trillion tokens—between two and four-and-a-half times fewer than what similarly sized models typically use—our evaluation results show that our model surpasses similar sized models in text generation and comprehension tasks, while performing on par on tasks requiring parametric knowledge. These findings are further supported by human evaluations of linguistic quality, which reveal that our model produces up to more than 10 times fewer errors per 100 words than Gemma 2 for lower-resource languages.

2. Tokenizer

Our model supports 34 European languages. We group languages into two categories: 1) focus languages – languages for which we want to achieve equitable support in the language model, and 2) other supported languages. The focus languages are Bosnian, Bulgarian, Croatian, Czech, Estonian, Finnish, German, Latvian, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovak, Slovenian, and Ukrainian. Other languages include Albanian, Danish, Dutch, English, French, Hungarian, Icelandic, Irish, Italian, Latgalian, Maltese, Montenegrin, Norwegian, Portuguese, Spanish, Swedish, and Turkish. In addition to human languages, we also include programming languages from The Stack dataset (Kocetkov et al., 2022).

Aim and Motivation We design our tokenizer to achieve equitable language representation for the focus languages, ensuring that the same content translates into a similar number of tokens across languages. Multilingual tokenizers often encode low-resource languages far less efficiently than high-resource ones. This means that the same sentence may require several times as many tokens when written in a lower-resourced language. This inflates the cost of inference, reduces the effective duration of the context, and increases the training computation per unit of meaning for these languages.

Implementation We train the tokenizer on 4.38B bytes, comprising 3.78B bytes from the HPLT multilingual corpus (de Gibert et al., 2024; Arefyev et al., 2024), 400M bytes of code from The Stack dataset (Kocetkov et al., 2022), and 20M bytes of LaTeX from arXiv. In total, 3B bytes are allocated to focus languages, and 1.48B are allocated to other languages and programming languages.

We achieve tokenization equity for focus languages by iteratively adjusting the data proportions during tokenizer training and evaluating the resulting tokenization efficiency. We measure this tokenization equity using parallel translations from the FLORES 200 development dataset (NLLB Team, 2022), adjusting proportions until equivalent content yields similar token counts across languages. Tokenization equity for our focus languages is depicted in Figure 1. It compares tokenizers of various LLMs and shows that our tokenizer tokenizes text in any of our focus languages in a relatively equal number of tokens.

Due to their large dataset sizes, the tokenization efficiency of French, English, and Turkish impacts the downstream LLM's training efficiency. Therefore, we also include substantial amounts of data for these languages in the tokenizer training data, resulting in 250M bytes for French and 200M bytes each for the other. We also allocate 400M bytes to code data. Finally, we include the remaining languages, allocating 50M bytes to Spanish and 20M bytes each to the rest, to provide baseline support.

We implement the tokenizer using SentencePiece (Kudo and Richardson, 2018) with the Byte Pair Encoding (Gage, 1994) algorithm and set the vocabulary size to 131,072 tokens. During training, we use a character coverage of 0.99995, disable multiword tokens, split text at Unicode script changes, and tokenize numbers into separate digits. We enable byte-level fallback and do not add the beginning-of-sequence token during training, nor do we use text normalization.

3. Model and Training Details

Model Architecture Our model is a 30B parameter dense decoder-only transformer based on the Llama 3 architecture (Grattafiori et al., 2024). It has $n_{\text{layers}} = 60$ layers and a model dimension of $d_{\text{model}} = 6144$. We employ a variation of RMSNorm (Zhang and Sennrich, 2019) for layer normalization

We use Group Query Attention (Ainslie et al., 2023) for self-attention with Rotary Position Embeddings (RoPE) (Su et al., 2024) using $\theta = 200000$ for positional encoding. We set the attention head size

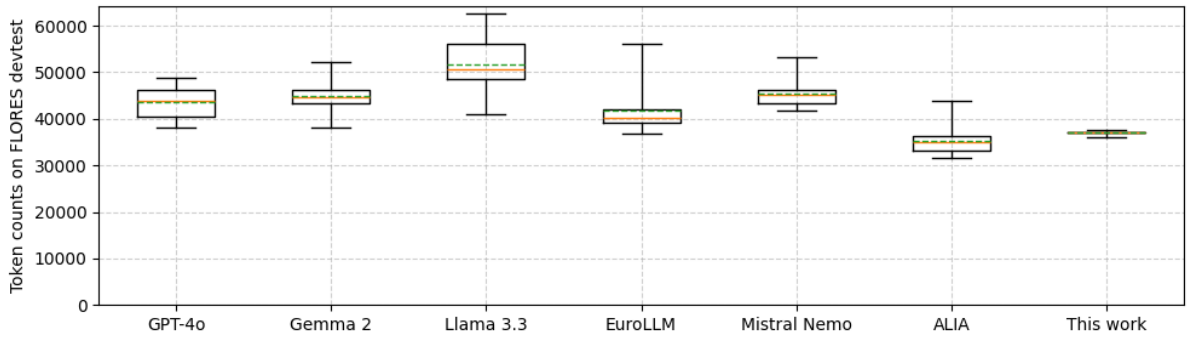


Figure 1: Comparison of tokenization efficiency of various LLMs - boxplot over all focus languages of [Anonymized] on the FLORES 200 devtest dataset (except Albanian, which is not included in the dataset).

to 128 and configure the model with 8 key-value heads and 48 query heads. We design the feed-forward layers to be mathematically equivalent to the $\text{FFN}_{\text{SwiGLU}}$ architecture described in Shazeer (2020). The intermediate layer size is set to 21,504. We do not add bias terms to any layers.

Training Hyperparameters We initialize all model weights using the SmallInit method from Nguyen and Salazar (2019) with an exception for FFN output layer weights, for which we follow Wang et al. (2024)

We train the model with a batch size of 576 samples, each containing 8,192 tokens, resulting in approximately 4.7M tokens per batch. We employ 8× tensor parallelism and 192× data parallelism during training. Adam optimizer (Kingma and Ba, 2017) is used, with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1 \cdot 10^{-8}$.

We follow Liu et al. (2024) and Martins et al. (2025) and use a trapezoidal learning rate scheduler with linear warmup to $1.8 \cdot 10^{-4}$ over the first 2,000 steps[1]. Followed by a constant learning rate phase and a cooldown phase. For the cooldown phase, we follow Hägele et al. (2024) (1-sqrt)-cooldown schedule and decrease the learning rate to $8 \cdot 10^{-6}$ and also increase the gradient clipping norm back to 1.0. We apply a weight decay of 0.1 throughout training and do not employ dropout.

We use the open-source LLM training framework GPT-NeoX (Black et al., 2022) on 768 AMD MI250x GPUs on the Large Unified Modern Infrastructure (LUMI) supercomputer. The training took approximately 1.5M GPUh.

[1]Due to loss spike frequency, we reduce the learning rate to $1.6 \cdot 10^{-4}$ and lower the gradient clipping threshold to 0.4.

4. Data

4.1. Monolingual Data and Filtering

The bulk of our data comes from large Web datasets MADLAD-400 (Kudugunta et al., 2023), HPLT 1 and 2 (de Gibert et al., 2024; Arefyev et al., 2024), Cultura-X (Nguyen et al., 2024), FineWeb 2 (Penedo et al., 2025) and the Common Pile (Kandpal et al., 2025). We also use specialist resources such as The Stack (Kocetkov et al., 2022), MathPile Commercial (Wang et al., 2024), Tezaurs (Klints et al., 2024).

Examining the raw dataset's 300 most frequent top-level Web domains for each language, we find low-quality sources that are either unsafe or unwanted content for LLM training or low-quality machine-translated materials. Thus, our data filtering process comprises document URL filters to remove certain sources entirely, followed by deduplication, which aims to eliminate repetitive content from the beginning and end of the document. We filter the remaining content using manually selected heuristics to remove noisy and low-quality text, as well as personally identifiable information.

URL Filtering We develop URL filtering criteria based on analysis of the most frequent top-level domains for each language. We remove data from sources with more than four subdomains, as these typically contain spam and low-quality text. We also filter out documents from blacklisted domains using both public and custom lists, as well as domains containing keywords related to pornography.

Deduplication We implement a two-step deduplication process to remove repetitive content in documents: (1) exact line removal and (2) similar line removal. Exact line removal removes ad text, references to articles, repetition between image caption and alt-text and boilerplate found at beginning and end of documents. The exact duplicate removal was done in a case-insensitive fashion,

and ignoring all non-alphanumeric characters except spaces. For similar line removal we use the ONe Instance ONly (Onion) tool [2]. Onion generates all paragraphs' n-grams and marks paragraphs as duplicates when the proportion of previously seen n-grams within a given paragraph exceeds a threshold. We configure Onion with the following parameters: an n-gram size of 5, a duplicate n-gram threshold of 0.5, and smoothing disabled. Whole document is marked as duplicate if duplicate paragraph to total paragraph ratio within a given document exceeds 0.5.

For all languages except English, French, and German, Onion was applied to the full dataset of each language. These three languages were exceptions due to the size of their respective datasets. For French and German we applied deduplication per data source, processing each source independently. For English, the dataset size precluded deduplication Onion, and we instead relied entirely on the deduplication schemes applied by the original dataset curators.

Heuristic and PII Filters We remove low-quality documents that meet any of the following criteria: punctuation character to all character ratio is less than 0.012 or is greater than 0.08; uppercase character to all character proportion is greater than 0.23; digit character to all character proportion is greater than 0.11; one letter word proportion to all word proportion (after removing all punctuation and digits) is greater than 0.22; stop-word to all word proportion (after removing all punctuation and digits) is less than 0.08; total word count is less than 50; or average word length is greater than 1.44 times the global average word length. We anonymize personally identifiable information by replacing emails, phone numbers, IBANs, credit card numbers, and language-specific personal identification numbers with synthetic equivalents from the Faker library[3].

Content Filtering The Russian state-controlled media spread propaganda that pollutes the public Web (Lesplingart and Châtelet, 2025) and consequently LLM-based systems (Dylan and Grossfeld, 2025) with falsehoods and misinformation. In the EU, many Web domains that are linked to spreading propaganda and (mis/dis)information are banned thanks to the efforts of organizations like the National Electronic Mass Media Council of Latvia[4]. We use URL blacklists targeting such Web domains. Examining the remaining data, we still found problematic Russian content with strong anti-Western pro-Russian sentiment and anti-Ukrainian propaganda. Therefore, we filter Russian language data by first constructing 200 clusters using Latent Dirichlet Allocation (Blei et al., 2003) and removing data belonging to clusters with keywords related to geopolitics, history, war, and LGBT. In total, this step removed about 5% of Russian documents.

	Unique	Ups.Ratio.	Total
LTG	0.01	2.34	0.03
GA	0.3	2.30	0.6
CNR	0.5	2.38	1.2
MT	0.5	2.16	1.1
IS	1.7	2.24	3.9
MK	3.6	2.33	8.4
SQ	6.7	2.29	15.3
SR	7.2	2.17	15.6
LV	9.8	2.35	22.9
NO	10.8	2.41	25.9
DA	14.2	1.84	26.1
BS	14.5	1.80	26.1
ET	15.4	1.70	26.1
SL	16.7	1.56	26.1
LT	18.0	1.45	26.1
SK	21.9	1.19	26.1
HR	22.9	1.14	26.1
RO	23.1	1.13	26.1
SV	26.1	–	26.1
UK	26.6	–	26.6
BG	26.9	–	26.9
HU	33.6	–	33.6
TR	35.6	–	35.6
FI	38.0	–	38.0
ES	40.6	–	40.6
NL	41.6	–	41.6
CS	44.6	–	44.6
PT	47.6	–	47.6
IT	47.8	–	47.8
Math	62.9	–	62.9
Parallel	71.0	–	71.0
RU	77.5	–	77.5
PL	99.1	–	99.1
FR	176.6	–	176.6
DE	176.6	–	176.6
Code	226.1	–	226.1
EN	397.4	–	397.4
Total	1,884		2,000

Table 1: Dataset statistics in billions of unique number of tokens and tokens after upsampling.

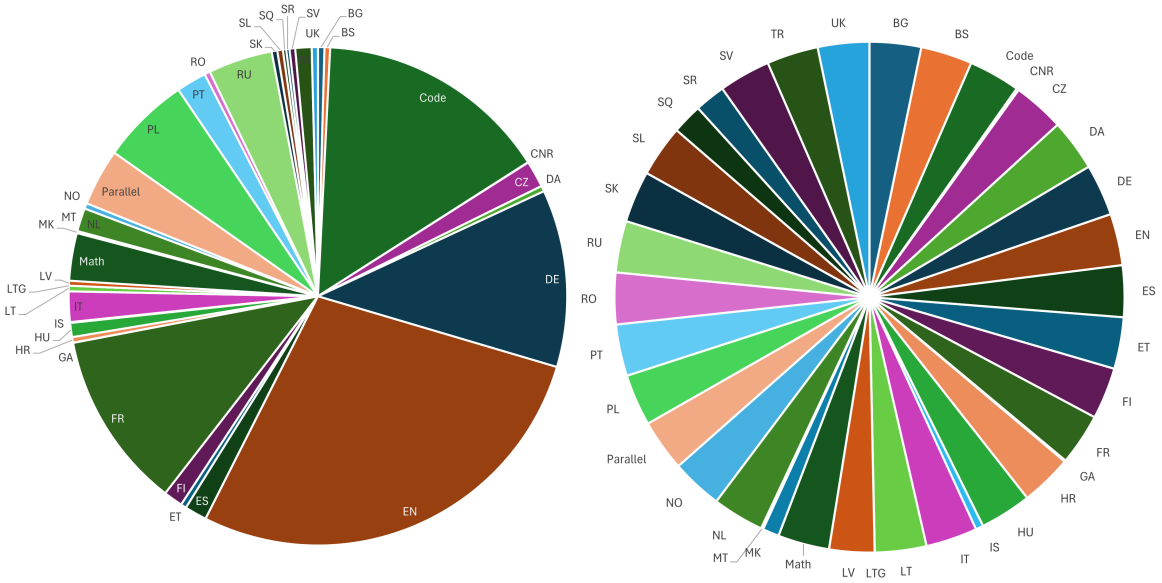
4.2. Parallel Data

We use parallel data from OPUS (Tiedemann, 2012). We use all parallel data where both languages are in the list of our supported languages. For filtering sentence-level data, we use

[2]Onion: https://corpus.tools/wiki/Onion

[3]Faker library: https://fakerjs.dev/

[4]Restricted access Websites by National Electronic Mass Media Council of Latvia: https://www.neplp.lv/en/restricting-access-websites



(a) Intermediate phase

(b) Initial and final phases

Figure 2: Data distribution during different phases of training.

the COMET quality estimation model[5] (Rei et al., 2022). As the scale of the scores assigned by the model varies drastically depending on the source and target languages, we calibrate the acceptability threshold to be 25% higher than the average score of the Flores Dev set (Goyal et al., 2022) sentence pairs for that language pair. To avoid using the same sentences repeatedly across different language pairs, we allow each sentence to be used once as the source and once as the target. We randomize the sentence pair selection but prioritize locally salient language pairs featuring languages of neighbouring countries and globally salient ones like those involving French, German, and English.

We use XML structure to create synthetic documents of varying lengths up to 8k tokens in total. Documents feature parallel sentences originating from the same dataset and belonging to the same language pair. To stimulate the model to learn long-range dependencies, we organize all source sentences in the first half of the XML document and their translation in the second half.

4.3. Data Sampling

Our objective is to achieve uniform model performance across all target languages. Ideally, this requires training data to be uniformly distributed across languages. The available data, however, is much different, with some languages having orders of magnitude more data than others (see second column of Table 1).

To mitigate this imbalance, we follow Muennighoff et al. (2023) and Luukkonen et al. (2023) and apply upsampling to under-represented languages up to 2.5 times.[6] Despite this intervention, significant disparities persist across the language distribution (see last column of Table 1). We therefore adopt a **curriculum learning approach with three phases**: uniform language exposure during the initial phase (7.5% of training, 150B tokens) and final phase (25%, 500B tokens (b) Figure 2), while allowing more *natural data distribution* during the intermediate phase (67.5%, 1.35T tokens (a) Figure 2).

5. Base Model Evaluation

We evaluate models using five benchmarks. **Multi-BLiMP** 1.0 (Jumelet et al., 2025) assesses the grammatical acceptability judgment. **Belebele** (Bandarkar et al., 2024) evaluates multilingual reading comprehension through contextual multiple-choice questions. For parametric knowledge and reasoning, we use Eurolingua translations (Thellmann et al., 2024) of **ARC** (Clark et al., 2018) and **MMLU** (Hendrycks et al., 2021). Since these translations retain American cultural context and exhibit translationese effects, we additionally include the **Exam** dataset (Hardalov et al., 2020), which contains national exams in their original languages. We use the LLM Evaluation Harness (Gao et al.,

[5]COMET-Kiwi: https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[6]Findings of Muennighoff et al. (2023) permit even higher upsampling factors, however, we leave some data aside for further model development activities, such as long-context ability extension.

2024) for all base model comparisons.

We use the **Borda** count method for result aggregation to compare models across benchmarks with different scoring scales. Models are ranked by performance on each dataset, with the top three receiving 3, 2, and 1 points, respectively. The average Borda score across all tasks provides a scale-independent measure of overall performance.

We compare our work with similar-sized highly multilingual models: EuroLLM-22b-preview[7] (**EuroLLM**), ALIA-40b[8] (**ALIA**) (Gonzalez-Agirre et al., 2025) and Gemma 2 27b (**Gemma 2**) (Team et al., 2024). We compare our model against these models because they are all trained from scratch instead of distilled from much larger models. This allows us to answer the question of to what extent our training methods and design choices were justified.

5.1. Results

Lang. Fam.	EuroLLM	ALIA	Gemma 2
N-Germanic	4.0%	2.7%	4.6%
W-Germanic	6.4%	1.5%	4.6%
Romance	11.2%	2.3%	5.6%
Baltic	7.8%	7.9%	13.8%
Slavic (Lat.)	7.7%	4.8%	8.6%
Slavic (Cyr.)	5.6%	3.4%	5.5%
Finno-Ugric	8.4%	5.9%	11.2%
Turkic	5.4%	5.9%	6.6%

Table 2: *Relative improvement* of our model's per-character perplexity compared to other open foundational LLMs on WMT24pp. Higher results are better. The Northern Germanic comparison excluded Icelandic (some models do not support it).

5-shot	EuroLLM	Gemma 2	ALIA	This work
MultiBLiMP	96.4%	95.7%	96.7%	**99.0%**
Belebele	82.5%	79.5%	76.8%	**84.7%**
ARCx	65.6%	**72.4%**	65.9%	65.3%
MMLUx	59.3%	**69.3%**	60.7%	59.9%
Exams	62.5%	**71.2%**	62.7%	66.6%
AVG Borda	0.8	**2.0**	1.4	1.8

Table 3: 5-shot performance and average Borda scores across benchmarks. Higher results are better.

[7]EuroLLM-22b-preview:https://huggingface.co/utter-project/EuroLLM-22B-Preview

[8]ALIA-40b: https://huggingface.co/BSC-LT/ALIA-40b

Intrinsic Evaluation Language models employ heterogeneous tokenization schemes, which preclude token-level perplexity comparisons across models. Per-character perplexity enables cross-model evaluation by computing token-level perplexity and converting it to character-level equivalents. Table 2 summarizes the relative improvements of our model with respect to other foundational LLMs in per-character perplexity per language family on WMT24pp (Deutsch et al., 2025).

Table 2 summarizes our model's relative per-character perplexity improvement over three strong multilingual baselines (EuroLLM, ALIA, and Gemma 2) on WMT24pp. Across all language families, our model consistently achieves lower perplexity, indicating more efficient text modeling. The improvements are particularly large for Baltic (+13.8%), Romance (+11.2%), and Finno-Ugric (+11.2%) languages, and remain substantial for Slavic (Latin script, +8.6%) languages. These results demonstrate that our model generalizes well across diverse European languages, surpassing prior multilingual LLMs in both high- and low-resource settings. We attribute this result to our data sampling strategy (see Section 4.3).

Benchmark Results Table 3 summarizes results over five standard benchmarks. Our model excels in tasks measuring language generation and understanding. For example, our model is better than other models on 20 out of 23 languages as measured by MultiBLiMP, resulting in the best result on average. We attribute this result to our data sampling strategy (see Section 4.3), uniformly presenting the model with different languages during training. Similarly, our model is best at answering multiple-choice questions based on text passages provided in context, as measured by Belebele. In this task, it marginally comes second to EuroLLM and Gemma 2 only on a few languages, namely the larger Western European languages (English for Gemma 2, German, French, and Spanish for EuroLLM), for which these models have seen more data in absolute and relative terms.

Curiously, we observe little difference among the three European models on tasks measuring their parametric knowledge—question answering without context (ARC, MMLU and Exams). The exception is Gemma 2, which stands out as the best model by a substantial margin. It's worth noting that the ARC and MMLU results for our model, which has been trained on the least amount of data, 2 trillion tokens, are on par with those of EuroLLM and ALIA, with 4 and 6.7 trillion tokens, respectively. This suggests that parametric knowledge is not improved by merely showing more data. In fact, when translations of American multiple-choice questions are replaced by national exams taken by pupils in

Model	ET	LV
EuroLLM	2445	2116
Gemma 2	2216	3957
This work	3168	2652

Table 4: Word count in Estonian (ET) and Latvian (LV) texts generated for linguistic error analysis.

European countries, as done in the Exams set, our model shows substantially better results than EuroLLM and ALIA. This either suggests that these models are somehow overfit to America-centric topics featured in ARC and MMLU or that they struggle with the more complex language of questions written in the original language. However, the superior results of Gemma 2 suggest that there are some critical differences between it and the other three models.

Overall, as the average Borda score indicates, our model comes second, bested by Gemma 2 due to its superior parametric knowledge, yet comes on top of the other two recent European LLMs of similar size.

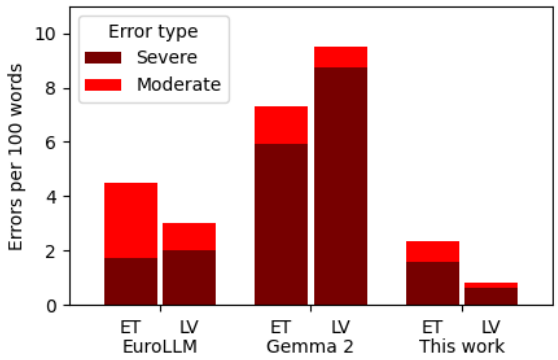


Figure 3: Error analysis results - averaged errors per 100 words over two language editors (professional linguists and native speakers).

Linguistic Error Analysis We conduct a small-scale linguistic error analysis task for Latvian and Estonian to better understand the linguistic quality of texts generated by our model and other open foundational models of comparable size—EuroLLM and Gemma 2. To that end, for each language we ask two annotators, professional editors and native speakers, to mark errors in texts generated by three foundational LLMs (see statistics in Table 4).

We instruct annotators to mark linguistic errors of two severity levels – severe and moderate. Severe errors are punctuation, grammar, lexical choice errors (inappropriate choice of words or made-up words, non-existing surface forms), and anything else that annotators deem gross linguistic mistakes. Moderate errors are capitalization errors, style errors, phrasing borrowed from English (or any other language), and anything else that annotators deem a moderate linguistic mistake.

We summarise results in Figure 3 by analyzing the average number of errors per 100 words. The results show that our model produces fewer than one mistake per 100 words for Latvian, compared to three mistakes for EuroLLM and almost 10 mistakes for Gemma 2. The error rate is slightly higher for Estonian, but still almost twice as low as for EuroLLM and over three times as low as for Gemma 2.

5.2. Training Data Memorization

Experimental setting To evaluate the degree to which the language model has memorized its training data, we test on the final slice of the training data. The test set consists of 965 documents that span all languages and data types present in the training corpus. Each document is split into two equal halves: the first half serves as the generation prompt and the second half as the reference. The model generates a continuation, which is then compared against the reference using three complementary metrics: 1) ChrF++ (Popović, 2017; Post, 2018), which measures word and character n-gram overlap; 2) edit distance, which quantifies the minimum number of word-level edits required to transform the generated text into the reference. A model with no memorization would be expected to produce continuations that are fluent and topically coherent but lexically distinct from the reference. Elevated ChrF++ or a low edit distance are therefore indicative of potential memorization.

Results Table 6 in Appendix A details memorization evaluation results per language and data type. On average, the model generates continuations that are lexically distinct from the references across all languages, suggesting that even when prompted with a long passage (i.e., 50% of the example) from the training data, the model is unlikely to reproduce the remainder verbatim. As expected, the highest similarity scores are observed for code and parallel translation documents. This is attributable to the structural properties of these data types: programming languages are more constrained than natural languages, producing more predictable and repetitive token sequences, whereas parallel documents contain the same content expressed in different languages, making partial overlap between prompt and reference more likely regardless of memorization.

While average ChrF++ scores indicate little overall memorization, the maximum values reveal that for some individual documents, the model generates continuations that are lexically very similar

to the reference. To examine these cases more closely, we select all documents with a ChrF++ score above 45, a threshold that, in machine translation evaluation, roughly corresponds to low-quality but intelligible output, indicating substantial lexical overlap. This yields 51 documents. The majority are parallel translation documents (23 out of 51) and code (13 out of 51), with the remainder spread across math (6), and a small number of individual documents from FR, EN, DE, CS, SK, BS, ET, and RO languages. The concentration in parallel and code data is expected, given their structural properties, as discussed above. The presence of math documents is similarly unsurprising given the highly formulaic and symbol-heavy nature of mathematical text.

Inspection of the flagged natural language documents reveals three distinct cases. The Estonian document is a long medical instruction text with medium lexical similarity (ChrF++ 47.0) that was included multiple times in the training data: three times in the Estonian subset, twice in the parallel data, and in translated variants across other languages. This repeated exposure is specific to the EuroLex and EMA subsets of our data. The BS, SK, DE, and RO documents are short texts of around 75 words, where the elevated scores are explained by topical similarity between the generated text and the reference rather than verbatim reproduction. Finally, the EN and FR documents originate from FineWeb, for which whole-dataset deduplication was not performed, resulting in some boilerplate text (e.g., HTML headers) being reproduced verbatim while the content portion, though drawn from the same domain, covers a different topic.

6. Instruction-Tuned Model Comparison

The goal of this section is to evaluate base models as a starting point for post-training by comparing the resulting downstream models. We use a multilingual translation task as a representative model task for downstream applications.

Direct comparison of already-tuned models would not provide meaningful base model comparison because of possible differences in the instruction-tuning data used to train the models. Unfortunately, the authors of most existing models have not publicly released their instruction-tuning datasets. Therefore, we construct our own training corpus encompassing translation directions between all supported languages and apply instruction-tuning to both our model and a selected baseline.

Dataset Our data sources are publicly available and include OpenSubtitles (Lison and Tiedemann, 2016) and TED Talks (Salesky et al., 2021) from OPUS, as well as newly recrawled EUROLEX (Baisa et al., 2016) parallel data. We use the document-level versions of these datasets and select consecutive segments (lines or paragraphs) such that the source and target texts each contain no more than 3k space-separated words.

Additionally, we collect monolingual news articles in all supported languages from HPLT-v2 and backtranslate them into English using the Gemma 2 27b IT model[9]. We apply COMET-based filtering (see Section 4.2) to the OpenSubtitles and TED datasets, retaining only those segments with average COMET scores exceeding the language-pair-specific thresholds.

To further improve data quality, we apply LangID filtering to exclude texts with low language confidence or identical source and target segments. For backtranslated data, we verify that the generated English texts preserve paragraph structure, numerals, and terminal punctuation. We allow each sentence or text to appear only once as a translation target across all language pairs.

We then sample a total of 2M documents, balancing the number of target-language tokens across all languages. Each language serves as the target approximately 3% of the time. We augment parallel data with simple English language instructions in the form `"Translate from {source language} to {target language}: {source text}\n{target text}"`.

The resulting dataset composition is as follows: 65% backtranslated news, 25% OpenSubtitles, 5% TED, and 5% EUROLEX. The dataset is approximately 462M tokens large. We publish this data on Hugging Face to support reproducibility of our experiments.

Instruction-tuning We instruction-tune our model for the translation task in a supervised fashion using the aforementioned dataset. We stick to the GPT-NeoX framework and use 768 AMD MI250x GPUs on the LUMI supercomputer. We train the model for 100 updates using our final cooldown LR value of $8 \cdot 10^{-6}$ and a batch size of approximately 4.7M tokens.

Baseline We compare against EuroLLM 22b, as its instruction-tuned version showed the best translation results for all language pairs when compared to other models in our preliminary experiments. For instruction-tuning, we employ the Hugging Face TRL framework[10], which enables direct

[9]Gemma 2 27b IT: https://huggingface.co/google/gemma-2-27b-it

[10]TRL: https://huggingface.co/docs/trl/index

reuse of the original Hugging Face checkpoints released by the model's authors. We adopt the post-training learning rate reported by the authors for their smaller models (Martins et al., 2025). For consistency across experiments, the baseline model is instruction-tuned for the same number of update steps (100) and uses the same relatively large batch size of 4.7M tokens as our model.

WMT24++	GPT-4.1	EuroLLM	This Work
EN-XX	0.865	0.834	**0.843**
DE-XX	0.847	0.831	**0.840**
FR-XX	0.839	0.823	**0.832**
PL-XX	0.838	0.826	**0.833**
UK-XX	0.833	0.820	**0.829**
LT-XX	0.830	0.813	**0.825**
Average	0.842	0.825	**0.834**

Table 5: Results of automatic translation quality evaluation as measured by COMET on the WMT24pp dataset. Results for GPT-4.1 added for reference.

Evaluation We use WMT24pp (Deutsch et al., 2025) to evaluate the resulting models' translation ability. WMT24pp is an English-centric multilingual dataset featuring the same English texts human translated into multiple other languages. We use this data to create non–English-centric translation pairs to evaluate the models' ability to translate between non-English languages as well. Specifically, we evaluate the models on their ability to translate out of English (EN–XX), German (DE–XX), French (FR–XX), Polish (PL–XX), Ukrainian (UK–XX), and Lithuanian (LT–XX). We select these source languages because they represent different levels of resource availability and speaker population size. We translate into languages supported by both models: Bulgarian, Czech, Danish, German, English, Estonian, Finnish, French, Croatian, Hungarian, Italian, Lithuanian, Latvian, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Swedish, Turkish, and Ukrainian, where appropriate, and report the average COMET[11] score for each model–source language combination. To provide context for the results, we also report scores for ChatGPT-4.1 obtained in July 2025.

6.1. Results

Table 5 summarizes the automatic translation quality evaluation results. The results indicate that our model's translation quality, as measured by COMET, surpasses EuroLLM's when both models are instruction-tuned for translation on the same dataset. These findings are consistent across all source language groups and individual language pairs.

We validate these results by comparing the EuroLLM version we instruction-tuned to the version published by the model's authors[12]. We find an average difference of 0.003 COMET points between the two versions. We attribute this to (a) differences in the instruction-tuning datasets and (b) differences in the instruction-tuning procedures. These results make us reasonably confident that our instruction-tuning of EuroLLM has been successful and performed optimally.

Next, we examine whether models' performance differences can be explained by the 8B parameter difference alone. To assess this, we compare our results with those of GPT-4.1 and note that, for most translation directions, except for EN–XX, our model's scores are much closer to GPT-4.1's (a model estimated to be around 60 times larger) than to EuroLLM's. We interpret this as *some* evidence that the observed differences are driven by the underlying training techniques rather than by mere parameter count alone.

7. Conclusions

We presented a 30B-parameter multilingual foundational LLM trained on 2T tokens covering 34 European languages. Our goal was to develop an LLM that handles languages equally. For this, we proposed an iterative rebalancing process for training data that allows training a tokenizer that guarantees language equity in language representation. We also devised a three phase curriculum for data sampling during training of the LLM that allows reaching equality for low-resource languages.

We showed that despite being trained on considerably less data than comparable open-weight models, our model exhibits strong performance in multilingual text generation and comprehension tasks, particularly for low-resource languages. Small-scale linguistic error analysis for two low-resource languages showed that the model generates text with up to 10 times fewer mistakes than comparable open weights models.

In future work, we plan to extend this model in several directions. First, we aim to increase its context length and evaluate its capabilities on document-level reasoning and translation tasks. Second, we intend to instruction-tune the model for various downstream tasks to understand whether language equity during pre-training helps in downstream fine-tuning.

[11] COMET: https://huggingface.co/Unbabel/wmt22-comet-da

[12] EuroLLM-22B-Instruct-Preview: https://huggingface.co/utter-project/EuroLLM-22B-Instruct-Preview

8. Acknowledgements

This work was supported as part of the Large AI Grand Challenge organized by the AI-BOOST project and funded by the European Union under Grant Agreement No. 101135737. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Part of this work was carried out within the scope of the Innovation Study Locally Deployable Enterprise Search and Q and A solution (1212), which has received funding through the FFplus project, funded by the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101163317. The JU receives support from the European Union's Horizon Europe Programme.

We acknowledge that this work also benefited from the EuroHPC JU computing award, granting access to LUMI HPC in Finland.

The project is co-financed by the Recovery and Resilience Facility within the framework of the activity 5.1.1.2.i "Support Instrument for Research and Internationalisation" under reform 5.1.1 "Innovation Governance and Motivation of Private R&D Investments" of investment direction 5.1 "Increasing Productivity through Increased Investment in R&D" of Latvia's Recovery and Resilience Plan.

9. Ethics Discussion

9.1. Data and Copyright Considerations

We carried out our work under the changing European regulatory framework. Some of our training data, such as the Common Pile and The Stack, come from public domain or permissively licensed sources. We also used large-scale Web datasets commonly used in the European LLM research community. While these datasets are standard for LLM training, they include Web-crawled and archived content that could be protected by EU copyright regulations.

Within the European Union, the use of copyrighted material for training AI models is governed by text and data mining (TDM) exceptions introduced in Articles 3 and 4 of the EU Copyright Directive (Directive 2019/790). Article 3 establishes a mandatory exception for TDM conducted by research organizations and cultural heritage institutions for scientific research. Article 4 extends this exception to TDM for any purpose, provided that rightsholders have not expressly reserved their rights through machine-readable means. The EU AI Act (Regulation 2024/1689), applicable to general-purpose AI models from August 2025, reinforces these provisions. Specifically, Article 53(1)(c) requires general-purpose AI model providers to implement policies to identify and comply with opt-out reservations expressed under Article 4(3) of the Copyright Directive. As we did not perform our own Web crawling and instead relied on established datasets compiled by reputable research institutions and organizations, we depend on these upstream providers to have conducted appropriate due diligence regarding rightsholder opt-outs during data collection.

The GPAI Code of Practice, published in July 2025, puts these rules into practice by requiring providers to use legal data sources, avoid known piracy sites, and establish mechanisms for rightsholders to file complaints. To meet these recommendations for Web-crawled data, we used URL filtering to remove content from blacklisted domains. This aligns with the Code of Practice's rule on excluding known copyright-infringing websites (Measure 1.2). We describe our data sources and filtering steps in detail, following the transparency requirements in Article 53(1)(d) of the AI Act.

The Code of Practice (Measure 1.4) says providers must use technical safeguards to stop models from generating outputs that copy copyrighted training content in a way that breaks the law. To reduce memorization risks, we followed advice from research on data-constrained language model training and limited each document to a maximum of 2.5 presentations during training. This is below the 3–4-repetition threshold, where memorization risk increases (Muennighoff et al., 2023). We also used several layers of deduplication beyond what the dataset curators had already done. We removed documents from the same Web addresses using URL-based deduplication and used the Onion tool for n-gram-based deduplication to find and remove paragraphs with high overlap. Our empirical evaluation of training data memorization (Section 5.2) confirms that the risk of verbatim training data reproduction is low.

9.2. Propaganda Filtering and the Russian-Language Data

Our data preparation pipeline includes filtering and selection steps to address Russian state propaganda. To the best of our knowledge, such an approach has not been used in LLM training before and therefore warrants broader discussion. We motivate this intervention by the well-documented large-scale Russian state-sponsored Web content manipulation. The Russian state maintains extensive infrastructure for producing and disseminating propaganda across the open Web and closed social media platforms in dozens of languages (VIGINUM, 2024; American Sunlight Project, 2025; VIGINUM, 2025; Lesplingart and Châtelet, 2025; Dylan and

Grossfeld, 2025).

The short summary of the issue is as follows: In February 2024, the French government agency VIGINUM exposed the Portal Kombat network, a coordinated system of over 190 websites disseminating pro-Russian propaganda across Europe in multiple languages (VIGINUM, 2024). Subsequent investigations found that this network, also known as the Pravda network, published over 3.6 million articles in 2024 alone, with the apparent primary objective of flooding LLM training data rather than reaching human readers, a tactic later termed as LLM grooming (American Sunlight Project, 2025; Sadeghi and Blachez, 2025). An audit of 10 leading AI chatbots found that they repeated false narratives from this network in approximately one-third of responses (Sadeghi and Blachez, 2025).

We therefore believe that extra-special filtering of Russian-language data is warranted and justified.

URL-based domain filtering URL filtering is applied across all languages in our corpus. It removes content from domains that host unsafe, low-quality, or prohibited content. This includes domains appearing on public and custom blacklists targeting known sources of disinformation, fabricated news, and other harmful content, regardless of language. In this work, it also includes domains sanctioned under EU law, including the Pravda network. Specifically, since March 2022, the Council of the European Union has prohibited the broadcast and distribution of content from several Russian state-controlled media outlets within the EU (Council Regulation (EU) 2022/350;[13] Council Regulation (EU) 2022/879), a measure subsequently upheld by the General Court of the European Union.[14]. Additional domains have been restricted by national regulatory authorities acting within this EU framework. For instance, the National Electronic Mass Media Council of Latvia[15] has identified and restricted access to domains linked to Russian state-controlled media; these decisions are binding across the European Union. Compliance with these restrictions is a legal obligation and not a research decision or an editorial choice.

Topic-level Filtering of Russian data. Most of the Russian-language web content comes from Russia. Russia's domestic legal environment, however, systematically suppresses content that contradicts official state narratives. For example, public opposition to the military campaign in Ukraine is a criminal offense under amendments to Russia's Criminal Code adopted in March 2022, carrying sentences of up to 15 years.[16] Similarly, Russia's 2023 expansion of its restrictions on so-called LGBT propaganda effectively criminalizes any public expression that does not align with the state's position on LGBT rights[17]. In the EU, on the other hand, sexual orientation is one of the attributes protected by EU anti-discrimination laws.

The consequence for Web-crawled data is that Russian-language content on these topics is disproportionately likely to reflect a single, state-endorsed viewpoint that arises not from organic societal consensus but from legal suppression of the alternatives. Including such data without intervention risks encoding systematically one-sided content into the model's learned representations, including anti-LGBT hate speech and content justifying military aggression against a sovereign nation. In contrast, content in other languages in our corpus is produced in media environments where a plurality of viewpoints on these topics can be and are expressed.

10. Limitations

Underrepresented Languages Our work excludes a range of Europe's regional and minority languages, such as Catalan, Galician, Welsh, and Basque. This exclusion primarily results from the limited availability of high-quality written data for these low-resource languages, constraining both model performance and reliability. Other languages—including Greek, Armenian, and Georgian—were omitted due to the additional complexity of incorporating disjoint scripts used exclusively by these languages.

Design Considerations We employed substantial data upsampling to promote language parity among the low-resource languages. This data augmentation followed best practices established in prior research but introduced several trade-offs. Upsampling inherently reduces training efficiency, as the same tokens are repeatedly presented to

[13] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0350

[14] Case T-125/22, *RT France v. Council of the European Union*, judgment of 27 July 2022.

[15] https://www.neplp.lv/en/restricting-access-websites

[16] Russian Federal Law No. 32-FZ of 4 March 2022, amending Article 207.3 of the Criminal Code of the Russian Federation.

[17] Russian Federal Law No. 478-FZ of 5 December 2022, amending Federal Law No. 149-FZ On Information, Informational Technologies and the Protection of Information: http://en.kremlin.ru/acts/news/70004, https://www.wipo.int/wipolex/en/legislation/details/23335

the model—effectively expending compute without proportionate informational gain.

Given the substantial computational cost of training models at this scale, conducting extensive scaling experiments to identify optimal data-related hyperparameters was infeasible. Consequently, many design decisions were guided by established findings in the literature rather than empirical exploration. Likewise, the same computational cost precluded meaningful ablation studies.

Data Contamination and Safety Despite rigorous filtering efforts—including automated screening and extensive manual review—some degree of data contamination is likely to persist. As with many foundational LLMs, our model has not undergone human preference alignment, increasing the risk of producing unsafe or biased outputs. Consequently, model generations may occasionally reflect harmful content, stereotypes, or political biases (e.g., Russian propaganda). No formal safety or bias evaluation has yet been performed to assess the prevalence or potential impact of such outputs.

Russian-language data By applying especially heavy and strict filtering to Russian data, due to the previously explained propaganda concerns, we risk missing societal and cultural nuances that may be present in filtered domains, such as forums, as a form of public discourse.

10.1. Bias Evaluations

We have not performed systematic safety checks, toxicity tests, political bias reviews, or cross-lingual fairness assessments. This is an important gap, but the absence of culturally and linguistically relevant benchmarks makes it unfeasible for a multilingual foundational model of this size. We have not used English-language datasets based on American culture—as Europeans are not Americans.

Most established bias and fairness benchmarks for language models are only available in English. CrowS-Pairs (Nangia et al., 2020), a common benchmark for measuring social stereotypes, is only in English, with some versions in French (Névéol et al., 2022) and Dutch (Pagliai et al., 2025). StereoSet (Nadeem et al., 2021), which looks at stereotypes about gender, jobs, race, and religion, is also only in English. The Bias Benchmark for Question Answering (BBQ) (Parrish et al., 2022), which has over 58,000 items in nine social bias categories, has been extended to Dutch, Spanish, and Turkish through MBBQ (Weber et al., 2024) and to Japanese through JBBQ (Yanaka et al., 2025). BOLD (Dhamala et al., 2021), RealToxicityPrompts (Gehman et al., 2020), ToxiGen (Hartvigsen et al., 2022), and TruthfulQA (Lin et al., 2022) are all only in English. DecodingTrust (Wang et al., 2023), the most thorough trustworthiness evaluation so far, was designed for and tested only on English. The Political Compass Test has been used in several languages in research (Röttger et al., 2024; Helwe et al., 2025), but there is no standard, downloadable political bias benchmark for the main languages in our study.

Importantly, none of these benchmarks include Baltic languages like Latvian and Lithuanian, most Finno-Ugric languages such as Estonian, Finnish, and Hungarian, or most of the Slavic languages in our model. The few multilingual bias resources that exist, like SeeGULL (Bhutani et al., 2024), which covers 20 languages, and FLAMES (Huang et al., 2024), a Chinese-language value alignment benchmark, also do not include our target low-resource European languages.

Simply translating English-based benchmarks by machine is not a reliable method. Bias is shaped by culture, and stereotypes, political issues, and ideas about toxicity can vary widely across languages and cultures (Röttger et al., 2024; Naous et al., 2025). For example, a CrowS-Pairs item about racial stereotypes in the U.S. would not give a valid measure of bias in Latvian or Estonian, because neither Latvians nor Estonians are Americans.

Building culturally relevant bias benchmarks for 34 European languages is a major research task, since it needs experts who understand local stereotypes, politics, and social issues.

11. References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, et al. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *arXiv preprint arXiv:2410.03730*.

American Sunlight Project. 2025. The pravda network: Russian propaganda may be flooding AI models. Technical report, American Sunlight Project.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding,

Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Huw Dylan and Elena Grossfeld. 2025. Revisionist future: Russia's assault on large language models, the distortion of collective memory, and the politics of eternity. *Dialogues on Digital Society*, page 29768640251377941.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

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. 2024. The language model evaluation harness.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, et al. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li,

Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. 2024. Scaling laws and compute-optimal training beyond fixed training durations.

Jurgita Kapočiūtė-Dzikienė, Toms Bergmanis, and

Mārcis Pinnis. 2025. Localizing AI: evaluating open-weight language models for languages of Baltic states. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 287–295, Tallinn, Estonia. University of Tartu Library.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Amaury Lesplingart and Valentin Châtelet. 2025. Russia-linked pravda network cited on wikipedia, llms, and x. *Digital Forensic Research Lab (DFRLab)*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Manuel Faysse, et al. 2025. Eurollm-9b: Technical report. *arXiv e-prints*, pages arXiv–2506.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

McKenzie Sadeghi and Isis Blachez. 2025. Russian disinformation network floods AI chatbots. Technical report, NewsGuard.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

VIGINUM. 2024. Portal kombat: A structured and coordinated pro-russian propaganda network. Technical report, Secrétariat Général de la Défense et de la Sécurité Nationale.

VIGINUM. 2025. War in Ukraine: Three years of Russian information operations. Technical report, Secrétariat Général de la Défense et de la Sécurité Nationale.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2024. Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6761–6774.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32*, volume 32, pages 12360–12371. Curran Associates Inc. 33rd Conference on Neural Information Processing Systems, NeurIPS 2019 ; Conference date: 08-12-2019 Through 14-12-2019.

12. Language Resource References

Nikolay Arefyev, Mikko Aulamo, Pinzhen Chen, Ona De Gibert Bonet, Barry Haddow, Jindřich Helcl, Bhavitvya Malik, Gema Ramírez-Sánchez, Pavel Stepachev, Jörg Tiedemann, et al. 2024. Hplt's first release of data and models. In *The 25th Annual Conference of The European Association for Machine Translation*, pages 53–54. European Association for Machine Translation (EAMT).

Vít Baisa, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.

Nishtha Bhutani et al. 2024. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. 2024. A new massive multilingual dataset for high-performance language technologies. In *LREC/COLING*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, et al. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *CoRR*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 862–872.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5427–5444.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Chadi Helwe et al. 2025. Political Compass Test Evaluation Spanning 50 Countries.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuxia Huang et al. 2024. FLAMES: Benchmarking Value Alignment of LLMs in Chinese.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv e-prints*, pages arXiv–2504.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben, Elie Bakouch, John David, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, John Kirchenbauer, Tom Goldstein, Brian R, Bhavya Kailkhura, and Tyler Murray. 2025. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. *arXiv preprint*.

Agute Klints, Mikus Grasmanis, Gunta Nšpore-Bērzkalne, Lauma Pretkalniņa, Madara Stāde, Normunds Grūzītis, Ilze Lokmane, Pēteris Paikens, Laura Rituma, and Andrejs Spektors. 2024. Tēzaurs as a digital multifunctional lexical resource. *Baltic Journal of Modern Computing*, 12(4).

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code. *Preprint*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Tarek Naous, Anagha Savit, Carlos Rafael Catalan, et al. 2025. Camellia: Benchmarking Cultural Biases in LLMs for Asian Languages. ArXiv preprint arXiv:2510.05291.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extending a Challenge Dataset for Measuring Social Bias in Masked Language Models to a Language Other Than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Irene Pagliai, Goya van Boven, Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Isabella Söder-

gren, and Elisa Barney. 2025. Dutch CrowS-Pairs: Adapting a Challenge Dataset for Measuring Social Biases in Language Models for Dutch. In *Proceedings of RANLP*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv e-prints*, pages arXiv–2506.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech 2021*, pages 3655–3659.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards cross-lingual llm evaluation for european languages.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024. Mathpile: A billion-token-scale pretraining corpus for math. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rickard Weber, Michael Lepori, et al. 2024. MBBQ: A Dataset for Cross-Lingual Comparison of Stereotypes in Generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. 2025. JBBQ: Japanese Bias Benchmark for Analyzing Social Biases in Large Language Models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP) at ACL 2025*.

Appendix A: Memorization Evaluation Results

Lang.	Docs	Avg. tok.	ChrF++ Avg.	ChrF++ Max.	Edit dist.
BG	31	650	21.9	35.3	0.94
BS	32	423	23.0	42.9	0.94
CNR	2	2094	17.5	22.6	0.93
CODE	31	2469	43.6	86.7	0.74
CS	31	512	20.7	54.1	0.94
DA	31	603	25.7	35.1	0.94
DE	31	805	24.6	52.7	0.93
EN	35	872	24.8	77.7	0.90
ES	31	618	23.1	35.7	0.93
ET	31	879	22.0	44.7	0.96
FI	31	772	23.8	36.7	0.95
FR	31	645	27.6	79.5	0.90
GA	2	2231	18.9	19.4	0.96
HR	31	652	23.7	35.5	0.94
HU	31	1008	20.6	34.0	0.94
IS	4	1379	20.2	21.1	0.94
IT	32	731	23.9	35.7	0.96
LT	31	715	21.1	35.2	0.96
LTG	2	2142	17.8	20.1	0.99
LV	28	644	24.0	41.5	0.96
MATH	34	2514	32.8	66.4	0.88
MK	10	691	25.0	39.7	0.94
MT	2	2082	25.3	27.6	0.95
NL	32	438	25.0	37.6	0.93
NO	31	667	21.2	31.3	0.94
PAR.	31	1085	54.3	87.0	0.66
PL	31	597	19.7	29.2	0.95
PT	31	547	23.8	38.8	0.94
RO	31	673	22.3	49.1	0.94
RU	30	1588	20.7	31.2	0.96
SK	31	451	23.6	49.2	0.94
SL	31	1031	24.1	39.9	0.95
SQ	19	480	22.3	38.2	0.94
SR	18	665	23.4	35.8	0.94
SV	32	620	21.4	35.7	0.95
TR	31	517	22.7	35.9	0.96
UK	31	594	20.7	33.8	0.95

Table 6: Memorization evaluation on the final training data slice. We report word-level ChrF++ (higher indicates greater similarity to the reference) and averaged edit distance (higher indicates greater divergence from the reference).