Multilingual Language Model Pretraining using Machine-translated Data

Jiayi Wang[α∗] **Yao Lu**[α∗] **Maurice Weber**[β] **Max Ryabinin**[β]
David Ifeoluwa Adelani[γ] **Yihong Chen**[α] **Raphael Tang**[α] **Pontus Stenetorp**[α,δ]
[α]Centre for Artificial Intelligence, University College London [β]Together AI
[γ]Mila, McGill University, Canada CIFAR AI Chair
[δ]LLMC, National Institute of Informatics
{jiaywang,yao.lu,yihong.chen,p.stenetorp}@cs.ucl.ac.uk,
maurice.weber@hotmail.com, mryabinin0@gmail.com,
raphael.tang.25@ucl.ac.uk, david.adelani@mila.quebec

Abstract

English, as a very high-resource language, enables the pretraining of high-quality large language models (LLMs). However, the same can not be said for most other languages, likely due to a gap in the quality and diversity of available multilingual pretraining corpora. In this work, we find that documents machine-translated from a high-quality English corpus, can contribute significantly to the pretraining quality of multilingual LLMs. Concretely, we translate *FineWeb-Edu*, a high-quality English web corpus, into nine languages. resulting in a 1.7-trillion-token corpus, which we call *TransWebEdu* and pretrain a 1.3B-parameter model, *TransWebLLM*, from scratch on this corpus. Across *Non-English* understanding and reasoning tasks, we show that *TransWebLLM* matches or even outperforms multilingual LLMs of similar size, including *Llama3.2*, *Qwen2.5*, and *Gemma3*, despite being trained on an order of magnitude less data. Moreover, we show that adding fewer than 5% of *TransWebLLM*'s training tokens as domain-specific data for continued pretraining yields state-of-the-art results in Arabic, Indonesian, Swahili, and Welsh for understanding and commonsense reasoning tasks. To promote reproducibility, we release our corpus and models under Open Source Initiative-approved licenses.[1]

1 Introduction

Multilingual language models have shown remarkable potential for natural language processing (Dubey et al., 2024; Yang et al., 2025b; Gemma et al., 2024), yet their development faces a fundamental challenge: the scarcity of high-quality training data for most languages (Joshi et al., 2020; Kreutzer et al., 2022). Current practices of collecting and filtering multilingual web data leads to most languages lagging behind English performance due to the Internet's English-centric nature (Bender et al., 2021; Imani et al., 2023).

To address this issue, previous work has used pretrained LLMs to generate high-quality synthetic data (Maini et al., 2024; Abdin et al., 2024). However, this is not applicable to most languages due to limited language coverage. For example, one of the popular multilingual LLMs, *Llama 3.2* (Dubey et al., 2024), officially supports fewer than 20 languages. Thus, for low-resource languages like Welsh and Yorùbá, the limited language coverage of LLMs presents a challenge for data generation.

In this work, we explore two research questions: (i) *Can machine translation serve as a viable approach to diversify medium- and low-resource corpora?* (ii) *Is it feasible to rely entirely on machine-translated synthetic data for pretraining, and what are the limitations of this approach?* These questions are grounded in the wide accessibility and adoption of neural machine translation (NMT) models, particularly for medium- and low-resource languages, the result of years of dedicated research (Stahlberg, 2020; Costa-jussà et al., 2022). Despite its potential, the use of machine-translated data for multilingual language model (LM) pretraining remains largely underexplored (Urbizu et al., 2023a; Doshi et al., 2024; Boughorbel et al., 2024). Motivated by this, we conduct an empirical study that investigates this hypothesis for the pretraining of a multilingual LLM foundation model.

We introduce *TransWebEdu*, a large-scale multilingual corpus created by translating a subset of *FineWeb-edu* (Lozhkov et al., 2024), a high-

[1]Corpus: `hf.co/datasets/britllm/TransWebEdu`; Models: `hf.co/britllm/TransWebLLM-*`.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28087–28107

quality English corpus, into nine languages using *NLLB-200-1.3B* (Costa-jussà et al., 2022). *TransWebEdu* spans ten languages (Arabic, French, German, Indonesian, Italian, Russian, Spanish, Swahili, Welsh, and English) with more than 100B tokens per language and a total of 1.7 trillion tokens. We evaluate the efficiency of *TransWebEdu* by pretraining a 1.3B-parameter language model on the dataset. Although sentence-level NMT for document translation suffers from limited context (compared to document-level translation) that may affect the translation quality, we show that the translated documents yield substantial improvements in pretraining performance. For example, *TransWebEdu* yields improvements of 13%, 19%, and 1.5% for Swahili, Welsh, and Arabic, respectively, when compared to *Qwen3 (1.7B)* (Yang et al., 2025a), a top-performing multilingual LLM of similar size, based on overall performance across the ten languages.

In summary, our contributions are as follows:

1. We translate a high-quality, pretraining-scale English corpus into nine languages, including three medium- and low-resource languages, using a sentence-level NMT model, creating one of the largest machine-generated multilingual datasets to date, *TransWebEdu*, containing 1.7T tokens.

2. We pretrain *TransWebLLM*, a 1.3B-parameter model, from scratch on *TransWebEdu*. Despite using significantly fewer tokens, it achieves state-of-the-art multilingual performance on a broad range of reasoning tasks across nine non-English languages, outperforming or matching models of similar size trained on closed-source data, such as *Llama3.2*, *Qwen2.5*, and *Gemma3*.

3. We release our corpus, models, and training pipeline under open licenses to advance reproducibility in multilingual NLP.

2 Related Work

Recently, there has been growing interest in using synthetic data, particularly machine-translated data, to enhance multilingual capabilities of LLMs. For example, *Llama3* (Dubey et al., 2024) translated synthetic quantitative reasoning data into multiple languages to improve multilingual supervised fine-tuning. Bornea et al. (2021) enhanced cross-lingual QA transfer by augmenting English training data with machine-translated QA pairs.

However, research on large-scale translated synthetic data for multilingual LLM pretraining remains limited. Early efforts include the work of Urbizu et al. (2023b), who explored pretraining BERT models for Basque using machine-translated data from Spanish and showed that models trained solely on translated data can achieve competitive results. Similarly, Boughorbel et al. (2024) examined the limitations of pretraining using TinyStories (Eldan and Li, 2023) machine-translated into Arabic. Doshi et al. (2024) extended this line of work to low-resource Indic languages by applying quality filtering to translated corpora and pretraining models with 28M and 85M parameters from scratch. These studies, however, focused on either relatively small translated datasets (e.g., 3B Basque words and 2M Arabic stories), or evaluated only small models (ranging from 1M to 125M parameters).

In this work, we translate a 100B-token, high-quality, pretraining-scale English corpus into nine languages, including three medium- and low-resource ones, resulting in one of the largest machine-generated multilingual datasets to date with 1.7 trillion tokens. We pretrain a 1.3B-parameter model from scratch on this data and evaluate it on multilingual benchmarks covering ten languages to investigate the feasibility and limitations of our approach to multilingual LLM pretraining using translated synthetic data.

3 Pretraining with Machine-translated Multilingual Data

This section describes our pipeline for constructing a machine-translated corpus and pretraining a multilingual language model using it. Our process consists of the following steps: **(i)** We select a high-quality English pretraining dataset; **(ii)** We segment English documents into sentences, translate each sentence into target languages using a sentence-level NMT model, and reconstruct the documents by concatenating the translated sentences; **(iii)** We pretrain a language model from scratch on the resulting multilingual data mixture and validate the effectiveness of the corpus.

3.1 Pretraining Data Curation

Large language models, such as Llama (Dubey et al., 2024) and Gemma (Gemma et al., 2024), are typically trained on document-level data. In line

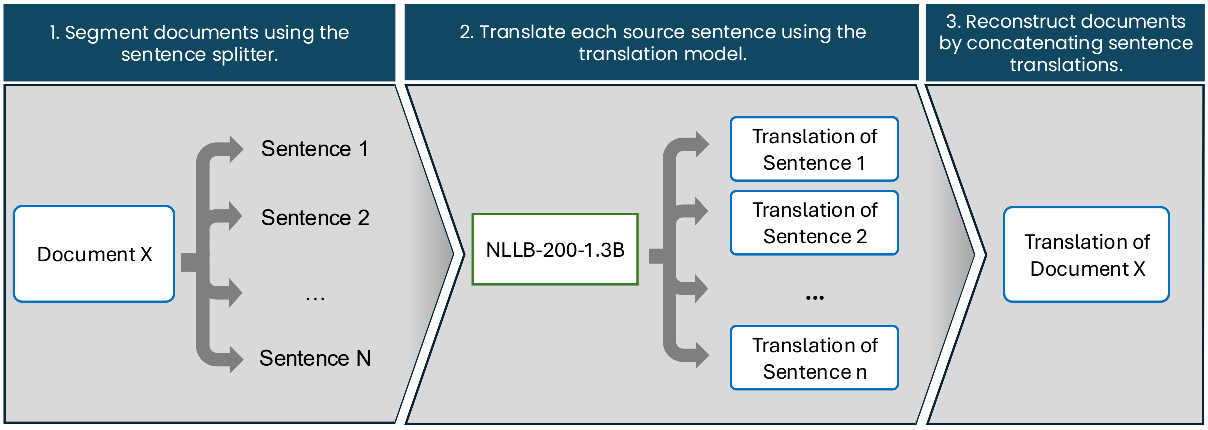

Figure 1: Step-by-step illustration of the *TransWebEdu* translation pipeline.

with this practice, we construct document-level pretraining data using four key components: source data, target languages, a translation model, and a strategy for composing document-level translations.

Source data The quality of a pretraining dataset significantly influences the performance of LLMs trained on it. Among high-resource languages, English stands out due to its linguistic diversity and coverage of topics (Joshi et al., 2020; Kreutzer et al., 2022). This makes English an excellent choice for high-quality web data. The *FineWeb-Edu* dataset[2] (Lozhkov et al., 2024), a subset of *FineWeb* (Penedo et al., 2024) consists of 1.3 trillions tokens of educational content data. Constructed using scalable automated high-quality annotations for educational value, it has been used to train both English-centric models like *GPT-2* (Karpathy, 2022, 2024) and multilingual models such as *EuroLLM* (Martins et al., 2024). Thus, we deem it a suitable candidate as a source dataset and use a randomly sampled 100B-token subset in order to comply with computational constraints for the translation process.

Target Languages We select nine target languages from several linguistic families to ensure a broad representation. From the ***Indo-European*** family, we include Germanic languages: English (en) and German (de); Romance languages: French (fr), Spanish (es), and Italian (it); a Celtic language: Welsh (cy); and a Slavic language: Russian (ru). Additionally, we include languages from distinct families: ***Afroasiatic*** (Arabic (ar)), ***Niger-Congo*** (Swahili (sw)), and ***Austronesian*** (Indonesian (id)). According to Joshi et al. (2020) and Ezeani et al. (2019), Indonesian is categorized as a medium-resource language, while Swahili and Welsh are classified as low-resource languages. The remaining languages are considered high-resource languages (although none with as many resources as English). We translate the 100B-token *FineWeb-Edu* corpus subset from English into these target languages, aiming to transfer the knowledge encoded in the English data into the other languages.

Translation Model While both NMT models and LLMs support translation (Stahlberg, 2020; Alves et al., 2024; Martins et al., 2024), LLM performance on low-resource languages remains underexplored. For instance, TowerLLM (Alves et al., 2024), a multilingual LLM for translation, covers only ten languages. In contrast, NMT models are more accessible and widely used for these languages as they benefit from years of focused development (Stahlberg, 2020). A prominent example is NLLB-200 (Costa-jussà et al., 2022), a model suite built for high-quality *sentence-level* translation across 200 languages, with strong performance even in low-resource settings.

We investigate whether document construction from sentence-level translations can yield robust pretraining performance for LLMs. Our hypothesis is that key linguistic and semantic patterns in high-quality source data can be preserved despite the potential incoherence introduced by constructing documents from translated sentences, which will offer a feasible approach for cold-start pretraining in medium- and low-resource languages and expanding access to multilingual data. Specifically, we segment English documents into sentences us-

[2] hf.co/datasets/HuggingFaceFW/fineweb-edu

Language	Tokens (B)	Avg. Doc Length (tokens)
Arabic	311.35	3,201
English	114.95	1,182
French	143.71	1,479
German	140.70	1,447
Indonesian	174.12	1,792
Italian	140.32	1,447
Russian	157.40	1,618
Spanish	140.99	1,449
Swahili	183.55	1,887
Welsh	201.49	2,071
Total	1,708.58	1,757

Table 1: Statistics of the *TransWebEdu* dataset, measured using the *Llama2* tokenizer.

GPT2-style Training Sequence
<random-fr-doc><eos><random-en-doc><eos>....<random-ru-doc>

Table 2: An illustration of our pretraining sample containing multiple non-parallel documents.

ing the NLTK sentence splitter (Bird et al., 2009), translate them using *NLLB-200-1.3B*, and reassemble the translations into documents, while preserving the original structure (e.g., newline characters). The pipeline is shown in Figure 1.

TransWebEdu We construct *TransWebEdu*, a machine-translated pretraining corpus spanning ten languages and totaling 1.7 trillion tokens. Each language is translated with a batch size of 4,096, a beam size of one, and is completed within 168 GPU hours on a single 4×GH200 node. Table 1 provides statistics for the English source and translated outputs. To the best of our knowledge, *TransWebEdu* is the largest publicly available multiway parallel, document-level corpus to date.

3.2 Multilingual LM Pretraining

This section outlines the technical details of pretraining a multilingual LM with *TransWebEdu*.

Model Architecture and Hyper-parameters We pretrain a multilingual LM from scratch using *TransWebEdu*, referred to as *TransWebLLM*. Its 1.3B-parameter architecture and hyperparameters (Appendix A) are inspired by the *Llama* models and open-source GPT-2 reproductions (Karpathy, 2024). Following these efforts, we use a constant learning rate of 2×10^{-4}, a sequence length of 2,048, and a batch size of 2,048, yielding approximately 4 million tokens per iteration.

Tokenization Alves et al. (2024) extends the multilingual capabilities of *Llama2* models (Touvron et al., 2023), demonstrating that the *Llama2* tokenizer remains a practical choice for ensuring efficiency across several languages. Building on their findings, we use the *Llama2* tokenizer in our experiments. For non-Latin script languages, such as Arabic and Russian, the *Llama2* tokenizer tokenizes the text while representing it using Unicode-based embeddings.

Pretraining Data During pretraining, we adopt a GPT-style setup by randomly sampling documents from *TransWebEdu*. This leads to a low chance of the same document appearing in multiple languages in the same batch. Table 2 shows the typical structure of a training sample.

Framework and Training We train *TransWebLLM* from scratch using the *Megatron-LM* framework (Shoeybi et al., 2019) with accelerated attention (Dao, 2023) on an NVIDIA GH200 cluster (McIntosh-Smith et al., 2024), for a total of 8,366 GPU hours. Pretraining covers approximately 1.5T tokens, roughly one epoch, and no performance degradation was observed on the validation set, which is consistent with findings from Muennighoff et al. (2024).

4 Experiments

This section presents our evaluation of model performance across various multilingual benchmarks.

4.1 Evaluation Benchmark Datasets

Our evaluation spans all ten languages in our corpus, focusing on natural language understanding and commonsense reasoning. All benchmarks are open-source, ensuring transparency and reproducibility.[3] Our evaluation framework includes the following tasks: **ARC** (Clark et al., 2018; Lai et al., 2023; Bayes et al., 2024): grade-school level multiple-choice science questions; **Hellaswag** (Zellers et al., 2019; Lai et al., 2023): commonsense reasoning benchmarks for contextually appropriate sentence endings prediction; **PAWS-X** (Yang et al., 2019): a cross-lingual adversarial dataset for paraphrase identification, sourced from Wikipedia and Quora; **PIQA** (Bisk et al., 2020): physical commonsense reasoning benchmarks; **SciQ** (Welbl et al., 2017): a multiple-choice scientific QA dataset; **TruthfulQA** (Lin

[3]Evaluations are conducted using https://github.com/EleutherAI/lm-evaluation-harness.

et al., 2021a; Bayes et al., 2024): QA evaluation tasks for the truthfulness and factual accuracy of model responses;[4] **XCOPA** (Ponti et al., 2020): a cross-lingual adaptation of COPA (Roemmele et al., 2011) for commonsense reasoning evaluation; **XNLI** (Conneau et al., 2018): a multilingual extension of Williams et al. (2018), assessing textual entailment prediction; **XStoryCloze** (Lin et al., 2021b): a multilingual adaptation of Mostafazadeh et al. (2016) for cross-lingual story ending prediction; and **XWinograd** (Tikhonov and Ryabinin, 2021): a cross-lingual adaptation of the Winograd Schema challenge[5] for coreference resolution evaluation. Benchmark availability varies for the ten languages and the specific datasets used for each language are listed in Table 9 in Appendix B.[6] We use a five-shot evaluation, report *accuracy*,[7] and the evaluations are repeated with three different seeds to ensure statistical significance.

4.2 Baselines

We benchmark *TransWebLLM* against several open-source LLMs with similar model size, but varying multilingual pretraining mixtures and data sources. Our ***multilingual LLM baselines*** include: mGPT (1.3B) (Shliazhko et al., 2022), BLOOM (1.1B) (Workshop et al., 2022), Llama3.2 (1.3B) (Dubey et al., 2024), Qwen2.5 (1.5B) (Yang et al., 2025b), Qwen3 (1.7B) (Yang et al., 2025a), Gemma3 (1B) (Team et al., 2025), and Gemma (2.6B) (Gemma et al., 2024). Additionally, we compare against ***language-specific LLM baselines***: Afriteva_v2_large (1B) (Oladipo et al., 2023) for Swahili, BritLLM (3B)[8] for Welsh, CroissantLLM (1.3B) (Faysse et al., 2024) for French, EuroLLM (1.7B) (Martins et al., 2024) for Arabic, French, German, Italian, Russian, and Spanish, Jais-family-1p3b (1.3B) (Sengupta et al., 2023) for Arabic, Sailor (1.8B) (Dou et al., 2024a) and Sailor2 (1B) (Dou et al., 2024b) for Indonesian. Furthermore, we include two ***English-centric baselines*** in our evaluation: TinyLlama (1.1B) (Zhang et al., 2024) and Pythia (1.4B) (Biderman et al., 2023). An overview of baseline models and our *TransWebLLM* is shown in Table 10 in Appendix C.

[4]We utilize the truthfulqa_mc1 for the evaluation.

[5]https://cs.nyu.edu/~davise/papers/WinogradSchemas/WS.html

[6]For Welsh, we use the *BritEval* benchmarks: https://llm.org.uk.

[7]For TruthfulQA for English and Welsh, we adopt the default lm-evaluation-harness configuration of six few-shot examples.

[8]hf.co/britllm/britllm-3b-v0.1.

4.3 Main Results

Table 3 shows the average performance of *TransWebLLM* and the baseline models across the benchmark datasets for each of the ten languages. Per-task results for each language can be found in Tables 17 to 26 in Appendix G. The last four columns of Table 3 summarizes the average performance across: (i) all languages, (ii) non-English languages, (iii) high-resource languages, and (iv) medium- and low-resource languages. *TransWebLLM* ranks among the top three models in terms of average performance for ***all languages*** and ***non-English languages***, with accuracy scores of 45.11 and 43.86, respectively. On average across *all languages*, it outperforms similarly sized multilingual LLMs, including *mGPT*, *BLOOM*, *Llama3.2*, *Qwen2.5*, and *Gemma3*. Notably, it achieves the best performance on ***medium- and low-resource languages***, with an average accuracy score of 43.25.

For ***high-resource languages***, *TransWebLLM* outperforms *Llama3.2* on average (45.90 vs. 44.13), despite being trained on significantly less data (1.5T vs. 9T tokens) and performs comparably to *Gemma3* (45.90 vs. 46.04). It ranks among the top three models for Arabic and Italian. For French, *TransWebLLM* surpasses *CroissantLLM* (46.57 vs. 45.17), which is trained on 3T tokens with half in French, while *TransWebLLM* uses only 150 billion French machine-translated tokens.

For ***medium- and low-resource languages***, *TransWebLLM* outperforms *Qwen2.5* on Indonesian (47.48 vs. 46.65), despite *Qwen2.5* being trained on 18T tokens. For Swahili and Welsh, *TransWebLLM* ranks first among all baselines, achieving accuracy scores of 43.76 and 38.52; outperforming *Gemma* (2.6B) and *BritLLM* (3B). These results suggest that training with translation data can be a viable cold-start strategy for pretraining LLMs in medium- and low-resource languages.

5 Discussion and Ablations

This section investigates (1) the impact of LLM-generated translation data on pretraining performance, (2) the effects of incorporating additional sources such as general web data, rephrased synthetic text, QA and code data, and (3) multilingual

				High				Medium	Low			Average		
	ar	en	fr	de	it	ru	es	id	sw	cy	All	Non-eng	High	Med.& Low
English LLMs														
Pythia (1.4B)	32.95	54.71	41.43	36.25	34.71	38.30	40.04	36.87	37.34	31.45	38.41	36.59	39.77	35.22
TinyLlama (1.1B)	32.50	57.12	43.52	36.36	36.84	41.36	42.02	36.13	36.94	31.48	39.43	37.46	41.39	34.85
Multilingual LLMs														
mGPT (1.3B)	32.52	45.36	39.95	35.17	34.71	40.45	39.33	39.32	38.71	31.11	37.66	36.81	38.21	36.38
BLOOM (1.1B)	34.90	51.16	43.52	34.69	33.76	37.49	42.61	43.25	37.17	31.18	38.97	37.62	39.73	37.20
Llama3.2 (1.3B)	34.64	58.12	44.89	39.84	41.04	45.56	44.85	44.81	37.76	31.55	42.31	40.55	44.13	38.04
Qwen2.5 (1.5B)	37.22	63.94	49.44	42.25	43.84	47.36	48.87	46.65	37.72	31.93	44.92	42.81	47.56	38.77
Qwen3 (1.7B)	38.83	64.90	53.09	46.25	48.32	49.77	51.84	50.28	38.56	32.32	47.42	45.47	50.43	40.39
Gemma3 (1B)	37.85	58.35	47.85	40.39	44.69	46.85	46.29	49.85	38.55	31.90	44.26	42.69	46.04	40.10
Gemma (2.6B)	37.19	62.42	49.61	43.50	44.22	48.35	49.13	48.71	40.23	31.99	45.54	43.66	47.77	40.31
Language-Specific LLMs														
AfriTeVa (1B)		37.70							40.25					
BritLLM (3B)		60.06								37.26				
CroissantLLM (1.3B)		53.34	45.17											
EuroLLM (1.7B)	38.59	57.95	48.23	42.03	47.29	46.68	47.10						46.84	
Jais-family-1p3b (1.3B)	39.59	56.31												
Sailor (1.8B)		55.53						48.84						
Sailor2 (1B)		54.38						49.84						
Ours														
TransWebLLM (1.3B)	39.41	56.32	46.57	41.59	45.51	46.08	45.84	47.48	43.76	38.52	45.11	43.86	45.90	43.25

Table 3: Evaluation of LLMs for ten languages. For each language, the scores are the averaged performance over benchmarks. Detailed model performance for each benchmark is described in Appendix G. The last four columns report mean scores for all languages (All), non-English languages (Non-Eng), high-resource languages (High), and medium- and low-resource languages (Med.&Low). The top three models in each column are underlined, and the best one is highlighted.

behavior analysis for interpretability.

5.1 Pretraining with LLM-generated Translation Data

Recent work (Alves et al., 2024; Martins et al., 2024) show that LLMs can perform translation tasks. These results prompt the question: *How does pretraining performance differ between LLM and NMT translations, as used in TransWebLLM, given their translation quality differences?*

Dubey et al. (2024) showed Mistral's potential for multilingual NLP, while Moslem et al. (2023) and Kocmi et al. (2024) demonstrate its effectiveness for machine translation. In our preliminary evaluation, *Mistral-7B-Instruct-v0.1* achieved BLEU scores of 28.75 on WMT14 EN-FR and 23.88 on WMT16 EN-DE in a zero-shot setting, outperforming supervised NMT systems trained on 30 million parallel sentences, which score 27.97 and 21.33, respectively (Lample et al., 2017). Based on these findings, we adopt *Mistral-7B-Instruct-v0.1*[9] for translation, focusing on *English*, *French*, *German*, and *Spanish* due to its limited language coverage. Details on data generation are provided in Appendix D. A key distinction between Mistral- and NLLB-generated translations lies in the text segmentation: Mistral is prompted to translate chunked documents, better preserving context, while NLLB operates at the sentence level, which may lead to reduced document-level coherence.

Due to computational constraints, we translate 64B English tokens from the sample-100BT subset of *FineWeb-Edu*. We then pretrain a new model from scratch with the same framework and hyperparameters as *TransWebLLM*, referring to it as *CuatroLLM*. For fair comparison, we pretrain a model on the same 64B English tokens and their corresponding NLLB-translated French, German, and Spanish data, which is referred to as *TransWebLLM*-4. Both models are evaluated at the same training step, after processing 470B tokens for the four languages. Due to space constraints, the results are presented in Table 12 in Appendix D. On average, *CuatroLLM* and *TransWebLLM*-4 achieve comparable performance (46.47 vs. 46.57), both outperforming *mGPT* and *BLOOM*, and almost matching *Llama3.2*. This suggests that the choice of translation method, using either *Mistral-7B-Instruct-v0.1* or NLLB, has only a limited impact on pretraining performance. However, the NLLB model offers a key advantage: its support for 200 languages enables scalable multilingual pretraining across a much wider language spectrum.

5.2 Beyond Pretraining with Translation Data

In this section, we assess whether incorporating specialized data offers additional benefits beyond

[9] hf.co/mistralai/Mistral-7B-Instruct-v0.1

Model	# tokens	Method	Data
TransWebLLM	1.5T	Train from scratch	*TransWebEdu*
TransWebLLM-web	+90B	Continue train on *TransWebLLM*	*TransWebEdu* + Real web data
TransWebLLM-cool	+62B	Continue train on *TransWebLLM-web*	*TransWebEdu* + Real web data + MC synthetic data + Cooldown Data

Table 4: Models used in data impact ablations.

pretraining with machine-translated data.

5.2.1 Impact of General Web Data

TransWebEdu is primarily composed of educational content, a highly specialized domain. We explore whether incorporating general web data can further improve multilingual reasoning capabilities.

We construct a general web dataset by sampling English, French, German, Italian, and Spanish data from *RedPajama-v2* (*RPv2*) (Weber et al., 2024); Arabic, Russian, and Indonesian from *mC4* (Xue et al., 2021); Swahili from *Wura* (Oladipo et al., 2023); and Welsh from *CC100* (Wenzek et al., 2020). For *RPv2*, we filter each subset using its built-in quality signals, as described in Appendix E; for *mC4*, we apply random sampling. Given the limited availability of Swahili and Welsh data in *Wura* and *CC100*, we include their entire datasets. We balance the general web data by sampling a nearly equal number of tokens per language, upsampling Indonesian, Swahili, and Welsh as needed to match their proportions in *TransWebEdu*. We then merge it with *TransWebEdu* at a nearly 1:0.8 ratio[10] for continued pretraining. Building on *TransWebLLM*, we extend training for an additional 20,800 steps, processing approximately 90B tokens during this phase, with general web data accounting for only around 40B tokens (less than 3%). We refer to this continued pretraining model as *TransWebLLM-web*, as detailed in Table 4.

Understanding and Reasoning Evaluation The evaluation results of *TransWebLLM-web* are presented in Table 13 in Appendix F, with per-task averaged results over three random seeds in Tables 17 to 26 in Appendix G. As shown, *TransWebLLM-web* outperforms *TransWebLLM* with consistently higher average scores. The last row of the table summarizes these performance gains. These results underscore the value of incorporating even a limited amount of web data during continued pretraining for multilingual understanding and reasoning.

[10]We aimed to balance the data across all ten languages based on general web sources. However, for Welsh and Swahili, the available data is extremely limited, compared with other languages. We avoid excessive upsampling to maintain training performance.

Model	*fr-grammar*	*fr-vocab*	Avg.
Baselines			
EuroLLM*	79.83	78.99	79.41
Qwen2.5	71.43	73.95	72.69
Qwen3	78.99	78.15	78.57
Gemma	73.11	72.27	72.69
CroissantLLM*	79.83	78.15	78.99
Ours			
TransWebLLM	67.23	63.03	65.13
TransWebLLM-web	73.11	76.47	74.79

Table 5: French grammar and vocabulary proficiency evaluation of *TransWebLLM-web*, measured in accuracy, compared to the top French-performing models from Table 13 and French-specific LLMs. Models marked with * are regional models trained with French support.

Model	*colloquial*	*standard*	Avg.
Baselines			
Qwen3	53.31	56.71	55.01
Gemma3	56.53	61.18	58.86
Sailor*	57.60	65.47	61.54
Sailor2*	58.86	66.37	62.62
Ours			
TransWebLLM	48.12	49.55	48.84
TransWebLLM-web	55.46	59.75	57.61

Table 6: COPAL-ID evaluation of *TransWebLLM-web*, measured in accuracy, compared to the top Indonesian-performing models from Table 13 and Indonesian-specific LLMs. Models marked with * are regional models trained with Indonesian support.

Linguistic Proficiency Evaluation We also evaluate the model's linguistic proficiency, focusing on its ability to understand and generate coherent, grammatically accurate sentences. Faysse et al. (2024) introduced the *fr-grammar* and *fr-vocabulary* test sets in French to assess models' grammar and vocabulary capabilities through structured language evaluations. We test both *TransWebLLM* and *TransWebLLM-web* on these benchmarks in a 5-shot setting to measure their proficiency in French linguistic competence. As shown in Table 5, *TransWebLLM-web* outperforms *TransWebLLM* by nearly 10 accuracy points (74.79 vs. 65.13) on average, demonstrating that even a small addition of general web data in continued pretraining can significantly enhance linguistic proficiency.

Reasoning Evaluation for Local Culture Local culture reasoning reflects causal understanding within specific cultural contexts. COPAL-ID (Wibowo et al., 2023) is an Indonesian causal reasoning dataset written from scratch by native speakers in both standard and Jakartan Indonesian, a widely spoken dialect. We evaluate both *TransWebLLM* and *TransWebLLM-web* on this benchmark in a 5-shot setting to assess their ability to reason within the Indonesian cultural sphere. As shown in Table 6, *TransWebLLM-web* improves Indonesian cultural reasoning by over an averaged 8 accuracy points (57.61 vs. 48.84) by incorporating a limited amount of general web data in continued pretraining on *TransWebLLM*.

5.2.2 Impact of Special Data

Yang et al. (2023) shows that rephrasing MMLU (Hendrycks et al., 2021) samples enhances model reasoning performance across various domains. Motivated by these findings, we explore the impact of ***rephrased synthetic data*** on *TransWebLLM*. Instead of rephrasing MMLU test cases (Yang et al., 2023), we rephrase English web data into a multiple-choice (MC) style using an LLM, aligning with reasoning structure while maintaining its open-ended nature. We extract 10BT English data from SlimPajama (Soboleva et al., 2023), generate about 8BT MC synthetic data using *Mistral-7B-Instruct-v0.1*,[11] and upsample and integrate it into *TransWebEdu* with general web data, ensuring MC data constitutes about 5% of the corpus. Given the improved performance of *TransWebLLM-web*, we continue pretraining for 9,000 steps, processing 38B tokens, including 2B tokens from the MC data.

Prior works (Faysse et al., 2024; Zhang et al., 2024; Martins et al., 2024) highlight the importance of a cooldown phase for enhancing model capabilities. While *TransWebEdu* emphasizes educational content, it lacks code and instruction data, such as question-answering (QA), compared to other LLMs. To address this, we introduce ***cooldown data*** during this phase: Python-Edu (Ben Allal et al., 2024), an educational Python dataset from The Stack (4.4B tokens), and WebInstruct (Yue et al., 2024), a curated QA dataset (0.8B tokens) from the web. They are up-sampled and mixed with the previous-stage data (Table 4), forming 30% of the total. The model undergoes an additional 24B-token training phase using a reduced learning rate.[12] Notably, cooldown data constitutes about 7B tokens, accounting for about 0.4% of total training tokens. We denote this final cooldown-trained model as ***TransWebLLM-cool***.

We evaluate *TransWebLLM-cool* on all benchmarks used in Sections 4.1 and 5.2.1, as well as Global-MMLU (Singh et al., 2024), covering nine languages (excluding Welsh), in a 5-shot setting. As shown in Table 14 (Appendix F), *TransWebLLM-cool*, trained with additional rephrased synthetic and cooldown data, ranks among the top three models on Global-MMLU, on average across all and high-resource languages.

Furthermore, Table 7 shows that *TransWebLLM-cool* surpasses *TransWebLLM-web* across nine Non-English languages for understanding and reasoning tasks, ranking as the best LLM on average. Remarkably, it is the ***best-performing*** LLM for ***Arabic, Indonesian, and Welsh***. In addition, Tables 15 and 16 in Appendix F demonstrate that *TransWebLLM-cool*, despite being trained with limited amount of special data, further improves both French linguistic proficiency and Indonesian local cultural reasoning over *TransWebLLM-web*. These findings underscore the effectiveness of rephrased synthetic and cooldown data in enhancing multilingual pretraining built on NLLB-translated data and limited web data.

5.3 Multilingual Behavior Analysis

Wendler et al. (2024) used the logit lens (Nostalgebraist, 2020) to show that multilingual LLMs trained on English-heavy data develop a latent space biased toward English. We apply a similar method by projecting intermediate layer outputs through the final-layer linear transformation to examine whether *TransWebLLM* models rely on English as a pivot when processing Non-English languages. To control for semantic variation, we use the devtest set of FLORES-200 (Costa-jussà et al., 2022), which provides semantically aligned texts across languages. For each layer, we apply the logit lens and use FastText (Joulin et al., 2016) to identify the language distribution of the generated tokens. We then plot the predicted probabilities of English and the target language. Our analysis compares *TransWebLLM* models with *Llama3.2*, *Qwen2.5*, and *Qwen3*.

Results (detailed in Appendix H) show that *TransWebLLM* models generally exhibit lower English

[11] We use the prompt template as "*Write multiple-choice questions and answers based on the document: [doc]*".

[12] We apply a constant learning rate schedule: 2×10^{-4} for earlier pretraining phases and 6×10^{-5} for cooldown, where we also reduce the batch size to 1024.

	High							Medium	Low		Average			
	ar	en	fr	de	it	ru	es	id	sw	cy	All	Non-eng	High	Med.& Low
English LLMs														
Pythia (1.4B)	32.95	54.71	41.43	36.25	34.71	38.30	40.04	36.87	37.34	31.45	38.41	36.59	39.77	35.22
TinyLlama (1.1B)	32.50	57.12	43.52	36.36	36.84	41.36	42.02	36.13	36.94	31.48	39.43	37.46	41.39	34.85
Multilingual LLMs														
mGPT (1.3B)	32.52	45.36	39.95	35.17	34.71	40.45	39.33	39.32	38.71	31.11	37.66	36.81	38.21	36.38
BLOOM (1.1B)	34.90	51.16	43.52	34.69	33.76	37.49	42.61	43.25	37.17	31.18	38.97	37.62	39.73	37.20
Llama3.2 (1.3B)	34.64	58.12	44.89	39.84	41.04	45.56	44.85	44.81	37.76	31.55	42.31	40.55	44.13	38.04
Qwen2.5 (1.5B)	37.22	63.94	49.44	42.25	43.84	47.36	48.87	46.65	37.72	31.93	44.92	42.81	47.56	38.77
Qwen3 (1.7B)	38.83	64.90	53.09	46.25	48.32	49.77	51.84	50.28	38.56	32.32	47.42	45.47	50.43	40.39
Gemma3 (1B)	37.85	58.35	47.85	40.39	44.69	46.85	46.29	49.85	38.55	31.90	44.26	42.69	46.04	40.10
Gemma (2.6B)	37.19	62.42	49.61	43.50	44.22	48.35	49.13	48.71	40.23	31.99	45.54	43.66	47.77	40.31
Language-Specific LLMs														
AfriTeVa (1B)		37.70							40.25					
BritLLM (3B)		60.06								37.26				
CroissantLLM (1.3B)		53.34	45.17											
EuroLLM (1.7B)	38.59	57.95	48.23	42.03	47.29	46.68	47.10						46.84	
Jais-family-1p3b (1.3B)	39.59	56.31												
Sailor (1.8B)		55.53						48.84						
Sailor2 (1B)		54.38						49.84						
Ours														
TransWebLLM (1.3B)	39.41	56.32	46.57	41.59	45.51	46.08	45.84	47.48	43.76	38.52	45.11	43.86	45.90	43.25
TransWebLLM-web (1.3B)	39.96	56.26	48.25	42.10	46.83	46.35	46.93	50.17	44.28	39.96	46.11	44.98	46.67	44.80
TransWebLLM-cool (1.3B)	40.13	57.80	48.48	42.88	48.15	47.02	47.62	50.93	44.21	40.43	46.77	45.54	47.44	45.19
Δ (Cool - Base)	+0.72	+1.48	+1.91	+1.29	+2.64	+0.94	+1.78	+3.45	+0.45	+1.91	+1.66	+1.68	+1.54	+1.94
Δ (Cool - Web)	+0.17	+1.54	+0.23	+0.78	+1.32	+0.67	+0.69	+0.76	-0.07	+0.47	+0.66	+0.56	+0.77	+0.39

Table 7: Evalution of *TransWebLLM-cool* across ten languages, measured in accuracy. In each column, the top three models are underlined and the best one is highlighted.

probabilities and reduced English dominance in intermediate representations compared to *Llama3.2*, *Qwen2.5*, and *Qwen3* on Non-English languages. When examining the probabilities of the target language, *TransWebLLM* models show a gradual increase starting in the middle-to-late layers, suggesting more stable alignment with the target language in deeper layers. In contrast, *Llama3.2*, *Qwen2.5*, and *Qwen3* exhibit a U-shaped trend, with higher probabilities in the early and final layers but a noticeable dip in the middle. This pattern might reflect a stronger reliance on an English-centric intermediate representation before shifting back to the target language toward the output layers.

6 Conclusion

We introduce *TransWebEdu*, a multilingual dataset at pretraining scale, created by machine-translating a high-quality English corpus. Our model, *TransWebLLM*, trained from scratch on this data, achieves competitive performance on understanding and reasoning benchmarks across nine Non-English languages, outperforming multilingual LLMs trained on closed data, such as *Gemma3*, *Llama3.2*, and *Qwen2.5* with similar model size. Furthermore, we show that adding fewer than 5% of *TransWebLLM*'s training tokens as domain-specific data for continued pretraining yields new state-of-the-art results in Arabic, Indonesian, Swahili, and Welsh, and leads to the best overall average performance across Non-English benchmarks. Our approach offers a scalable method for creating multilingual pretraining data, with promising results particularly for medium- and low-resource languages.

Limitations

Our study yields promising results while also identifying areas for future exploration.

TransWebLLM, trained on *TransWebEdu*, achieves competitive average performance across 10 multilingual benchmarks to state-of-the-art multilingual LLMs of similar size, such as *Qwen2.5* and *Gemma3*. Further improvements are observed with the addition of general web data, rephrased synthetic data, and code and web-instruct data. However, due to computational constraints, we haven't conducted ablation studies to determine the optimal data mixture beyond pretraining with *TransWebEdu*. Future work will extend the experiments beyond pretraining with translation data in Section 5.2 to explore optimal data mixing strategies from diverse sources.

Moreover, our experiments focus on *TransWebLLM*, a 1.3B-parameter model that has shown promising results at this scale. However, it remains unclear whether the benefits of our translated pretraining data would persist or amplify in substantially larger models (e.g., 70B+ parameters). Scaling up could provide deeper insights into multilingual learning dynamics and data efficiency. Future research will explore these aspects to validate and enhance the scalability of our multilingual pretraining approach.

Acknowledgments

We acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR), which is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via the UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. We would like to thank Jesse Lee Coombs for feedback on earlier versions of the paper and are grateful to Together AI for providing us with API credits for data generation. ChatGPT was used solely to correct grammatical errors during the writing of the paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Edward Bayes, Israel Abebe Azime, Jesujoba O Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A Etori, Shamsuddeen Hassan Muhammad, et al. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948*.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Smollm-corpus.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12583–12591.

Sabri Boughorbel, MD Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. *arXiv preprint arXiv:2405.14277*.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

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. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024a. Sailor: Open language models for south-east asia. *arXiv preprint arXiv:2404.03608*.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, Hynek Kydlíček, Zeyi Liu, Qunshu Lin, Sittipong Sripaisarnmongkol, Kridtaphad Sae-Khow, Nirattisai Thongchim, Taechawat Konkaew, Narong Borijindargoon, Anh Dao, Matichon Maneegard, Phakphum Artkaew, Zheng-Xin Yong, Quan Nguyen, Wannaphong Phatthiyaphaibun, Hoang H. Tran, Mike Zhang, Shiqi Chen, Tianyu Pang, Chao Du, Xinyi Wan, Wei Lu, and Min Lin. 2024b. Sailor2: Sailing in south-east asia with inclusive multilingual llm.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Ignatius Ezeani, Scott Piao, Steven Neale, Paul Rayson, and Dawn Knight. 2019. Leveraging pre-trained embeddings for Welsh taggers. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 270–280, Florence, Italy. Association for Computational Linguistics.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Team Gemma, 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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Andrej Karpathy. 2022. NanoGPT. `https://github.com/karpathy/nanoGPT`.

Andrej Karpathy. 2024. llm.c. `https://github.com/karpathy/llm.c`.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Preliminary wmt24 ranking of general mt systems and llms. *arXiv preprint arXiv:2407.19884*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021a. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021b. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. 2024. Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence. *arXiv preprint arXiv:2410.11199*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *arXiv preprint arXiv:2312.12740*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2024. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36.

Nostalgebraist. 2020. Interpreting GPT: The Logit Lens.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *arXiv preprint arXiv:2106.12066*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023a. Not enough data to pre-train your language model? mt to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023b. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy S Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 116462–116492. Curran Associates, Inc.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Anirudh Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods*

Hyperparameter	Value
Sequence Length	2048
Number of Layers	24
Embedding Size	2048
FFN Hidden Size	5504
Number of Heads	16
Position Encodings	RoPE
Activation Function	SwiGLU
Layer Norm	RMSNorm
Learning Rate	2E-4
Batch Size	2048
Vocabulary Size	32000
Embedding Parameters	0.13B
Non-Embedding Parameters	1.21B
Total Parameters	1.34B

Table 8: Model and pretraining hyperparameters.

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Appendix

This appendix provides additional technical details on our approach and supplementary evaluation results for the main paper.

A Hyperparameters Settings of Model Pretrainning

Pretraining hyperparameter settings are shown in Table 8.

B Specific Evaluation Benchmarks for Each Language

Specific evaluation benchmarks for each of the 10 languages are shown in Table 9.

C An Overview of Baseline Models

An overview of baseline models are shown in Table 10.

D Translation Data Generation from the Mistral-7B-Instruct LLM

We employ *Mistral-7B-Instruct-v0.1* as our translation model. However, its efficacy when prompted for document-level translation, particularly with long-context English source documents, has not yet been verified. A recent related work by Maini et al. (2024) has empirically demonstrated that prompting an LLM to rephrase more than 300 tokens could lead to information loss when rephrasing web data.

Following their setup, we first segment the English source documents from the sample-100BT subset of *FineWeb-Edu* into shorter pieces, prompt Mistral to translate these segments sequentially, and subsequently reconstruct the whole translated document by concatenating the translated segments.The detailed translation pipeline is shown in Figure 2.

Adhering to the instruction format[13] specified for *Mistral-7B-Instruct*, the chat template employed to prompt Mistral model for translation (using English-French as an example) is illustrated in Figure 3.[14] To maintain translation integrity, any sentence not fully translated to a terminal punctuation is omitted, based on the NLTK sentence splitter (Bird et al., 2009).

We translate English documents from *FineWeb-Edu* (Lozhkov et al., 2024) into three major European languages: French, German, and Spanish via prompting the Mistral-7B-Intruct model. To optimize memory efficiency and accelerate the inference process of *Mistral-7B-Instruct-v0.1*, we employ *vLLM* (Kwon et al., 2023), a library specifically designed for efficient large language model inference and serving. Using this setup, we translate approximately 54 million English documents (a subset of sample-100B of *FineWeb-Edu*) into the three target languages by prompting *Mistral-7B-Instruct-v0.1*. Table 11 presents the statistics of the original English data and the translated French, German, and Spanish. Leveraging *vLLM*'s efficiency, we estimate the total computational cost to be approximately 6.03×10^{22} FLOPs.

Table 12 compares the performance of the model trained on Mistral-generated translation data (*CuatroLLM*) with the model trained on NLLB-generated data (*TransWebLLM*-4) across English,

[13] hf.co/mistralai/Mistral-7B-Instruct-v0.1

[14] The highlighted portions in the template are adjusted according to the target language.

Language	Evaluation Datasets
Arabic	ARC-C, Hellaswag (Lai et al., 2023), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b)
English	ARC-E, ARC-C (Clark et al., 2018), Hellaswag (Zellers et al., 2019), PAWS-X (Yang et al., 2019), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), TruthfulQA (Lin et al., 2021a), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b)
French	ARC-C, Hellaswag (Lai et al., 2023), PAWS-X (Yang et al., 2019), XNLI (Conneau et al., 2018), XWinograd (Tikhonov and Ryabinin, 2021)
German	ARC-C, Hellaswag (Lai et al., 2023), PAWS-X (Yang et al., 2019), XNLI (Conneau et al., 2018)
Indonesian	ARC-C, Hellaswag (Lai et al., 2023), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2021b)
Italian	ARC-C, Hellaswag (Lai et al., 2023), XCOPA (Ponti et al., 2020)
Russian	ARC-C, Hellaswag (Lai et al., 2023), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b), XWinograd (Tikhonov and Ryabinin, 2021)
Spanish	ARC-C, Hellaswag (Lai et al., 2023), PAWS-X (Yang et al., 2019), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b)
Swahili	ARC-C, TruthfulQA (Bayes et al., 2024), XCOPA (Ponti et al., 2020), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b)
Welsh	ARC-E, ARC-C, PIQA, TruthfulQA, and XNLI from *BritEval*

Table 9: Specific evaluation benchmarks for each language.

Model	# Param.	Corpus	Corpus Size	Training Tokens	Data Avail.	Languages
Monolingual LLMs						
TinyLlama	1.1B	SlimPajama (Soboleva et al., 2023) and StarCoder training data (Li et al., 2023)	1T	3T	✔	Primarily English
Pythia	1.4B	The Pile (Gao et al., 2020)	207B	300B	✔	Primarily English
Multilingual LLMs						
mGPT	1.3B	mC4,Wiki	488B	440B	✘	61 languages
BLOOM	1.1B	BigScience Catalogue, Common Crawl, Github Code, and OSCAR (Ortiz Su'arez et al., 2019)	350B	366B	✔	46 languages
Llama3.2	1.3B	Web data, Code, and Math	-	9T	✘	At least 8 languages
Qwen2.5	1.5B	Web data, High-quality Reasoning Data	-	18T	✘	At least 30 languages
Qwen3	1.7B	Web data, High-quality Reasoning Data	-	36T	✘	119 languages
Gemma3	1B	Web data, Code, Science Articles, Parallel Data	-	2T	✘	Over 140 languages
Gemma	2.6B	Web data, Code, and Science Articles	-	2T	✘	-
Language-specific LLMs						
afriteva_v2_large	1B	Wura (Oladipo et al., 2023)	30 million	136B	✔	20 African languages
BritLLM	3B	SlimPajama (Soboleva et al., 2023), QA and MC Synthetic Data, Wiki, NLLB	668B	-	✘	5 British languages
CroissantLLM	1.3B	Croissant (Faysse et al., 2024)	1T	3T	✔	English, French
EuroLLM	1.7B	Web data, Parallel data, Code/Math, Wiki, ArXiv, Books, Apollo, Annealing Data	-	4T	✘	35 languages
Jais-family-1p3b	1.3B	Jais Model Family training data (Sengupta et al., 2023)	395B	1.6T	✘	Arabic, English
Sailor	1.8B	CC100 (Wenzek et al., 2020), MADLAD-400 (Kudugunta et al., 2024), OpenSubtitles, and Wiki	395B	400B	✔	English, Chinese, and 5 South-East Asian languages
Sailor2	1B	CC100 (Wenzek et al., 2020), MADLAD-400 (Kudugunta et al., 2024), OpenSubtitles, Wiki, Fineweb-Pro, Chinese-Fineweb-Edu, Open-Web-Math-Pro, and Synthetic data	-	500B	✔	15 languages
***TransWebLLM* (Ours)**	1.3B	*TransWebEdu*	1.7T	1.5T	✔	10 languages

Table 10: Overview of pretraining data across LLMs.

German, French, and Spanish benchmarks.[15]

E Sampling General Web Data from *RedPajama-v2*

We use the English, French, German, Italian, and Spanish subsets of the *RedPajama-v2* (RPv2) (Weber et al., 2024) as web data. Given that web data is inherently noisy, we make further use of the quality signals provided for RPv2 and filter each subset down to a smaller, high-quality subset. Specifically, we use the six most recent dumps from 2022 and 2023 and apply quality filtering using the Gopher rules (Rae et al., 2021). Additionally, web data often contains near duplicates, stemming from

[15]We apply the default seed of the lm-evaluation-harness framework for this evaluation.

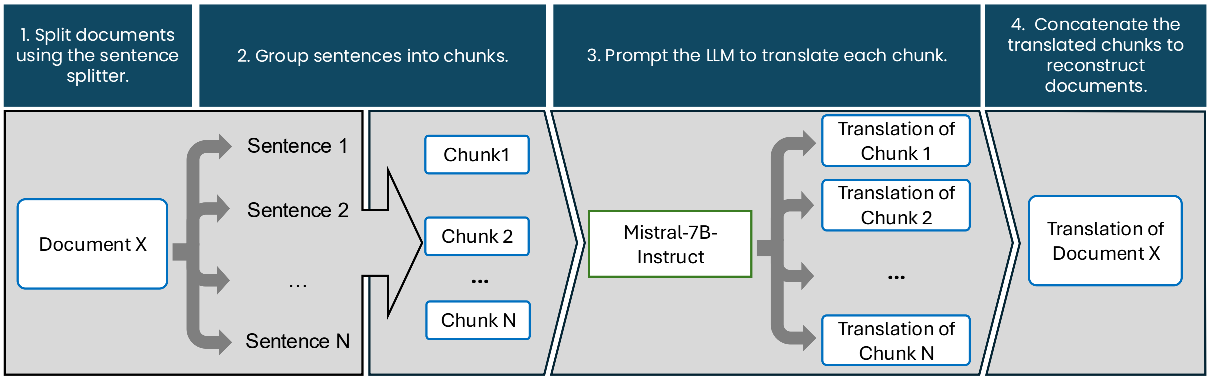

Figure 2: Step-by-step illustration of the translation pipeline with the Mistral-7B-Instruct model.

```
"<s>[INST] translate document from English to French:
{source text} [/INST] Voici les documents en français
:\n\n"

{to be prompted}
```

Figure 3: Chat template used for prompting *Mistral-7B-Instruct-v0.1* for English-French translation.

Language	Tokens (B)	Avg. Doc Length (tokens)
English	63.41	1,171.48
French	76.25	1,408.74
German	73.91	1,365.41
Spanish	72.93	1,383.25
Total	286.50	1,331.89

Table 11: Statistics of Translation Data generated with the Mistral-7B-Model, measured in Llama2 tokenizer.

	en	fr	de	es	**Avg.**
English LLMs					
Pythia (1.4B)	54.79	40.56	36.16	40.12	42.91
TinyLlama (1.1B)	57.26	43.96	36.19	42.40	44.95
Multilingual LLMs					
mGPT (1.3B)	45.63	40.22	34.94	39.43	40.06
BLOOM (1.1B)	51.47	42.31	34.93	42.72	42.86
Llama3.2 (1.3B)	58.20	44.80	40.13	45.04	47.04
Qwen3 (1.7B)	65.05	54.16	46.19	52.06	54.37
Gemma3 (1B)	58.45	48.79	40.74	46.37	48.59
Language-Specific LLMs					
CroissantLLM (1.3B)	53.45	45.45	-	-	-
EuroLLM (1.7B)	58.07	48.14	42.05	47.11	48.84
Ours					
CuatroLLM (1.3B)	55.48	45.32	40.38	44.70	46.47
TransWebLLM-4 (1.3B)	55.15	45.72	40.55	44.87	46.57

Table 12: Performance comparison between *CuatroLLM*, trained on LLM-translated data, and *TransWebLLM*-4 across four selected languages.

boilerplate text, ads, and other computer-generated text that only differs by a few words, and removing these has been shown to positively affect training efficiency and reduce the amount of memorization (Lee et al., 2021). We therefore adopt the MinHash algorithm with locality-sensitive hashing (Broder, 1997) to perform near-deduplication. We identify documents as near duplicates if their Jaccard similarity is greater than 0.8 and use 128 hash functions.

F Evaluation for Impact of Special Data

The evaluation results of *TransWebLLM-web* are presented in Table 13.

The evaluation results of *TransWebLLM-cool* on Global MMLU, French linguistic proficiency, and reasoning for Indonesian local culture are presented in Table 14, 15, and 16, respectively.

G Detailed Results per Language for Understanding and Reasoning Benchmarks

Tables 17 to 26 present detailed benchmark results for each language, as outlined in Section 4.1. All results are averaged over three runs with different random seeds to ensure stability and statistical significance. The average scores across benchmarks (shown in the last column of each table)[16] correspond to those reported in Tables 3,13, and 7, respectively.

H Logit Lens Analysis for Interpretability

Detailed logit lens visualizations for *TransWebLLM* models and other baselines are provided in Figure 4 and 5, respectively, as discussed in Section 5.3.[17]

[16] The last decimal digit in the average column may differ by 0.01 because the original benchmark results were reported with higher decimal precision and subsequently rounded.

[17] We include layer_0 in our plots, which corresponds to the embedding projection prior to the first decoder block.

	High							Medium	Low		Average			
	ar	en	fr	de	it	ru	es	id	sw	cy	All	Non-eng	High	Med.& Low
English LLMs														
Pythia (1.4B)	32.95	54.71	41.43	36.25	34.71	38.30	40.04	36.87	37.34	31.45	38.41	36.59	39.77	35.22
TinyLlama (1.1B)	32.50	57.12	43.52	36.36	36.84	41.36	42.02	36.13	36.94	31.48	39.43	37.46	41.39	34.85
Multilingual LLMs														
mGPT (1.3B)	32.52	45.36	39.95	35.17	34.71	40.45	39.33	39.32	38.71	31.11	37.66	36.81	38.21	36.38
BLOOM (1.1B)	34.90	51.16	43.52	34.69	33.76	37.49	42.61	43.25	37.17	31.18	38.97	37.62	39.73	37.20
Llama3.2 (1.3B)	34.64	58.12	44.89	39.84	41.04	45.56	44.85	44.81	37.76	31.55	42.31	40.55	44.13	38.04
Qwen2.5 (1.5B)	37.22	63.94	49.44	42.25	43.84	47.36	48.87	46.65	37.72	31.93	44.92	42.81	47.56	38.77
Qwen3 (1.7B)	38.83	64.90	53.09	46.25	48.32	49.77	51.84	50.28	38.56	32.32	47.42	45.47	50.43	40.39
Gemma3 (1B)	37.85	58.35	47.85	40.39	44.69	46.85	46.29	49.85	38.55	31.90	44.26	42.69	46.04	40.10
Gemma (2.6B)	37.19	62.42	49.61	43.50	44.22	48.35	49.13	48.71	40.23	31.99	45.54	43.66	47.77	40.31
Language-Specific LLMs														
AfriTeVa (1B)		37.70							40.25					
BritLLM (3B)		60.06								37.26				
CroissantLLM (1.3B)		53.34	45.17											
EuroLLM (1.7B)	38.59	57.95	48.23	42.03	47.29	46.68	47.10						46.84	
Jais-family-1p3b (1.3B)	39.59	56.31												
Sailor (1.8B)		55.53						48.84						
Sailor2 (1B)		54.38						49.84						
Ours														
TransWebLLM (1.3B)	39.41	56.32	46.57	41.59	45.51	46.08	45.84	47.48	43.76	38.52	45.11	43.86	45.90	43.25
TransWebLLM-web (1.3B)	39.96	56.26	48.25	42.10	46.83	46.35	46.93	50.17	44.28	39.96	46.11	44.98	46.67	44.80
Δ Gain	+0.55	-0.06	+1.68	+0.51	+1.32	+0.27	+1.09	+2.69	+0.52	+1.44	+1.00	+1.12	+0.77	+1.55

Table 13: Performance comparison between *TransWebLLM* and *TransWebLLM-web* in a 5-shot setting across ten languages. The last row (Δ Gain) shows the performance difference, with positive values indicating improvements of *TransWebLLM-web* over *TransWebLLM*. In each column, the top three models are underlined and the best one is highlighted.

	High							Medium	Low	Average	
	ar	en	fr	de	it	ru	es	id	sw	All	High
Multilingual LLMs											
mGPT (1.3B)	25.02	25.27	26.10	24.05	25.70	25.48	25.64	25.10	24.11	25.16	25.32
BLOOM (1.1B)	26.36	26.25	26.65	26.51	27.25	26.76	26.09	25.86	26.61	26.48	26.55
Llama3.2 (1.3B)	27.72	31.17	27.69	27.94	27.67	27.54	28.19	27.86	26.39	28.02	28.27
Qwen3 (1.7B)	45.45	62.23	55.18	53.71	54.51	51.69	55.58	52.49	33.28	51.57	54.05
Gemma3 (1B)	25.57	26.16	26.27	26.87	26.58	26.71	26.28	26.22	26.13	26.31	26.35
Language-Specific LLMs											
AfriTeVa (1B)		26.87							26.93		
CroissantLLM (1.3B)		25.35	25.36								
EuroLLM (1.7B)	26.23	27.13	26.79	26.47	26.25	27.61	26.37				26.69
Jais-family-1p3b (1.3B)	25.94	25.06									
Sailor (1.8B)		28.62						26.39			
Sailor2 (1B)		37.03						33.34			
Ours											
TransWebLLM (1.3B)	26.63	24.66	25.69	25.46	25.32	26.42	26.21	25.28	25.35	25.67	25.77
TransWebLLM-web (1.3B)	26.49	26.41	26.84	26.11	26.16	26.56	26.58	26.68	26.48	26.48	26.45
TransWebLLM-cool (1.3B)	30.44	34.26	32.58	32.27	31.95	32.50	32.53	33.18	31.11	32.31	32.36

Table 14: Evaluation on Global-MMLU full sets (Singh et al., 2024), measured in accuracy. The rightmost columns report the average scores across all languages (All) and high-resource languages (High). Top 3 models are underlined.

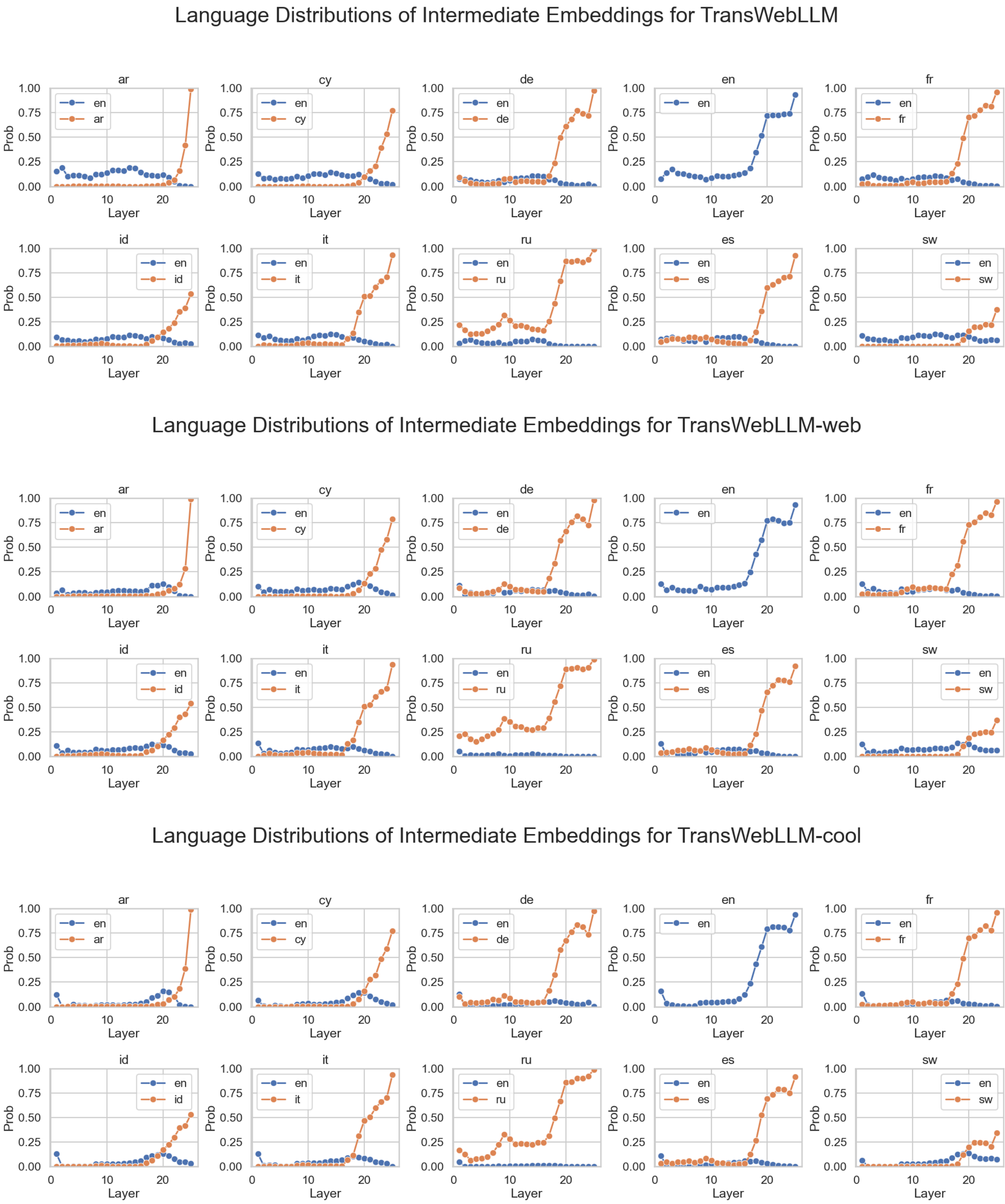

Figure 4: Logit lens outputs for *TransWebLLM* models across 10 languages.

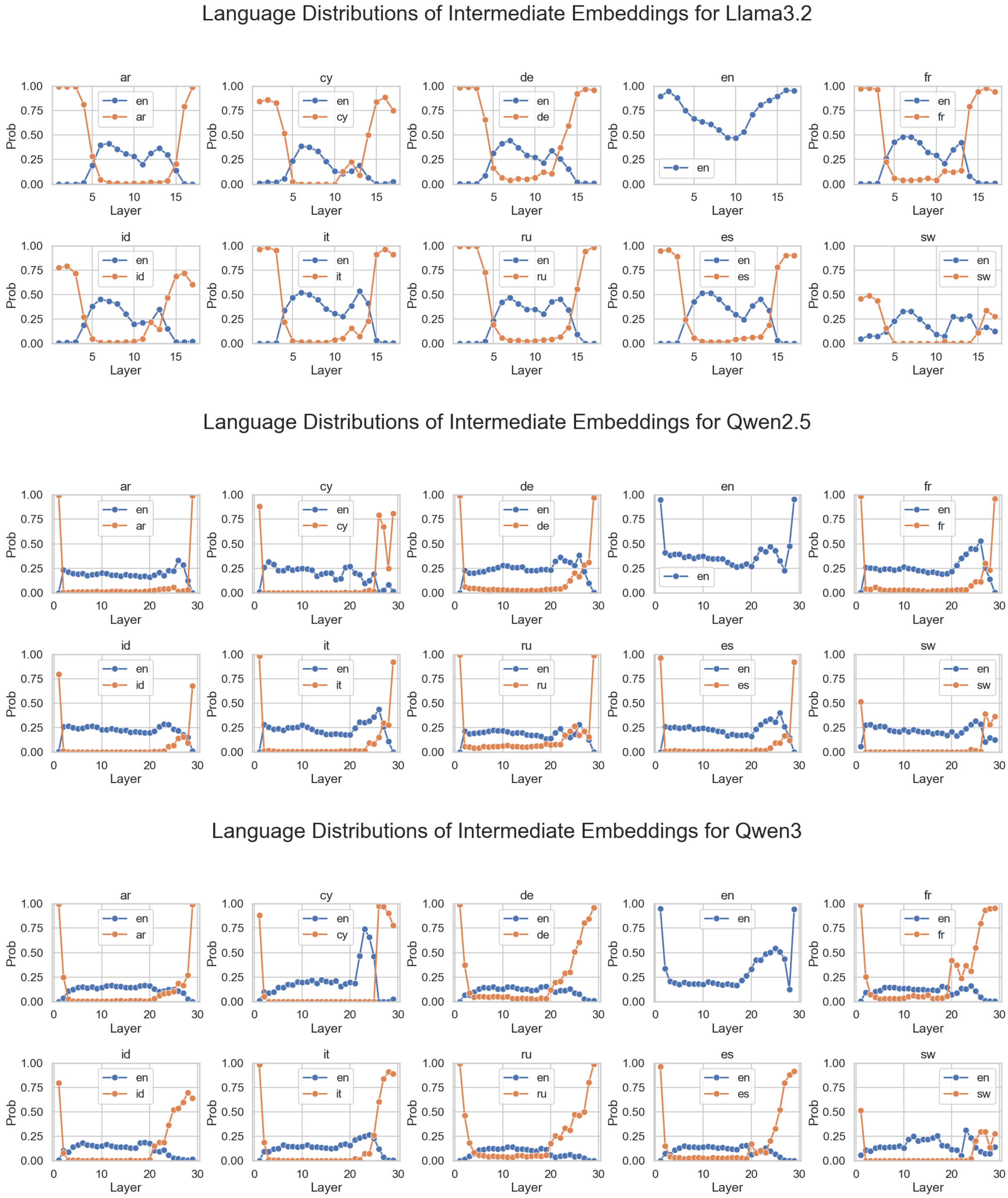

Figure 5: Logit lens outputs for baseline models across 10 languages.

Model	***fr-grammar***	***fr-vocab***	**Avg.**
Baselines			
EuroLLM*	79.83	78.99	79.41
Qwen2.5	71.43	73.95	72.69
Qwen3	78.99	78.15	78.57
Gemma	73.11	72.27	72.69
CroissantLLM*	79.83	78.15	78.99
Ours			
TransWebLLM	67.23	63.03	65.13
TransWebLLM-web	73.11	76.47	74.79
TransWebLLM-cool	78.15	73.95	76.05

Table 15: French grammar and vocabulary proficiency evaluation of *TransWebLLM-cool*, measured in accuracy, compared to the top French-performing models from Table 7 and French-specific LLMs. Models marked with * are regional models trained with French support.

Model	***colloquial***	***standard***	**Avg.**
Baselines			
Qwen3	53.31	56.71	55.01
Gemma3	56.53	61.18	58.86
Sailor*	57.60	65.47	61.54
Sailor2*	58.86	66.37	62.62
Ours			
TransWebLLM	48.12	49.55	48.84
TransWebLLM-web	55.46	59.75	57.61
TransWebLLM-cool	55.99	61.90	58.95

Table 16: COPAL-ID evaluation of *TransWebLLM-cool*, measured in accuracy, compared to the top Indonesian-performing models from Table 7 and Indonesian-specific LLMs. Models marked with * are regional models trained with Indonesian support.

Model	ARC-C	Hellaswag	XNLI	XStoryCloze	Avg.
Pythia	21.10	27.16	35.65	47.89	32.95
TinyLlama	20.25	26.87	34.54	48.33	32.50
mGPT	20.27	25.99	34.03	49.79	32.52
BLOOM	22.01	29.74	35.30	52.55	34.90
Llama3.2	22.47	30.53	34.18	51.38	34.64
EuroLLM	26.15	33.90	36.53	57.76	38.59
Qwen2.5	26.86	32.15	34.36	55.51	37.22
Qwen3	30.85	33.75	35.75	54.97	38.83
Gemma3	24.89	33.40	35.10	58.02	37.85
Gemma (2.6B)	27.06	32.29	35.10	54.31	37.19
jais-family-1p3b	27.86	35.62	35.28	59.61	39.59
TransWebLLM	30.25	34.35	36.00	57.02	39.41
TransWebLLM-web	29.17	36.04	35.73	58.88	39.96
TransWebLLM-cool	30.48	35.95	35.18	58.92	40.13

Table 17: Detailed Arabic Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	ARC-E	Hellaswag	PAWS	PIQA	SciQ	TruthfulQA	XNLI	XStoryCloze	Avg.
Pythia	28.27	63.95	40.53	57.52	71.07	92.07	22.85	48.23	67.88	54.71
TinyLlama	34.33	67.72	46.36	57.72	73.92	93.10	22.28	47.04	71.63	57.12
mGPT	21.42	49.06	30.66	54.98	64.31	61.50	23.26	42.40	60.62	45.36
BLOOM	24.63	54.71	34.70	54.77	67.56	89.67	25.58	46.47	62.36	51.16
EuroLLM	36.92	71.66	44.82	55.80	73.40	94.60	24.03	49.13	71.19	57.95
Llama3.2	35.29	69.00	48.12	55.40	75.37	95.13	23.30	48.77	72.71	58.12
Qwen2.5	48.83	80.60	49.93	67.07	76.53	96.90	29.90	51.22	74.48	63.94
Qwen3	50.97	81.33	49.28	70.78	76.50	97.30	32.52	51.70	73.70	64.90
Gemma3	35.81	71.49	47.30	57.07	75.83	94.93	22.03	48.51	72.18	58.35
Gemma (2.6B)	47.33	77.27	52.81	63.62	76.88	96.50	22.07	48.53	76.77	62.42
Afriteva-v2-large	20.93	31.31	26.66	50.53	56.19	43.47	25.21	35.81	49.15	37.70
BritLLM	38.37	72.66	51.00	57.77	75.80	96.03	24.44	48.82	75.62	60.06
CroissantLLM	26.74	62.92	40.93	51.93	72.24	92.63	23.62	43.00	66.07	53.34
Jais-family-1p3b	29.78	64.87	42.58	60.42	72.65	93.97	25.46	48.21	68.90	56.31
Sailor	29.64	64.07	42.63	58.22	72.78	93.40	22.28	46.98	69.82	55.53
Sailor2	29.69	64.35	39.92	55.00	70.00	94.53	22.89	45.80	67.22	54.38
TransWebLLM	37.12	71.63	40.61	57.53	70.73	93.07	22.97	47.95	65.30	56.32
TransWebLLM-web	36.80	71.84	41.02	57.23	70.40	93.63	21.99	46.40	67.06	56.26
TransWebLLM-cool	39.05	72.70	42.13	59.90	71.40	93.83	25.66	48.06	67.51	57.80

Table 18: Detailed English Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	Hellaswag	PAWS	XNLI	XWinograd	Avg.
Pythia	20.99	29.74	53.22	43.76	59.44	41.43
TinyLlama	24.98	32.66	52.15	42.74	65.06	43.52
mGPT	20.13	27.17	52.95	40.86	58.64	39.95
BLOOM	22.87	33.79	53.45	46.04	61.45	43.52
Llama3.2	27.17	35.93	53.27	44.64	63.46	44.89
EuroLLM	32.13	40.20	52.85	45.68	70.28	48.23
Qwen2.5	33.33	38.31	62.15	44.73	68.67	49.44
Qwen3	39.75	40.11	65.73	46.75	73.09	53.09
Gemma3	29.63	39.24	54.78	45.34	70.28	47.85
Gemma (2.6B)	34.98	39.82	59.13	47.04	67.07	49.61
CroissantLLM	25.55	39.52	50.35	44.54	65.87	45.17
TransWebLLM	35.67	38.88	53.38	44.66	60.24	46.57
TransWebLLM-web	35.79	40.34	54.95	44.73	65.46	48.25
TransWebLLM-cool	36.50	40.65	56.37	45.86	63.05	48.48

Table 19: Detailed French Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	Hellaswag	PAWS	XNLI	Avg.
Pythia	19.76	28.76	55.05	41.45	36.25
TinyLlama	21.70	30.78	52.32	40.63	36.36
mGPT	19.65	27.65	52.42	40.95	35.17
BLOOM	20.65	27.12	53.52	37.48	34.69
Llama3.2	26.06	34.22	55.02	44.06	39.84
EuroLLM	28.97	37.73	54.97	46.47	42.03
Qwen2.5	28.91	34.99	61.12	43.98	42.25
Qwen3	36.27	37.77	64.87	46.09	46.25
Gemma3	26.04	37.11	54.83	43.59	40.39
Gemma (2.6B)	31.19	37.33	60.65	44.83	43.50
TransWebLLM	32.51	36.34	53.95	43.58	41.59
TransWebLLM-web	31.28	37.60	55.82	43.71	42.10
TransWebLLM-cool	32.68	37.92	56.23	44.70	42.88

Table 20: Detailed German Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	Hellaswag	XCOPA	Avg.
Pythia	20.64	29.10	54.40	34.71
TinyLlama	23.15	31.23	56.13	36.84
mGPT	19.28	27.64	57.20	34.71
BLOOM	20.45	28.43	52.40	33.76
Llama3.2	26.95	34.85	61.33	41.04
EuroLLM	33.02	39.59	69.27	47.29
Qwen2.5	32.31	35.82	63.40	43.84
Qwen3	40.58	38.53	65.87	48.32
Gemma3	29.40	37.99	66.67	44.69
Gemma (2.6B)	32.16	37.42	63.07	44.22
TransWebLLM	36.30	37.36	62.87	45.51
TransWebLLM-web	35.79	39.03	65.67	46.83
TransWebLLM-cool	36.84	39.35	68.27	48.15

Table 21: Detailed Italian Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	Hellaswag	XNLI	XStoryCloze	XWinograd	Avg.
Pythia	18.99	27.55	39.30	49.35	56.29	38.30
TinyLlama	22.59	30.53	39.23	54.05	60.42	41.36
mGPT	20.38	26.72	40.29	56.34	58.52	40.45
BLOOM	19.62	27.40	37.52	48.29	54.61	37.49
Llama3.2	25.24	34.24	42.76	59.74	65.82	45.56
EuroLLM	28.68	36.53	45.09	62.67	60.42	46.68
Qwen2.5	31.60	36.17	42.49	62.12	64.45	47.36
Qwen3	36.87	37.75	45.85	62.26	66.14	49.77
Gemma3	26.72	36.35	42.72	64.26	64.23	46.85
Gemma (2.6B)	32.59	36.77	44.93	62.08	65.40	48.35
TransWebLLM	32.02	35.62	41.39	58.64	62.75	46.08
TransWebLLM-web	31.82	36.88	41.04	60.18	61.80	46.35
TransWebLLM-cool	33.34	37.07	40.63	61.11	62.96	47.02

Table 22: Detailed Russian Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	HellaSwag	PAWS	XNLI	XStoryCloze	Avg.
Pythia	21.71	30.17	52.78	41.85	53.68	40.04
TinyLlama	23.76	33.39	54.43	41.35	57.16	42.02
mGPT	20.31	28.23	52.07	40.86	55.20	39.33
BLOOM	24.30	34.47	51.75	44.21	58.33	42.61
Llama3.2	29.11	37.22	53.42	42.62	61.88	44.85
EuroLLM	32.42	41.08	52.67	44.91	64.44	47.10
Qwen2.5	35.98	39.43	60.83	43.80	64.28	48.87
Qwen3	42.02	41.30	64.75	46.29	64.81	51.84
Gemma3	30.77	39.80	53.85	42.46	64.57	46.29
Gemma (2.6B)	36.10	41.46	58.13	44.50	65.47	49.13
TransWebLLM	34.84	39.12	54.07	43.02	58.15	45.84
TransWebLLM-web	35.16	40.65	55.60	42.69	60.53	46.93
TransWebLLM-cool	36.32	40.92	56.67	43.12	61.06	47.62

Table 23: Detailed Spanish Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	HellaSwag	XCOPA	XStoryCloze	Avg.
Pythia	17.64	27.86	53.33	48.67	36.87
TinyLlama	16.13	27.44	51.80	49.15	36.13
mGPT	19.17	27.11	57.33	53.65	39.32
BLOOM	21.42	31.70	62.20	57.69	43.25
Llama3.2	23.82	33.99	62.07	59.36	44.81
Qwen2.5	27.63	34.84	63.87	60.25	46.65
Qwen3	36.81	37.15	65.67	61.48	50.28
Gemma3	27.89	37.00	69.93	64.59	49.85
Gemma (2.6B)	32.11	36.35	64.93	61.46	48.71
Sailor	26.53	36.39	70.07	62.39	48.84
Sailor2	27.95	36.85	70.67	63.89	49.84
TransWebLLM	34.39	36.92	60.87	57.75	47.48
TransWebLLM-web	33.73	37.91	67.13	61.90	50.17
TransWebLLM-cool	35.07	37.93	68.80	61.90	50.93

Table 24: Detailed Indonesian Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	TruthfulQA	XCOPA	XNLI	XStoryCloze	Avg.
Pythia	24.64	24.46	54.80	33.80	48.98	37.34
TinyLlama	24.58	24.08	52.73	33.68	49.64	36.94
mGPT	26.55	24.58	55.80	35.18	51.45	38.71
BLOOM	23.08	25.07	53.13	34.34	50.21	37.17
Llama3.2	27.77	23.21	52.20	33.94	51.69	37.76
Qwen2.5	25.59	27.26	52.93	33.51	49.33	37.72
Qwen3	27.50	26.48	53.93	34.54	50.37	38.56
Gemma3	27.49	21.36	53.87	35.18	54.85	38.55
Gemma (2.6B)	28.45	24.37	56.20	36.99	55.15	40.23
Afriteva_v2_large	28.65	34.20	54.07	34.97	49.37	40.25
TransWebLLM	26.82	31.76	61.60	41.66	56.94	43.76
TransWebLLM-web	27.43	27.88	64.07	42.93	59.10	44.28
TransWebLLM-cool	34.22	21.85	63.73	42.41	58.86	44.21

Table 25: Detailed Swahili Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.

Model	ARC-C	ARC-E	PIQA	TruthfulQA	XNLI	Avg.
Pythia	17.71	26.24	52.18	27.60	33.54	31.45
TinyLlama	18.83	26.56	51.62	28.00	32.40	31.48
mGPT	18.03	26.16	52.65	24.80	33.91	31.11
BLOOM	18.14	26.35	52.05	24.93	34.43	31.18
Llama3.2	18.43	26.84	53.31	25.78	33.39	31.55
Qwen2.5	18.57	26.82	51.67	27.07	35.54	31.93
Qwen3	19.17	27.71	52.56	27.69	34.48	32.32
Gemma3	18.00	27.88	53.44	27.20	32.98	31.90
Gemma (2.6B)	18.25	27.91	52.78	27.24	33.76	31.99
BritLLM	22.35	40.58	58.88	24.22	40.28	37.26
TransWebLLM	27.10	43.37	56.20	27.34	38.61	38.52
TransWebLLM-web	28.67	46.52	57.81	26.49	40.31	39.96
TransWebLLM-cool	28.58	48.13	58.32	26.89	40.22	40.43

Table 26: Detailed Welsh Benchmark Results. For each task, the reported accuracy is averaged over three random seed configurations. "Avg." denotes the overall average across tasks.