AfriqueLLM: How Data Mixing and Model Architecture Impact Continued Pre-training for African Languages

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly multilingual, yet open models continue to underperform relative to proprietary systems, with the gap most pronounced for African languages. Continued pre-training (CPT) offers a practical route to language adaptation, but improvements on demanding capabilities such as mathematical reasoning often remain limited. This limitation is driven in part by the uneven domain coverage and missing task-relevant knowledge that characterize many low-resource language corpora. We present `AfriqueLLM`, a suite of open LLMs adapted to 20 African languages through CPT on 26B tokens. We perform a comprehensive empirical study across five base models spanning sizes and architectures, including Llama 3.1, Gemma 3, and Qwen 3, and systematically analyze how CPT data composition shapes downstream performance. In particular, we vary mixtures that include math, code, and synthetic translated data, and evaluate the resulting models on a range of multilingual benchmarks. Our results identify data composition as the primary driver of CPT gains. Adding math, code, and synthetic translated data yields consistent improvements, including on reasoning-oriented evaluations. Within a fixed architecture, larger models typically improve performance, but architectural choices dominate scale when comparing across model families. Moreover, strong multilingual performance in the base model does not reliably predict post-CPT outcomes; robust architectures coupled with task-aligned data provide a more dependable recipe. Finally, our best models improve long-context performance, including document-level translation.

1 Introduction

Large language models (LLMs) are becoming increasingly multilingual, with proprietary models pre-trained on hundreds of languages (Jaech et al., 2024; Comanici et al., 2025). Open models follow a similar trend, but the performance gap with proprietary LLMs is often larger for low-resource languages, particularly African languages (Adelani et al., 2025; Ojo et al., 2025; Adebara et al., 2025). This gap highlights an opportunity to develop language- or region-specific LLMs for these languages.

Since the advent of pretrained language models such as BERT (Devlin et al., 2019), continued pretraining has become a standard approach for adapting models to new domains and languages (Gururangan et al., 2020a; Chau and Smith, 2021; Alabi et al., 2022a), and has recently been scaled to modern LLMs (Nguyen et al., 2024; Ji et al., 2025a; Buzaaba et al., 2025a). While CPT often yields significant improvements for natural language understanding (NLU) and translation tasks, gains on more challenging tasks, such as mathematical reasoning or knowledge-based QA (e.g., MMLU (Hendrycks et al., 2021)), remain limited due to uneven knowledge coverage across languages, with low-resource languages often spanning fewer domains (Buzaaba et al., 2025a).

To further improve downstream performance, LLMs are increasingly trained on heterogeneous data sources such as math, code, and other knowledge-rich corpora. These sources are often scarce in low-resource languages, yet they can substantially boost performance across a wide range of downstream tasks (Aryabumi et al., 2024; Bakouch et al., 2025; Li et al., 2025b). Recent work also shows that multilingual capability can be improved by training on machine-translated English data covering diverse domains, achieving competitive results even without adding monolingual data in the target languages (Wang et al., 2025b). Despite these advances, we still lack a comprehensive empirical understanding of how incorporating heterogeneous sources affects CPT outcomes for low-resource languages. In this work, we address this gap by systematically studying CPT data mixtures and analyzing how base-model architecture

and prior language coverage influence downstream performance after adaptation.

We introduce **AfriqueLLM**, a suite of open language models adapted to 20 African languages via efficient continued pre-training (CPT) on 26B tokens. We perform CPT on several base model spanning different architectures and scales, including Llama 3.1 8B, Gemma 3 (4B and 12B), and Qwen 3 (8B and 14B). Across these backbones, we systematically vary the CPT data mixture to quantify its impact on downstream performance. AfriqueLLM achieves strong results on multilingual benchmarks for models with fewer than 15B parameters, while largely preserving English performance.

Our evaluation leads to four main findings. *(1)* The CPT data mixture is the strongest determinant of gains. Adding math, code, and synthetic translated data consistently improves performance. *(2)* Within a fixed architecture, larger models generally perform better. Across architectures, however, scale alone is not predictive; for example, CPT-adapted Qwen 3 8B is competitive with Gemma 3 12B. *(3)* Strong multilingual proficiency of the base model does not reliably translate into better post-CPT results. Instead, architectural choices and task-aligned data are more predictive. *(4)* Our best models, Qwen 3 (8B and 14B), better preserve performance in high-resource languages after CPT and achieve strong results on long-context tasks such as document-level translation.

We hope these findings inform more effective adaptation of LLMs to low-resource languages. To support future work, we will publicly release our CPT-adapted AfriqueLLMs.

2 Related Work

The landscape of LLMs has undergone a paradigm shift from model-centric architectures to data-centric methodologies. While early foundational work focused on scaling parameters and compute (Kaplan et al., 2020; Brown et al., 2020), recent advancements in 2024 and 2025 have demonstrated that data quality, mixture ratios, and curriculum learning are the primary drivers of performance (Team, 2024; Yang et al., 2025a; Bakouch et al., 2025; Olmo et al., 2025; NVIDIA et al., 2025). This section focuses on two important aspects of pre-training: (1) Data mixture and (2) Continued pre-training for low-resource languages.

2.1 Data Quality, Mixture, and Synthetic Data

Data Quality and Curation. Recent efforts have focused on improving the quality of web collected data such as FineWeb (Penedo et al., 2024) dataset in the English setting. In multilingual settings, FineWeb2 (Penedo et al., 2025) extends these pipelines to scale pre-training data processing to over 1,000 languages. In the African context, this focus on quality has led to the creation of specialized datasets, such as WURA (Oladipo et al., 2023) and MADLAD-400 (Kudugunta et al., 2023).

Data Mixture and Ratios The importance of dynamic data mixtures is exemplified by the training recipes of recent models like SmolLM2 (allal et al., 2025a) and SmolLM3 (Bakouch et al., 2025), which utilize multi-stage training curricula that adjust the ratio of web, code, and math data over time. OLMo2 and OLMo3 OLMo et al. (2025); Olmo et al. (2025) further validate this approach by introducing specialized data mixes (e.g., Dolmino Mix) during the annealing phase.

The scarcity of high-quality natural text for reasoning and low-resource languages has driven the adoption of synthetic data. Joshi et al. (2024) and NVIDIA (2024) demonstrate that synthetic data can effectively bridge the gap in model alignment and pre-training. Phi-4 (Abdin et al., 2024) relies heavily on synthetic data for reasoning capabilities. In the context of multilingual pre-training, Wang et al. (2025a) and Ji et al. (2025b) show that machine-translated data from high-resource languages can significantly enhance multilingual pre-training, effectively transferring missing knowledge to low-resource languages.

2.2 Continued Pre-training

The release of powerful open-weight models has broadened access to state-of-the-art language technology. Recent families such as Llama 3.1 (Team, 2024), Qwen 3 (Yang et al., 2025a), and Gemma 3 (Team et al., 2025a) provide strong foundations for downstream adaptation. For languages and domains that are underrepresented during initial pre-training, CPT remains a primary adaptation approach (Gururangan et al., 2020b). CPT has been used to build some of the strongest BERT-based models for African languages, including the AfroXLMR series (Alabi et al., 2022b; Adelani et al., 2024a; Li et al., 2025a).

In the LLM setting, recent work has studied more efficient CPT strategies, such as learning-

rate re-warming (Gupta et al., 2023a) and replay buffers (Ibrahim et al., 2024), to reduce catastrophic forgetting. Building on these ideas, Uemura et al. (2024) and Buzaaba et al. (2025b) adapt open LLMs to African languages, releasing AfriInstruct and Lugha-Llama and showing that CPT can yield substantial gains without training from scratch.

Our work builds on CPT and explores new CPT data mixtures to develop **AfriqueLLM**, a suite of models adapted to the linguistic and cultural diversity of Africa.

3 AfriqueLLM: Data & Training Recipe

3.1 Dataset Curation

High-quality and diverse training data is essential for effective language modeling. To mitigate data scarcity for African languages, we curate a 26B-token corpus designed for continued pre-training (CPT). Our corpus combines monolingual text with code, mathematics, and domain-specific synthetic data to better cover the knowledge and skill distributions needed for downstream tasks. We describe the resulting data pipeline below.

African Monolingual Data. We collect text for the 20 most resource-rich African languages by combining three complementary sources (Table 1): FineWeb2 (Penedo et al., 2025), WURA (Oladipo et al., 2023), and MADLAD-400 (Kudugunta et al., 2023). FineWeb2 provides the backbone of our corpus due to its scale and strong filtering. We add document-level data from WURA to increase contextual diversity and longer-range coherence, and we use MADLAD-400 to improve coverage for the lower-resource languages in our set. To mitigate catastrophic forgetting during CPT, we include four high-resource languages, English, French, Portuguese, and Arabic, capped at 1B tokens per language, following Oladipo et al. (2023). Detailed corpus statistics appear in Table 7 in Appendix A.

Sampling Strategy. African-language corpora are highly imbalanced, which can cause high-resource languages to dominate training. To mitigate this, we use UniMax samplin (Chung et al., 2023), which caps each high-resource language at approximately 1B tokens and upsamples lower-resource languages for up to five epochs. This produces a more balanced sampling distribution and increases coverage of underrepresented languages (see the *UniMax* column in Table 1).

Language	Code	Raw	Ep.	UniMax	Syn.
High-Resource (Non-African)					
English	eng_Latn	>1.00B	0	1.07B	16M
French	fra_Latn	>1.00B	0	1.07B	–
Portuguese	por_Latn	>1.00B	0	1.07B	–
Arabic	arb_Arab	>1.00B	0	1.07B	–
African Languages					
Afrikaans	afr_Latn	5.30B	0	1.07B	12M
Swahili	swh_Latn	2.92B	0	1.07B	13M
Moroccan Ar.	ary_Arab	3.29B	0	1.07B	–
Somali	som_Latn	1.78B	0	1.07B	14M
Amharic	amh_Ethi	989M	1	1.07B	24M
Egyptian Ar.	arz_Arab	953M	1	1.07B	–
Hausa	hau_Latn	500M	2	1.07B	13M
Kinyarwanda	kin_Latn	481M	2	1.07B	13M
Zulu	zul_Latn	350M	3	1.07B	12M
Igbo	ibo_Latn	318M	3	1.07B	13M
Plateau Malagasy	plt_Latn	310M	3	1.07B	14M
Xhosa	xho_Latn	268M	3	1.07B	15M
Shona	sna_Latn	263M	4	1.05B	11M
Yoruba	yor_Latn	258M	4	1.03B	17M
Nyanja	nya_Latn	230M	4	921M	11M
Southern Sotho	sot_Latn	203M	4	813M	14M
Tigrinya	tir_Ethi	142M	4	569M	76M
Tunisian Ar.	aeb_Arab	137M	4	547M	–
Oromo	gaz_Latn	93M	4	372M	22M
Tswana	tsn_Latn	92M	4	368M	16M
subtotal				*22.8B*	*324M*
CornStack-Python (Suresh et al., 2025) (Code)					*967M*
FineMath (Allal et al., 2025) (Math)					*1.07B*
NLLB-OPUS (NLLB Team et al., 2022) (Parallel)					*456M*
Total Tokens — CM					**24.9B**
— CMS					**25.2B**
— CMSP					**25.6B**

Table 1: **Token distribution for the 24 languages pre-trained**. High-resource languages are capped at 1B tokens. Syn. denotes synthetic data.

Code (C) and Mathematics (M) Reasoning and logical abilities are often weaker in models adapted to low-resource languages. To strengthen these skills, we incorporate approximately 1B tokens of Python code from CornStack (Suresh et al., 2025) and approximately 1B tokens of educational mathematics content from FineMath-4+ (Allal et al., 2025). We also hypothesize that such structured data acts as a cognitive anchor during CPT. Maintaining a substantial fraction of code and math may help preserve internal consistency and reduce the loss of previously acquired capabilities that can occur when adaptation data is dominated by noisy monolingual web text (Yang et al., 2025a; Bakouch et al., 2025; allal et al., 2025a).

Synthetic Data (S) We enrich our training corpus with 324M tokens of machine-translated content drawn from diverse web domains and mathematical reasoning questions to increase topical coverage. Following the domain-centric curation framework of Wettig et al. (2025), we select 10 domains from Web Organizer (Wettig et al., 2025), which span 20 topics. This design serves two goals. First, it introduces high-quality lexical and conceptual coverage for domains that are sparse in many African-language corpora. Second, it functions as a form

of distributional replay buffer (Gupta et al., 2023a): translating high-quality English sources into the target languages helps preserve broad, general-purpose knowledge and stabilizes continued pre-training by keeping the training distribution closer to that of high-resource pre-training.

We use GPT-4.1 for translation due to its strong performance on AfroBench. We translate the selected documents into 17 African languages, excluding Arabic dialects because they are already well represented in our corpus. The resulting translated dataset spans the 10 domains of Food and Dining, Health, History, Industrial, Politics, Science and Technology, Software Development, Travel, Education and Jobs, and Entertainment. In addition, we translate mathematical reasoning questions, thinking traces and solutions from OpenMathReasoning (Moshkov et al., 2025) (the `cot` split) and include them as an eleventh domain.

Translation Data (P) To refine cross-lingual alignment, we explored the integration of parallel data from the NLLB project (NLLB Team et al., 2022). Although we initially collected 1B bilingual pairs, quality control was paramount. We applied a rigorous filtering threshold of 0.7 using SSA-COMET (Li et al., 2025a)—a regression model for machine translation (MT) quality estimation (QE) specifically optimized for African languages. This process yielded a high-quality subset of 4M samples (approx. 456M tokens), ensuring that only the most reliable translation pairs contributed to the model's multilingual capabilities.

Overall, our curation process yields several dataset mixtures, like CMS and CMSP, totaling 25.2B and 25.6B tokens respectively, with the detailed per-language distribution across all 24 training languages presented in Table 1 and further elaborated in Appendix D.

3.2 Training Setup

Experiments were conducted using the LLaMA-Factory (Zheng et al., 2024) framework on a high-performance cluster (up to 16 nodes, 64 NVIDIA H100 GPUs). We maximized training throughput and memory efficiency by employing sequence packing, DeepSpeed ZeRO-1/ZeRO-2 (Rasley et al., 2020), Flash Attention 3 (Shah et al., 2024), and Liger Kernel (Hsu et al., 2025).

Hyperparameter Tuning Following the continual pre-training strategies of Gupta et al. (2023b) and Bakouch et al. (2025), we performed an extensive ablation study to tailor hyperparameters for the African language context. Our search yielded three key insights based on the `gemma-3-4b/12b-pt`:

1. **Learning Rate:** A sweep from 1e-6 to 2e-4 revealed that 5e-5 optimally balances the retention of prior knowledge with the acquisition of new linguistic features.
2. **Context Length:** Evaluating window sizes of 4k, 16k, and 32k tokens, we found that the 16k sequence length provided the best performance on reasoning tasks such as AfriMGSM.
3. **Learning Rate Scheduler:** We fine-tuned the cosine scheduler, setting a minimum learning rate ratio of 0.01 and a warmup ratio of 0.001 to ensure training stability.

We maintained a global batch size of 4M tokens across all runs, dynamically adjusting gradient accumulation steps to accommodate varying hardware configurations. Full configuration details and grid search results are available in Appendix B.2.

4 Evaluation Setting

We use a comprehensive evaluation suite to assess model performance across Africa's diverse linguistic landscape. Our primary benchmark is AfroBench (Ojo et al., 2025), which covers 64 languages across 15 tasks.

AfroBench-Lite To facilitate efficient yet comprehensive evaluation, we focus on the AfroBench-Lite subset, which selects 7 representative tasks/-datasets covering key capabilities: AfriMGSM (Math), AfriMMLU (Knowledge), AfriXNLI (natural language inference) (Adelani et al., 2024c), Belebele (Reading Comprehension) (Bandarkar et al., 2024), Flores (Translation) (Goyal et al., 2022), Injongo (Intent Classification) (Yu et al., 2025), and SIB (topic classification) (Adelani et al., 2024b). While the original AfroBench-Lite evaluated on only 14 languages, we expanded the coverage of our evaluation to all African languages covered in each dataset/task.

Metrics We strictly adhere to the `lm-eval` (Gao et al., 2024) tasks established by AfroBench to ensure comparability. Note that as our models are pre-trained checkpoints without any instruction tuning, we report few-shot (5-shots) results for all tasks except AfriMGSM (where the default setting is 8-shots). For translation tasks (Flores), we utilize SSA-COMET (Li et al., 2025a) rather than lexical overlap metrics like ChrF++ (Popović, 2017)

from the official AfroBench since recent studies indicate that SSA-COMET correlates significantly better with human judgment for African languages, offering a more accurate assessment of semantic quality (Li et al., 2025a). All evaluations utilize the Hugging Face or vLLM backend (Kwon et al., 2023) with "do_sample=False".

Baseline Models To evaluate the effectiveness of our data mixture and scaling laws, we selected several state-of-the-art open-weight models as baselines: the Google Gemma 3 series (Team et al., 2025b), Meta Llama 3.1 series (Team, 2024), and Alibaba Qwen 3 series (Yang et al., 2025b). Gemma 3 is renowned for its extensive multilingual support, while Llama 3.1 represents a highly optimized predecessor in the open-source landscape. We included the Qwen 3 series due to its strong performance in mathematical reasoning, despite its limited native support for African languages.[1] Our experimental pipeline first validates the data mixture using Gemma 3 (4B and 12B) and subsequently scales these findings to Llama 3.1 (8B) and Qwen 3 (7B and 14B) base models.

5 Experiments Results

5.1 Data Mixture Ablation

To identify the optimal recipe for African language adaptation, we perform an ablation study on Gemma 3 (4B and 12B), evaluating four benchmarks: Flores (MT), AfriXNLI (NLI), AfriMGSM (Math), and AfriMMLU (QA). Results are shown in Table 2.[2]

The Monolingual Trade-off Adding only monolingual data (**M**, 22B tokens) yields substantial gains on non-reasoning tasks, with MT (Flores) and NLI (AfriXNLI) improving by over 10% relative to the base model. However, on challenging reasoning datasets (AfriMGSM and AfriMMLU), performance declines slightly (e.g., 10.5 → 9.6 on 4B MGSM). We attribute this to catastrophic forgetting of reasoning priors when exposed to large volumes of raw web text that are less heterogeneous for low-resource languages.

Performance Recovery via Code and Math Integrating additional 2B tokens of Code and Math (CM) reverses this trend. For both models, CM improves performance across all tasks compared to monolingual data only, demonstrating the importance of adding datasets with structured reasoning such as Code. This finding aligns with prior work showing that CM enhances generalization to other tasks (allal et al., 2025b; Aryabumi et al., 2025).

Model	Flores	AfriMGSM	AfriMMLU	AfriXNLI
Baseline Models				
NLLB-200-1.3B	61.27	–	–	–
NLLB-200-3.3B	62.42	–	–	–
NLLB-MoE-54B	65.72	–	–	–
Gemma 3 4B PT	35.99	9.25	33.57	34.77
Gemma 3 4B IT	31.86	14.29	34.44	33.19
Gemma 3 4B Variants				
+ Monolingual (M)	62.72 ↑	10.68 ↑	35.41 ↑	**40.76** ↑
+ CM	62.30 ↑	14.68 ↑	36.08 ↑	40.19 ↑
+ CMP	63.21 ↑	14.29 ↑	35.20 ↑	40.10 ↑
+ CMS	63.17 ↑	**14.81** ↑	35.86 ↑	39.93 ↑
+ CMSP	**63.34** ↑	13.35 ↑	**36.72** ↑	40.44 ↑
Gemma 3 12B Variants				
Base (PT)	52.53	24.10	48.21	39.81
Gemma 3 12B IT	47.81	36.50	46.84	40.16
+ Monolingual (M)	65.78 ↑	23.78 ↓	46.72 ↓	**45.19** ↑
+ CMP	65.86 ↑	27.82 ↑	**48.49** ↑	42.45 ↑
+ CMS	**66.23** ↑	**30.87** ↑	48.46 ↑	44.57 ↑
+ CMSP	65.83 ↑	29.61 ↑	48.32 ↑	43.26 ↑

Table 2: **Ablation of CPT Data mixture**. We report result on African languages covered in CPT. C = Code, M = Math, S = Synthetic, P = Parallel. Best results among adapted models are in **bold**, and our final configurations are highlighted in green. Compare to base model: improvement with ↑ and degradation with ↓.

Data Quality vs. Scale At 12B scale, we observe a divergence regarding parallel data (P). While NLLB parallel data (CMP) provides marginal gains for the 4B model, it becomes detrimental for the 12B model compared to CMS. Specifically, CMS (Monolingual + Code/Math + Synthetic) achieves the highest scores on MGSM (27.5) and Flores (67.2), whereas adding parallel data (CMSP) causes performance reduction. Drawing from mid-training recipes in Zheng et al. (2025); Bakouch et al. (2025), we hypothesize that larger models are more sensitive to data quality: noisy parallel corpora like NLLB, even when filtered, benefit smaller models but harm larger ones. Accordingly, we adopt **CMS** as our primary recipe.

▷ Takeaway 1: The Quality-Scale Sensitivity

Leveraging synthetic data and datasets with structured reasoning such as Math & Code improves CPT generalization. For larger models (12B+), high-quality synthetic data is a more effective bridge than noisy parallel corpora.

[1] Qwen 3 supported languages

[2] For the data mixture ablation study, we use the HuggingFace backend with `lm-eval` for accuracy, while all other benchmarks use vLLM to reduce computation cost. As a result, relative trends are consistent, but absolute scores may differ between Table 2 and Table 3.

Model	AfriMGSM	AfriMMLU	AfriXNLI	Belebele	Flores	Injongo	SIB-200	Overall	Δ	Δ %
African Languages Adapted										
Lugha-Llama-8B-wura	9.46	37.00	39.24	47.86	48.27	62.30	75.81	45.71	-	-
Base Models										
Llama 3.1 8B	8.14	32.27	37.90	40.95	23.59	41.37	59.99	34.89	-	-
Gemma 3 4B	10.24	33.89	37.76	45.79	29.50	55.52	63.59	39.47	-	-
Gemma 3 12B	25.21	48.76	44.01	68.84	40.16	73.53	79.17	54.24	-	-
Qwen 3 8B	11.22	36.56	38.24	44.63	18.93	29.47	53.06	33.16	-	-
Qwen 3 14B	16.60	39.66	43.22	50.74	20.86	41.80	66.29	39.88	-	-
Afrique Models (Ours)										
AfriqueLlama-8B	17.51	36.57	37.39	50.51	64.88	71.17	69.14	49.60	+14.7	+42.2%
AfriqueGemma-4B	14.86	36.73	39.62	50.52	57.31	69.28	69.21	48.22	+8.7	+22.2%
AfriqueGemma-12B	32.14	49.47	44.60	68.65	**66.89**	76.79	75.08	59.09	+4.8	+8.9%
AfriqueQwen-8B	39.68	46.91	45.99	68.46	63.54	73.36	77.00	59.28	**+26.1**	**+78.8%**
AfriqueQwen-14B	**45.01**	52.22	**49.01**	74.63	65.26	77.80	82.63	**63.79**	+23.9	+60.0%
Gemma 3 27B	35.37	**55.47**	46.85	**74.81**	45.77	**79.70**	**84.34**	60.33	-	-

Table 3: **Task-level performance comparison between Base models and our Continued Pre-Trained (CPT) models**. Best results are **bolded**, second-best are underlined. ΔAbs and ΔRel show absolute and relative improvements over base models, with purple highlighting Qwen's superior gains.

5.2 Impact of Model Selection and Scaling

Table 3 shows the result of leveraging the CMS recipe across various model architectures and model sizes.

"Zero-to-Hero" Effect in Qwen 3 The most striking finding is the performance jump in Qwen 3 with a relative improvement of 78.8% over the base model while the Gemma 3 series achieved only 35.4% relative improvement, which we term the "Zero-to-Hero" effect. Despite minimal official support for African languages and the worst baseline performance (Qwen 3 8B avg.: 33.2), AfriqueQwen exhibits the highest relative gains, outperforming similarly-sized AfriqueGemma variants on all tasks except translation (Flores), where Gemma's native multilingual pre-training provides an expected advantage. And even notably, AfriqueQwen-14B (63.79) outperforms Gemma 3 27B (60.33) by +3.46 points overall, with significant advantages on AfriMGSM (+9.64) and Flores (+19.49), despite being less than half the size.

These results suggest that Qwen 3 models largely preserve their High-Resource Languages (HRLs) performance when adapted to Low-Resource Languages (LRLs) via CPT. Consistent with the Qwen 3 technical report, Qwen 3 14B outperforms Gemma 3 12B on HRLs. We hypothesize that Qwen 3 benefits from stronger latent fast adaptation capabilities that are more effectively unlocked through CPT, [3] highlighting that a strong HRL base model priors are more critical for cross-lingual adaptation than prior language familiarity.

Comparison with Other CPT African LLMs Compared to **Lugha-Llama-8B-wura** (Buzaaba et al., 2025b), adapted using only WURA monolingual data (Oladipo et al., 2023) on the same Llama 3.1 8B base, AfriqueLlama shows cost score to Lugha, and outperforms it in 4 of 7 tasks, particularly reasoning (MGSM: 17.51 vs. 9.46) and translation (Flores: 64.88 vs. 48.27).

Marginal Effects of Model Size As expected, relative improvement from CPT decreases with model size (Gemma 4B: +35.4% vs. 12B: +15.5%) with the same training data mixture, consistent with scaling laws in prior work (Ye et al., 2024; He et al., 2024). However, even at 14B parameters, Qwen shows substantial gains (+60.0%), indicating significant headroom for African language adaptation.

▷ Takeaway 2: Foundation Ability Matters

A base model's "strong ability" is a more potent starting point for CPT than its "language coverage." Strong foundation ability priors can be effectively mapped to new languages with high-quality data mixture.

5.3 Language-wise Analysis

Here, we analyze the impact of CPT across three language resource levels (Figure 1): High-Resource Pre-Trained (HRL-PT) language — English that is well-represented in base model pre-training; African Pre-Trained (Afr-PT) languages included in our CPT corpus (e.g., Swahili, Amharic); and African Non-Pre-Trained (Afr-NPT) languages absent from both base and CPT training (e.g., Ewe, Lingala).

Figure 1 reveals three key findings: *(1) Targeted gains on Afr-PT languages.* All models show substantial improvements on CPT-covered African languages, with Qwen 3 8B achieving the highest gain

[3] Probably because it was pre-trained on 119 languages

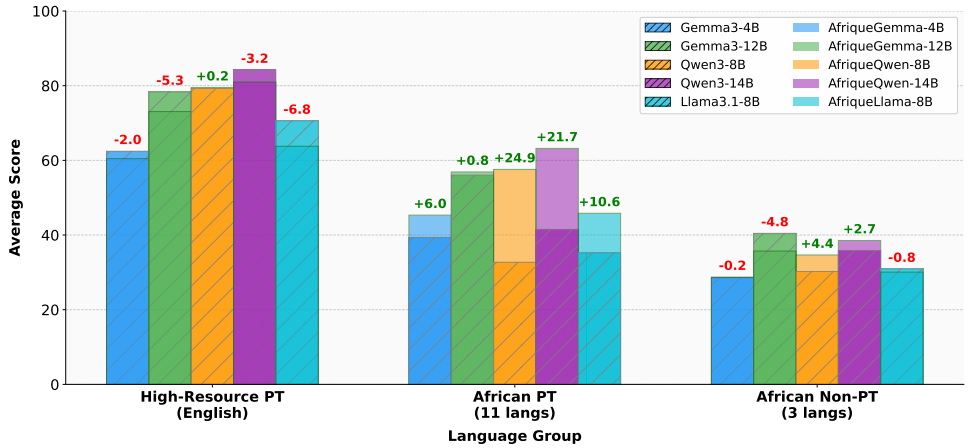

Figure 1: **Performance comparison across language groups: High-Resource PT (English), African PT, and African Non-PT**. We report the average score across all benchmarks excluding Flores. Hatched bars represent base models, while solid bars represent their Afrique-adapted counterparts. Values above the bars indicate the absolute improvement (Δ) after adaptation.

	Gemma3-4B			Gemma3-12B			Qwen3-8B			Qwen3-14B			Llama3.1-8B		
Language	Base	Afrique	Δ%	Base	Afrique	Δ%	Base	Afrique	Δ%	Base	Afrique	Δ%	Base	Afrique	Δ%
English	59.1	56.6	-4.2%	76.9	71.2	-7.4%	78.1	78.7	+0.8%	83.4	79.9	-4.1%	68.1	60.6	-11.0%
French	49.6	45.5	-8.3%	66.5	64.0	-3.8%	73.9	71.0	-4.0%	76.7	74.1	-3.5%	55.0	49.8	-9.4%
Avg.	54.3	51.1	-6.2%	71.7	67.6	-5.6%	76.0	74.8	-1.6%	80.0	77.0	-3.8%	61.5	55.2	-10.2%

Table 4: **High-Resource Language (HRL) performance comparison between base models and Afrique-adapted models**. Red values indicate performance drops, highlighting the trade-off when adapting models for African languages. Δ% denotes relative difference.

(+24.9 points). *(2) Minimal transfer to unseen languages.* Performance on Afr-NPT languages remains largely unchanged for most models, indicating that CPT primarily benefits explicitly covered languages. Interestingly, AfriqueQwens show modest positive transfer (+4.4, +2.7), suggesting that the CPT models leverage cross-lingual transfer from related languages from same family e.g. Lingala could benefit from other Bantu languages (like Swahili & Kinyarwanda) even when not covered. *(3) Less catastrophic forgetting for HRLs* While most models exhibit HRL decline, Qwen 3 8B maintains near-parity (+0.2), demonstrating that with a strong HRL base, CPT can enhance low-resource languages without sacrificing too much high-resource performance. This is further supported by AfriqueQwen-14B's +10.6 gain on Afr-PT languages with only -3.2 loss on HRL. Overall, the mixture of prior model capability and CPT data composition allows for balancing improvements in LRLs while controlling degradation in HRLs.

▷ Takeaway 3: Language Transfer Limits

"CPT favours seen languages in data mixtures." Including HRLs in the CPT data mixture mitigates catastrophic forgetting on HRLs, but yields limited transfer to unseen languages.

HRL degradation across models Table 4 quantifies catastrophic forgetting on English and French. Compared to the massive African language gains (up to +78.8% in Table 3), HRL performance drops are contained but noteworthy. Llama 3.1 8B shows the steepest average decline (-10.2% relative), followed by Gemma 3 models (-5.6% to -6.2%). In contrast, the Qwen 3 series exhibits the smallest average HRL degradation (-1.6% for 8B, -3.8% for 14B), showing they are slightly better in preventing catastrophic forgetting.

Granular Analysis on Afr-PT Languages Table 5 provides a detailed breakdown across 11 CPT-covered African languages, averaged across tasks. Gains are consistent across all languages and models, with **many low-resource languages benefiting most**: Oromo (orm) and Yoruba (yor) show the highest deltas (e.g., +16.1 and +17.2 for AfriqueGemma-4B). In contrast, Swahili (swa) shows more modest gains (+1.7 to +20.8). For Qwen 3, improvements are even more astounding: AfriqueQwen-8B exceeds +25 absolute points in 10 of 11 languages, peaking at +37.1 in Hausa. This confirms our hypothesis that previously underrepresented languages benefit most from our data mixture. The more detailed results across all

Model	amh	hau	ibo	kin	orm	sna	sot	swa	xho	yor	zul	Avg.
Llama 3.1 8B	29.0	41.0	37.5	29.2	25.8	28.7	28.0	48.7	28.7	29.5	28.9	32.3
AfriqueLlama-8B	47.9	50.3	47.2	47.5	45.3	50.1	48.0	55.0	48.6	47.4	47.0	48.6
Δ	+18.9	+9.3	+9.8	+18.3	+19.5	+21.5	+20.1	+6.3	+19.9	+17.9	+18.1	+16.3
Gemma 3 4B	43.3	42.6	37.3	36.3	26.4	38.2	33.2	52.4	38.0	26.6	38.8	37.6
AfriqueGemma-4B	48.7	50.0	46.4	46.9	42.5	49.8	43.0	54.1	48.4	43.8	46.5	47.3
Δ	+5.4	+7.4	+9.1	+10.6	+16.1	+11.6	+9.7	+1.7	+10.4	+17.2	+7.7	+9.7
Gemma 3 12B	59.9	57.8	52.8	52.7	41.4	56.7	51.1	66.8	53.2	43.6	53.7	53.6
AfriqueGemma-12B	60.8	60.4	55.8	56.1	54.5	59.8	58.2	66.3	58.0	54.6	57.5	58.3
Δ	+0.9	+2.6	+3.0	+3.4	+13.1	+3.0	+7.0	-0.5	+4.8	+11.0	+3.8	+4.7
Qwen 3 8B	34.6	24.6	25.6	24.9	28.4	27.4	28.3	47.3	27.2	24.8	25.8	29.0
AfriqueQwen-8B	61.0	61.7	54.8	56.8	55.1	59.6	57.3	68.2	57.2	55.0	56.4	58.5
Δ	**+26.3**	**+37.1**	**+29.2**	**+31.9**	**+26.8**	**+32.2**	**+29.0**	**+20.8**	**+30.0**	**+30.2**	**+30.6**	**+29.5**
Qwen 3 14B	42.0	32.2	32.9	31.0	35.4	33.3	35.8	58.7	36.7	33.7	35.1	37.0
AfriqueQwen-14B	64.7	66.1	61.0	61.0	62.0	64.5	62.4	73.0	61.8	61.2	61.5	63.6
Δ	+22.7	+34.0	+28.1	+30.0	+26.7	+31.2	+26.6	+14.3	+25.1	+27.5	+26.4	+26.6

Table 5: **Language-wise average performance improvement across all benchmarks on CPT covered languages**. Green values indicate gains after Afrique adaptation. **Bold** and underline denote the best and second-best improvements per language.

Model	amh	hau	swa	yor	zul	Avg.
English → African (eng2xx)						
Llama 3.1 8B SFT_{10}	27.6	49.7	64.1	**50.3**	47.0	47.8
Llama 3.1 8B	10.3	19.5	28.7	16.7	14.2	17.9
AfriqueLlama-8B	41.4	62.0	74.4	46.3	68.1	58.5
AfriqueGemma-12B	**42.1**	**64.2**	**78.1**	47.0	**69.8**	**60.2**
AfriqueQwen-14B	42.0	62.8	75.7	47.4	68.2	59.2
African → English (xx2eng)						
Llama 3.1 8B SFT_{10}	63.8	61.7	74.4	68.9	71.4	68.0
Llama 3.1 8B	20.0	53.9	71.2	30.7	37.0	42.6
AfriqueLlama-8B	44.7	58.2	66.6	53.5	63.4	57.3
AfriqueGemma-12B	72.7	67.7	**80.5**	68.8	**76.6**	73.3
AfriqueQwen-14B	**72.8**	**68.3**	79.7	**70.8**	76.1	**73.5**

Table 6: **Document-level translation (d-chrF,** $k = 10$**)** on AFRIDOC-MT health domain with 3-shot prompting. *Baseline*: Llama 3.1 8B SFT_{10}—fine-tuned on 4K documents from AFRIDOC-MT (Alabi et al., 2025).

languages and tasks are presented in Appendix D.

5.4 Document-Level Translation

To evaluate whether our models with 16K tokens sequence length improve long-context translation, we benchmark on AFRIDOC-MT (Alabi et al., 2025) (health domain), a document-level parallel corpus covering English and five African languages (Amharic, Hausa, Swahili, Yoruba, Zulu). We use pseudo-documents with $k = 10$ sentences and report document-level chrF (d-chrF) scores with 3-shot prompting. Table 6 shows that all AfriqueLLMs excel at document-level translation despite never seeing AFRIDOC-MT training data during CPT. We compare performance to Llama 3.1 SFT_{10} baseline that was instruction fine-tuned on 4,060 health documents (812 per language pair).

For *eng→xx*, AfriqueGemma-12B achieves the best average (60.2), outperforming the task-specific SFT model (47.8) by +12.4 points. AfriqueQwen-14B (59.2) and AfriqueLlama-8B (58.5) also substantially exceed the SFT baseline, demonstrating that CPT provides robust long-context translation capabilities.

For *xx→eng*, AfriqueQwen-14B leads with 73.5, closely followed by AfriqueGemma-12B (73.3). Notably, both surpass the task-specific SFT_{10} model (68.0), showing that CPT's general-purpose training can even exceed in-domain fine-tuning for certain translation directions.

6 Conclusion

We introduce **AfriqueLLM**, a suite of LLMs adapted for 20 African languages via efficient CPT on 26B tokens. Our key findings are: *(1)* data mixture matters most—combining monolingual text with code, math, and synthetic data (CMS) yields state-of-the-art results while preserving reasoning; *(2)* base model strong capability trumps multilingual coverage—Qwen 3, despite minimal African language support, achieves the highest performance after CPT, with AfriqueQwen-14B (63.79) outperforming Gemma 3 27B (60.33) at less than half the size; and *(3)* high-quality synthetic data provides a scalable bridge for low-resource languages, with AfriqueQwen-14B surpassing the 54B NLLB-MoE on translation. We will release our models to advance African language AI research.

As a future work, we plan to conduct a more comprehensive exploration and analysis on why Qwen 3 series models provides such a strong improvement after CPT, than similar architectures such as Gemma 3.

7 Limitations

Scope Constraints. Our study has several coverage limitations: (1) *Language coverage*: We cover 20 African languages, leaving hundreds unsupported—languages with minimal digital presence remain challenging. (2) *Model scale*: Resource constraints limited experiments to 14B parameters; larger models (30B+) may exhibit different adaptation dynamics and potential better performance, like "Qwen3-30B-A3B-Base" and "Gemma 3 27b PT" and potential better performance, like "Qwen3-30B-A3B-Base" and "Gemma 3 27b PT". (3) *Training stage*: We focus on base model CPT without instruction tuning—the scarcity of high-quality instruction data for African languages remains a bottleneck for downstream deployment. (4) *Hyperparameters*: Scaling to 12B+ prevents exhaustive search; we relied on heuristics from smaller model, which may not be optimal across architectures.

Training Stability and Efficiency. We observed intermittent gradient norm spikes during training, suggesting latent optimization instabilities. While these did not cause divergence, future work could explore matrix optimizers like Muon (Liu et al., 2025) for improved stability. Our framework achieves 31–34% Model FLOPs Utilization (Appendix B.3), competitive for general-purpose setups but leaving room for improvement via specialized frameworks like Megatron-LM (Shoeybi et al., 2019).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report.

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating LLM performance on African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024b. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Zhuang Yun Jian, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, and 1 others. 2024c. Irokobench: A new benchmark for african languages in the age of large language models. *ArXiv*, abs/2406.03368.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022a. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022b. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jesujoba O. Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025.

Afridoc-mt: Document-level mt corpus for african languages.

Loubna Ben allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, and 3 others. 2025a. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*.

Loubna Ben allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, and 3 others. 2025b. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2025. To code or not to code? exploring impact of code in pre-training. In *The Thirteenth International Conference on Learning Representations*.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`.

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, Bangkok, Thailand. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025a. Lugha-llama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025b. Lugha-llama: Adapting large language models for african languages.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, and 5 others. 2024. A framework for few-shot language model evaluation.

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.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023a. Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023b. Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020b. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2024. Scaling laws for multilingual language models.

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

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, Yanning Chen, and Zhipeng Wang. 2025. Liger-kernel: Efficient triton kernels for LLM training. In *Championing Open-source DEvelopment in ML Workshop @ ICML25*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Indraneil Paul, Hengyu Luo, and Jörg Tiedemann. 2025a. Massively multilingual adaptation of large language models using bilingual translation data. *arXiv preprint arXiv:2506.00469*.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Indraneil Paul, Hengyu Luo, and Jörg Tiedemann. 2025b. Massively multilingual adaptation of large language models using bilingual translation data. *arXiv preprint arXiv:2506.00469*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

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. *Preprint*, arXiv:2309.04662.

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*.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025a. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12990–13009, Suzhou, China. Association for Computational Linguistics.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. 2025b. Rethinking multilingual continual pretraining: Data mixing for adapting LLMs across languages and resources. In *Second Conference on Language Modeling*.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo

Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, and 9 others. 2025. Muon is scalable for llm training.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, and 340 others. 2025. Nvidia nemotron 3: Efficient and open intelligence.

NVIDIA. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious.

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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. *Preprint*, arXiv:2506.20920.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: fast and accurate attention with asynchrony and low-precision. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

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*.

Tarun Suresh, Revanth Gangi Reddy, Yifei Xu, Zach Nussbaum, Andriy Mulyar, Brandon Duderstadt, and Heng Ji. 2025. Cornstack: High-quality contrastive data for better code retrieval and reranking. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, and 1 others. 2025b. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Llama3 Team. 2024. The llama 3 herd of models.

Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, Yifei Sun, and En-Shiun Annie Lee. 2024. AfriInstruct: Instruction tuning of African languages for diverse tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585, Miami, Florida, USA. Association for Computational Linguistics.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025a. Multilingual language model pretraining using machine-translated data.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025b. Multilingual language model pretraining using machine-translated data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28075–28095, Suzhou, China. Association for Computational Linguistics.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2024. Data mixing laws: Optimizing data mixtures by predicting language modeling performance.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Zhuang Yun Jian, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing K. Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Bridget Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, and 3 others. 2025. Injongo: A multicultural intent detection and slot-filling dataset for 16 african languages. *ArXiv*, abs/2502.09814.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Data Details

A.1 Language Selection and Statistics

Table 7 provides a comprehensive overview of the language selection process and the final token counts for each language across our primary data sources: FineWeb2, WURA, and MADLAD-400. We applied a selection threshold of 90M tokens to ensure sufficient data for meaningful linguistic adaptation.

A.2 Synthetic Data and Translation Prompts

Table 8 details the distribution of synthetic data across 11 domains. The translation process was guided by the prompts shown in Section A.2.

General Translation Prompt

```
You are a professional translator.
    Translate the user text from {{
    source_lang}} into {{
    target_lang}}.
Preserve meaning, tone, formatting,
     inline markup, numerals, and
    named entities exactly.
For long texts, ensure the
    translation is fluent, coherent
     and complete. Make sure to
    translate all parts of the text
    . Return only the translation
    without additional commentary.
```

Mathematical Reasoning Translation Prompt

```
You are a {{source_lang}}-to-{{
    target_lang}} translator for
    mathematical content. Translate
     the provided math problem,
    reasoning, and answer while
    preserving:
- All numbers, formulas, and
    formatting
- Mathematical notation and markup
- Named entities and tone

Input structure:
<problem>[Original Problem]</
    problem>
<think>[Original Reasoning]</think>
[Final Answer]<eos>

Output structure:
<problem>[Translated problem]</
    problem>
<think>[Translated reasoning]</
    think>
[Translated Final Answer]<eos>

Ensure translations are fluent,
    coherent, and complete. Return
    only the translation without
    additional commentary.
```

B Training Details

B.1 Hyperparameter Search

We conducted an extensive ablation study to identify the optimal hyperparameters for continued pretraining on African languages.

Learning Rate We performed a learning rate sweep on the Gemma 3 4B PT model with rates ranging from 1e-6 to 2e-4. Table 9 identifies 5e-5 as the optimal rate based on average scores across low-resource languages.

Context Size Using the optimal learning rate (5e-5), we evaluated context lengths of 4k, 16k, and 32k. Table 10 shows that a 16k context window yields the best performance on AfriMGSM.

Cosine Scheduler We explored the impact of the minimum learning rate (min lr) and warmup steps. Table 11 presents the results using the Gemma 3 4B pretrained model with a fixed context size of 16k.

B.2 Training Configuration

The following YAML configuration was used for the continued pre-training of the AfriqueLLM models using the LLaMA-Factory framework.

LLaMA-Factory CPT config (YAML)

```yaml
### model
model_name_or_path: google/gemma-3-
    12b-pt

### method
stage: pt

### data
template: empty
packing: true
cutoff_len: 16384 # 16k
overwrite_cache: false
preprocessing_num_workers: 32
dataloader_num_workers: 32

### finetuning
do_train: true
finetuning_type: full
deepspeed: ds_z1_config.json
freeze_vision_tower: true
freeze_multi_modal_projector: true
freeze_language_model: false

### output
logging_steps: 10
save_steps: 1000
plot_loss: true
overwrite_output_dir: true
save_only_model: false
report_to: wandb
```

Language	Code	FineWeb2	Wura	Madlad400	Total Token	Rep.	Unimax Token	Synthetic	Other
***High-Resource (Non-African)* — Capped at 1B tokens**									
English	eng_Latn	>1000000000	865600280	–	1000000000	1×	1070793848	16,011,265	
French	fra_Latn	>1000000000	815336425	–	1000000000	1×	1070793848		
Portuguese	por_Latn	>1000000000	531069643	–	1000000000	1×	1070793849		
Arabic	arb_Arab	>1000000000	–	–	1000000000	1×	1070793848		
***African Languages* — Included in Training**									
Afrikaans	afr_Latn	2461214686	1357859486	1483495285	5302569457	1×	1070793849	12,113,273	
Swahili	swh_Latn	1051220388	1087449729	777825674	2916495791	1×	1070793849	12,503,168	
Moroccan Arabic	ary_Arab	3289564375	–	–	3289564375	1×	1070793849		
Somali	som_Latn	732191814	702650753	346518006	1781360573	1×	1070793849	13,572,904	
Amharic	amh_Ethi	403784914	276855513	308253510	988893937	2×	1070793848	23,943,363	
Egyptian Arabic	arz_Arab	821465539	131515140	–	952980679	2×	1070793848		
Hausa	hau_Latn	–	288353911	211672798	500026709	3×	1070793848	12,596,581	
Kinyarwanda	kin_Latn	136010710	69028912	275720285	480759907	3×	1070793848	12,707,048	
Zulu	zul_Latn	159037587	97653578	92982744	349673909	4×	1070793848	12,366,125	
Igbo	ibo_Latn	140734354	68796722	108189914	317720990	4×	1070793848	12,671,171	
Plateau Malagasy	plt_Latn	310443854	–	–	310443854	4×	1070793848	14,002,182	
Xhosa	xho_Latn	119027393	41737419	107219367	267984179	4×	1070793848	14,846,741	
Shona	sna_Latn	95516967	76561301	90581980	262660248	5×	1050640992	10,971,211	
Yoruba	yor_Latn	90126934	68250903	99303113	257680950	5×	1030723800	17,152,436	
Nyanja	nya_Latn	137607319	92652643	–	230259962	5×	921039848	11,481,563	
Southern Sotho	sot_Latn	122964390	–	80276553	203240943	5×	812963772	13,573,191	
Tigrinya	tir_Ethi	100865939	8661533	32703052	142230524	5×	568922096	75,525,088	
Tunisian Arabic	aeb_Arab	136652951	–	–	136652951	5×	546611804		
West Central Oromo	gaz_Latn	42916258	17619689	32493752	93029699	5×	372118796	21,619,016	
Tswana	tsn_Latn	9244373	72533425[4]	10215596	91993394	5×	367973576	16,313,360	
Additional Training Data									
FineMath (Math, M)	–	–	–	–	–	–	–	–	1,067,549,046
CornStack-Python (Code, C)	–	–	–	–	–	–	–	–	967,399,767
MT-NLLB (Parallel, P)	–	–	–	–	–	–	–	–	456,102,720
Subtotal of Tokens					**22.88B**		**22.80B**	**0.32B**	**2.49B**
***Excluded Languages* (<90M tokens)**									
Rundi	run_Latn	56775951		492969	57268920				
Ganda	lug_Latn	24162781		18022976	42185757				
Tsonga	tso_Latn	10782436		14451048	25233484				
Lingala	lin_Latn	16358800		7530450	23889250				
Ewe	ewe_Latn	3014541		15388319	18402860				
Wolof	wol_Latn	16527037		1839642	18366679				
Sango	sag_Latn	7104619		5590802	12695421				
Akan	aka_Latn			10824690	10824690				
Twi	twi_Latn	10648719			10648719				
Kabiye	kbp_Latn	1040478		8959130	9999608				
Bambara	bam_Latn	7335041		1426843	8761884				
Northern Sotho	nso_Latn	8630368			8630368				
Fon	fon_Latn	2281350		4439623	6720973				
Swati	ssw_Latn	2660736		2016953	4677689				
Tamazight	tzm_Tfng	4044801		260465	4305266				
Kabyle	kab_Latn	3860016			3860016				
Kabuverdianu	kea_Latn	3782732			3782732				
N'ko	nqo_Nkoo	3717948			3717948				
Mossi	Mos_Latn	3319912			3319912				
Kimbundu	kmb_Latn	1506689		1759056	3265745				
Kanuri (Arabic)	knc_Arab	3105431			3105431				
Dyula	dyu_Latn	2018490		960718	2979208				
Tamasheq (Latin)	taq_Latn	2640160			2640160				
Southwestern Dinka	dik_Latn	1144214		1420754	2564968				
Luo	luo_Latn	2010521			2010521				
Nigerian Fulfulde	fuv_Latn	1894553		95651	1990204				
Bemba	bem_Latn	1482559			1482559				
Kikuyu	kik_Tatn	1411871			1411871				
Kamba	kam_Latn	1018287			1018287				
Kikongo	kon_Latn			971858	971858				
Luba-Kasai	lua_Latn	908010			908010				
Umbundu	umb_Latn	540735			540735				
Tamasheq (Tifinagh)	taq_Tfng	401256			401256				
Kanuri (Latin)	knc_Latn	256317			256317				
Tumbuka	tum_Latn	228626			228626				
Nuer	nus_Latn	224103			224103				
Chokwe	cjk_Latn	33366			33366				
Non-Training Languages Subtotal					**303,325,401**				

Table 7: Complete dataset collection and language selection for training. This table presents all 60+ African and high-resource languages collected from FineWeb2, Wura, and Madlad400 sources, along with the selection criteria applied. Languages with ≥90M tokens are included in the final training set (24 languages, 22.8B tokens). The "Rep." column indicates the upsampling factor applied via UniMax to balance low-resource languages. Grayed rows indicate excluded languages due to insufficient data.

Domain	Tokens
Math	32,284,225
Science & Tech.	37,461,084
Politics	35,256,194
Health	31,213,028
Travel	29,751,012
History	28,386,610
Food & Dining	27,556,953
Education & Jobs	27,469,250
Software Dev.	26,446,379
Entertainment	25,148,472
Industrial	22,996,479
Total	**323,969,686**

Table 8: Synthetic Data domain distribution. Math data is sourced from (Moshkov et al., 2025), while other domains are from (Wettig et al., 2025).

LR	AfriMGSM	AfriXNLI	AfriMMLU	Flores
2e-4	5.1	39.0	27.5	-
1e-4	8.1	38.6	31.0	-
5e-5	**9.4**	**40.6**	34.7	**61.1**
2e-5	8.4	40.5	36.0	59.4
1e-5	8.0	39.9	**37.2**	57.5
5e-6	8.6	39.4	36.7	54.6
2e-6	9.0	37.8	36.8	50.3
1e-6	8.7	38.7	36.4	46.4

Table 9: Ablation on learning rate. Results are reported for the Gemma 3 4B pretrained model with a fixed 16k context length. We report average scores for AfriMGSM (8-shot CoT), AfriXNLI (Direct), AfriMMLU (Direct), and Translation (SSA-COMET) excluding English and French.

```
data_shared_file_system: true

### train
per_device_train_batch_size: 4
gradient_accumulation_steps: 8
learning_rate: 5.0e-5
num_train_epochs: 1.0
lr_scheduler_type:
    cosine_with_min_lr
lr_scheduler_kwargs:
  min_lr_rate: 0.01
warmup_ratio: 0.001
bf16: true
ddp_timeout: 180000000
resume_from_checkpoint: null
weight_decay: 0.1
adam_beta1: 0.9
adam_beta2: 0.95

enable_liger_kernel: true
flash_attn: fa3
```

Context Length	AfriMGSM
4k	7.5
16k	**9.4**
32k	7.8

Table 10: Ablation on context length. Results are reported for the Gemma 3 4B pretrained model with a fixed 5e-5 learning rate. We report average scores for AfriMGSM (8-shot CoT), excluding English and French.

Min lr	Warmup	AfriMGSM	AfriXNLI	AfriMMLU	Flores
0.01	0	9.4	**38.8**	**34.1**	60.6
0.01	0.001	**10.2**	38.2	**34.1**	60.5
0.1	0	7.7	38.5	34.0	60.9
0.1	0.001	9.4	38.5	33.4	**61.1**

Table 11: Ablation on cosine scheduler hyperparameters. Results are reported for the Gemma 3 4B pretrained model with a 16k context window. We report average scores for AfriMGSM (8-shot CoT), AfriXNLI (Direct), AfriMMLU (Direct), and Translation (SSA-COMET) excluding English and French.

B.3 Training Efficiency Analysis

Table 12 summarizes the computational metrics for our continued pre-training process.

Model	Steps	Tokens	Time (h)	TFLOPS	MFU (%)	Loss
Gemma 3 4B	6,008	25.2B	9.12	16,690	26.37	0.5098
Gemma 3 12B	6,008	25.2B	23.70	19,776	31.24	1.2942
Qwen 3 8B	6,872	27.5B	18.30	19,868	31.39	1.4338
Llama 3.1 8B	7,406	29.6B	18.06	21,516	33.99	1.2502

Table 12: Training efficiency metrics for Continued Pre-Training (CPT). Total Tokens are calculated based on the fixed batch size of 4M tokens per step.

C Evaluation Details

C.1 Benchmark Language Coverage

Table 13 lists the languages covered in each task of the AfroBench-Lite suite.

Task	Languages (Total Counts)
AFRIMGSM	Amharic†, English*, Ewe, French*, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Oromo†, Shona†, Sotho†, Swahili†, Twi, Vai, Wolof, Xhosa†, Yoruba†, Zulu† (19)
AFRIMMLU	Amharic†, English*, Ewe, French*, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Oromo†, Shona†, Sotho†, Swahili†, Twi, Wolof, Xhosa†, Yoruba†, Zulu† (18)
AFRIXNLI	Amharic†, English*, Ewe, French*, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Oromo†, Shona†, Sotho†, Swahili†, Twi, Wolof, Xhosa†, Yoruba†, Zulu† (18)
BELEBELE	Afrikaans†, Amharic†, Egyptian Arabic†, English*, French*, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Moroccan Arabic†, Nyanja†, Oromo†, Plateau Malagasy†, Portuguese*, Shona†, Somali†, Sotho†, Swahili†, Tigrinya†, Tswana†, Wolof, Xhosa†, Yoruba†, Zulu† (25)
FLORES	Afrikaans†, Amharic†, Egyptian Arabic†, Ewe, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Moroccan Arabic†, Nyanja†, Oromo†, Shona†, Somali†, Sotho†, Swahili†, Tigrinya†, Tswana†, Tunisian Arabic†, Twi, Wolof, Xhosa†, Yoruba†, Zulu† (24)
INJONGO	Amharic†, English*, Ewe, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Oromo†, Shona†, Sotho†, Swahili†, Twi, Wolof, Xhosa†, Yoruba†, Zulu† (17)
SIB-200	Afrikaans†, Amharic†, Egyptian Arabic†, English*, Ewe, Hausa†, Igbo†, Kinyarwanda†, Lingala, Luganda, Moroccan Arabic†, Nyanja†, Oromo†, Plateau Malagasy†, Portuguese*, Shona†, Somali†, Sotho†, Swahili†, Tigrinya†, Tunisian Arabic†, Twi, Wolof, Xhosa†, Yoruba†, Zulu† (26)

Table 13: Languages included in each benchmark task.
*: High-resource pretrained (4)
†: Pretrained African (20)

D Detailed Experimental Results

model	amh	eng	ewe	fra	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	vai	wol	xho	yor	zul
Llama3.1-8B	2.72	53.52	3.44	37.12	13.12	7.76	6.64	4.40	7.28	4.80	6.80	6.64	23.84	5.44	1.84	5.04	4.00	6.64	6.56
Lugha-Llama-8B-wura	4.72	40.88	2.00	20.32	12.16	10.24	9.76	2.72	4.56	8.16	10.08	7.68	19.28	3.28	2.88	1.92	6.40	7.44	8.16
AfriqueLlama-8B	7.84	55.52	3.12	36.88	20.96	15.04	18.96	6.40	11.52	17.52	20.72	18.80	24.48	4.08	1.68	3.92	13.04	19.52	15.76
Gemma3-4B	10.64	42.48	3.28	28.72	12.88	5.28	7.76	2.88	6.56	2.64	12.16	10.08	27.12	1.84	0.08	2.40	7.68	5.44	10.96
AfriqueGemma-4B	17.52	37.84	3.60	21.52	17.04	13.36	12.96	4.16	10.16	9.68	16.96	15.84	21.60	2.00	0.88	2.24	10.72	11.68	16.08
Gemma3-12B	38.64	72.40	6.08	50.16	26.00	22.08	22.08	13.60	19.84	14.32	29.52	20.00	46.40	5.92	1.52	4.88	17.60	13.28	27.36
AfriqueGemma-12B	36.00	68.08	4.48	57.20	34.88	30.72	24.72	7.76	19.84	28.48	31.52	33.44	47.36	6.40	0.80	2.64	26.32	26.48	33.60
Qwen3-8B	10.80	85.76	5.92	74.08	7.84	2.64	8.88	7.92	8.00	12.16	8.48	10.88	39.04	5.76	1.12	5.20	8.80	6.48	7.44
AfriqueQwen-8B	40.48	85.20	6.40	67.92	48.88	32.08	42.56	9.76	22.16	37.76	40.00	36.64	57.28	5.68	3.04	4.88	30.24	35.44	35.12
Qwen3-14B	12.88	88.00	8.80	76.56	13.68	3.60	15.68	12.08	12.96	19.04	11.52	16.16	50.40	6.64	1.92	5.84	13.92	13.84	11.92
AfriqueQwen-14B	35.28	82.24	7.68	72.24	52.32	41.44	46.96	12.24	27.36	47.12	46.80	45.20	67.04	5.68	2.88	5.52	31.84	42.32	38.80

Table 14: 8-shot performance on the AfriMGSM benchmark for multilingual grade school math. (Math)

model	amh	eng	ewe	fra	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul
Llama3.1-8B	34.16	65.56	27.48	50.64	33.84	31.72	32.80	34.84	30.84	32.24	29.84	29.24	39.08	28.48	30.20	27.76	32.08	32.20
Lugha-Llama-8B-wura	38.52	65.08	27.80	51.80	37.76	37.96	33.32	33.00	30.36	34.60	38.72	38.48	41.84	27.44	29.52	32.68	34.52	38.60
AfriqueLlama-8B	38.28	58.04	29.28	46.48	36.48	37.36	32.92	31.68	28.32	36.48	35.84	37.52	42.24	28.12	26.68	34.72	35.56	34.92
Gemma3-4B	34.40	58.00	28.24	49.68	34.08	35.36	34.04	29.92	27.48	28.00	33.88	34.04	41.48	28.92	26.04	33.84	29.00	34.72
AfriqueGemma-4B	38.80	54.48	29.16	47.56	37.92	37.04	34.72	28.04	27.88	34.16	36.56	36.24	40.40	26.64	25.08	37.80	36.24	34.12
Gemma3-12B	52.84	78.08	30.72	70.40	50.96	47.96	45.80	42.40	38.88	39.64	50.00	49.12	61.56	33.76	29.80	46.72	43.56	48.20
AfriqueGemma-12B	51.88	70.64	25.84	62.08	49.20	47.64	46.72	37.76	37.36	47.28	50.60	51.56	54.48	33.40	26.36	49.60	47.72	47.52
Qwen3-8B	40.96	77.80	33.56	69.68	34.48	35.68	32.08	41.16	31.40	37.64	36.28	35.60	42.00	33.28	33.60	34.00	36.80	36.60
AfriqueQwen-8B	56.32	78.12	31.00	67.80	48.20	45.76	40.92	39.44	33.32	46.44	45.76	47.24	52.00	31.24	31.04	43.88	43.52	46.00
Qwen3-14B	45.36	82.40	34.64	73.00	37.00	35.56	37.36	41.64	34.48	39.76	36.48	38.92	47.04	32.96	33.12	37.96	42.28	38.56
AfriqueQwen-14B	59.44	80.68	32.96	72.92	52.80	49.96	46.12	39.84	36.12	52.68	48.68	54.08	61.56	31.68	31.20	51.36	47.28	50.44

Table 15: 5-shot performance on the AfriMMLU benchmark for massive multilingual language understanding. (MMLU)

model	amh	eng	ewe	fra	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul
Llama3.1-8B	37.37	52.80	34.83	50.03	40.77	39.83	35.23	34.50	37.73	37.10	37.73	37.80	41.27	35.93	34.67	35.63	37.57	36.63
Lugha-Llama-8B-wura	38.17	50.23	35.07	48.37	40.20	41.17	35.83	34.33	36.90	39.57	38.33	40.10	41.20	35.23	34.47	39.73	39.03	38.27
AfriqueLlama-8B	37.17	43.80	32.70	42.10	37.13	38.13	35.93	32.57	34.97	37.73	36.87	37.83	37.80	33.07	31.87	39.10	36.93	36.63
Gemma3-4B	38.83	47.10	34.67	44.07	39.43	38.40	36.77	34.17	34.67	35.03	37.67	37.13	40.13	32.93	33.40	38.57	36.67	36.73
AfriqueGemma-4B	39.90	44.97	34.07	44.07	40.90	40.23	37.97	33.43	36.27	37.00	40.43	42.17	40.87	33.77	33.37	39.43	38.97	37.93
Gemma3-12B	43.23	58.07	34.30	55.23	47.70	44.87	39.70	32.43	42.63	41.97	45.80	44.80	48.73	36.27	33.33	44.90	41.90	40.50
AfriqueGemma-12B	43.47	54.20	34.10	50.67	47.27	45.57	38.47	32.80	41.43	45.93	46.00	46.07	46.60	35.43	33.83	45.30	45.03	40.93
Qwen3-8B	40.83	62.77	34.03	61.93	35.83	38.87	34.90	32.50	35.63	38.80	37.73	37.27	45.97	34.63	33.20	37.50	38.17	34.77
AfriqueQwen-8B	44.63	60.67	32.43	58.20	48.03	45.13	39.57	32.47	38.83	50.07	47.47	48.50	50.27	33.37	31.90	46.33	43.93	41.93
Qwen3-14B	43.70	66.10	35.93	64.10	43.03	43.33	37.63	32.80	38.40	45.07	43.37	42.87	50.30	35.87	33.60	41.13	45.43	39.60
AfriqueQwen-14B	48.77	60.60	34.87	58.80	49.90	49.87	41.40	34.03	42.90	51.07	51.70	49.20	52.57	34.53	32.10	48.57	49.73	46.33

Table 16: 5-shot performance on the AfriXNLI benchmark for cross-lingual natural language inference. (NLI)

model	afr	amh	ary	arz	eng	fra	hau	ibo	kin	lin	lug	nya	orm	plt	por	sna	som	sot	swa	tir	tsn	wol	xho	yor	zul
Llama3.1-8B	77.96	35.62	54.00	63.53	87.64	82.04	44.22	38.82	37.76	33.80	33.71	31.60	31.87	43.98	82.07	36.47	32.71	31.36	53.22	31.36	33.60	29.60	34.42	30.69	34.87
Lugha-Llama-8B-wura	79.16	47.40	51.11	62.47	84.51	79.87	53.98	42.13	44.60	34.58	34.18	41.18	37.04	57.98	79.60	47.04	45.89	40.44	61.96	35.80	39.16	27.89	42.82	35.42	43.82
AfriqueLlama-8B	71.31	54.62	54.49	59.69	79.31	73.78	49.16	41.49	49.51	31.87	33.80	43.69	42.00	58.09	73.47	50.31	46.69	47.40	61.53	43.56	47.96	27.13	48.60	41.40	48.24
Gemma3-4B	73.64	52.73	51.51	61.71	78.58	76.00	47.24	36.16	43.07	30.91	33.53	41.44	31.16	54.62	73.24	44.33	41.93	37.78	63.80	35.31	35.78	28.09	40.44	32.71	44.58
AfriqueGemma-4B	67.93	55.93	51.56	56.87	74.96	68.89	51.24	41.00	50.78	30.29	33.40	46.11	40.64	58.44	68.02	52.40	48.11	49.02	60.67	45.29	48.18	26.04	48.18	39.69	47.89
Gemma3-12B	90.71	76.98	77.98	82.53	92.53	90.33	73.36	55.80	70.91	43.02	49.07	63.07	52.33	79.18	89.69	71.00	69.33	64.93	86.24	51.27	55.73	32.38	67.82	50.36	68.49
AfriqueGemma-12B	86.29	74.71	70.87	76.11	89.18	85.96	68.58	55.87	70.89	35.36	44.93	62.76	60.29	79.69	84.80	69.53	66.16	67.76	79.11	62.42	65.11	28.89	66.09	55.27	66.87
Qwen3-8B	87.96	50.24	64.04	79.69	91.67	90.04	32.69	34.78	37.67	35.53	31.44	32.73	38.13	44.20	88.13	40.20	33.78	35.82	62.67	35.04	35.02	33.07	35.04	31.71	36.53
AfriqueQwen-8B	88.60	75.18	73.29	79.73	91.76	90.11	67.20	54.11	67.76	35.76	39.49	61.64	60.64	76.96	88.56	67.40	63.51	65.49	81.31	67.18	63.29	32.44	66.67	55.89	64.98
Qwen3-14B	90.78	55.42	72.27	84.78	94.56	93.16	39.87	39.29	42.31	38.42	36.58	38.44	41.69	55.31	91.53	43.93	35.13	43.69	73.33	39.53	41.27	33.82	44.18	37.44	45.40
AfriqueQwen-14B	91.04	82.33	78.47	84.69	93.40	92.22	75.49	62.71	72.67	38.02	45.38	68.24	66.24	82.62	90.93	73.56	69.49	73.82	85.58	74.09	69.11	32.87	72.51	62.80	72.58

Table 17: 5-shot performance on the Belebele reading comprehension benchmark. (RC)

model	aeb	afr	amh	ary	arz	ewe	hau	ibo	kin	lin	lug	nya	orm	sna	som	sot	swa	tir	tsn	twi	wol	xho	yor	zul
Llama3.1-8B	58.82	71.31	0.42	52.97	61.27	18.08	26.73	23.17	5.81	9.63	15.40	16.75	14.63	13.16	11.69	12.20	42.25	3.80	10.43	7.67	25.99	7.92	9.73	5.20
Lugha-Llama-8B-wura	55.34	75.15	26.57	49.53	57.50	18.37	56.46	49.92	50.50	10.30	24.40	57.72	29.20	49.69	52.43	49.17	63.98	18.30	42.96	4.81	26.80	43.78	42.69	46.33
AfriqueLlama-8B	60.71	76.90	55.50	59.31	68.20	18.32	63.77	62.01	68.71	12.97	39.16	72.52	61.77	68.59	63.38	69.35	72.42	50.42	65.00	6.41	27.71	65.17	63.15	65.86
Gemma3-4B	39.60	49.29	29.39	34.12	39.76	25.00	31.38	27.85	21.88	24.85	29.07	32.47	29.19	29.40	31.71	26.75	43.12	8.36	26.21	13.16	31.37	27.41	4.99	27.64
AfriqueGemma-4B	38.81	76.03	51.63	39.89	49.94	25.40	59.64	59.53	66.95	22.28	33.65	69.66	56.01	66.70	60.91	36.00	67.04	50.40	55.22	21.87	30.56	63.92	56.61	64.07
Gemma3-12B	49.49	59.16	42.99	45.83	47.95	20.24	40.88	47.56	40.96	21.72	28.31	43.40	24.70	41.76	41.42	43.34	49.16	19.31	29.81	17.57	30.22	35.28	22.04	38.04
AfriqueGemma-12B	61.99	77.91	62.35	62.67	69.15	21.43	64.07	63.01	70.26	23.04	50.22	73.67	65.17	70.09	64.72	70.19	73.87	56.58	66.56	19.46	30.48	66.55	64.83	67.18
Qwen3-8B	61.88	69.11	-0.70	55.73	64.22	22.04	5.17	4.89	1.87	9.26	14.01	14.37	13.78	11.94	5.10	10.65	15.57	5.28	10.65	5.11	25.11	6.03	0.02	4.10
AfriqueQwen-8B	61.14	76.23	57.18	58.60	67.38	22.18	62.20	60.60	66.49	11.44	28.40	71.34	60.75	67.09	62.01	67.92	70.87	45.58	64.05	7.16	27.55	63.29	60.88	63.69
Qwen3-14B	61.89	71.78	3.23	56.36	64.73	21.69	6.37	8.11	2.35	13.56	14.96	14.85	15.60	10.82	7.10	12.19	27.63	3.00	11.08	8.00	27.13	8.86	3.48	6.89
AfriqueQwen-14B	60.69	76.90	60.06	59.91	68.72	20.11	63.72	62.53	68.80	14.60	38.41	72.62	63.38	68.70	63.60	69.37	72.21	48.26	65.98	7.55	26.80	65.14	63.23	66.16

Table 18: Translation performance (SSA COMET score) from English to African languages on the FLORES-200 benchmark. (MT eng2xx)

model	aeb	afr	amh	ary	arz	ewe	hau	ibo	kin	lin	lug	nya	orm	sna	som	sot	swa	tir	tsn	twi	wol	xho	yor	zul
Llama3.1-8B	62.76	75.06	42.84	62.21	67.36	32.73	57.42	53.04	48.69	39.46	43.50	46.78	37.80	45.86	42.87	44.03	67.12	31.96	42.70	43.82	42.35	43.96	43.52	44.08
Lugha-Llama-8B-wura	62.15	75.77	60.02	60.97	66.84	32.26	64.34	57.67	63.15	38.46	49.35	64.50	51.48	63.27	61.42	64.33	70.47	52.85	60.66	38.46	41.03	62.90	55.23	64.27
AfriqueLlama-8B	65.30	75.77	64.50	64.15	68.58	31.78	65.35	59.52	65.63	38.87	51.96	66.32	59.33	65.55	63.74	67.96	71.18	59.34	64.76	39.67	39.54	65.80	58.89	66.48
Gemma3-4B	38.61	41.87	35.11	38.84	40.66	32.41	38.67	34.55	35.09	31.52	35.59	40.14	36.92	40.21	38.77	37.16	46.88	30.02	34.02	32.48	40.03	36.85	31.61	39.66
AfriqueGemma-4B	65.05	75.64	65.04	63.48	68.40	30.40	64.71	59.03	65.16	37.41	49.10	66.19	58.35	65.42	63.05	67.56	71.11	60.09	64.22	37.85	39.09	66.02	57.68	66.42
Gemma3-12B	48.19	46.88	39.95	42.59	53.33	32.77	35.77	38.61	36.55	33.89	40.52	39.91	43.38	36.90	46.12	38.32	42.41	36.41	38.00	33.15	39.24	39.46	39.85	42.21
AfriqueGemma-12B	66.29	76.31	68.07	65.47	69.51	32.69	66.63	61.09	66.88	43.16	55.44	67.37	61.90	66.60	65.04	68.94	72.15	63.04	65.57	44.61	41.67	67.44	60.26	67.18
Qwen3-8B	63.07	74.96	43.76	61.36	68.05	31.12	29.94	36.23	35.22	36.47	37.23	40.74	37.12	38.74	31.86	39.18	57.80	33.91	36.45	35.99	40.29	39.10	34.64	35.96
AfriqueQwen-8B	65.88	75.90	66.49	64.69	69.05	31.12	65.39	59.69	65.56	37.33	48.96	66.39	59.85	65.96	63.58	68.06	70.89	62.18	64.98	37.29	39.78	66.60	59.36	66.44
Qwen3-14B	64.59	75.62	46.49	62.91	69.07	32.34	34.07	41.13	39.03	38.49	39.54	44.05	41.30	41.51	34.64	44.72	63.36	36.76	40.26	38.40	41.57	45.28	39.85	42.87
AfriqueQwen-14B	66.65	76.24	67.82	65.59	69.70	32.24	66.39	60.75	66.33	40.28	52.82	67.18	61.48	66.70	64.47	68.90	71.76	63.01	65.81	39.17	40.84	67.36	60.91	67.48

Table 19: Translation performance (SSA COMET score) from African languages to English on the FLORES-200 benchmark. (xx2eng)

English → African Languages (eng2xx)																								
model	aeb	afr	amh	ary	arz	ewe	hau	ibo	kin	lin	lug	nya	orm	sna	som	sot	swa	tir	tsn	twi	wol	xho	yor	zul
Llama3.1-8B	33.87	62.46	7.43	28.37	34.66	6.64	28.59	19.42	12.22	13.12	10.98	12.51	10.79	11.40	17.38	13.25	37.69	4.12	14.83	15.92	9.14	12.36	12.72	12.41
Lugha-Llama-8B-wura	31.45	65.35	16.95	26.26	31.93	8.46	43.80	32.25	30.56	12.66	15.46	33.13	17.79	30.31	34.29	35.92	49.79	10.47	32.01	12.11	9.58	30.71	20.27	33.54
AfriqueLlama-8B	36.56	67.54	32.11	32.43	40.17	6.80	49.56	40.84	46.44	14.15	20.96	45.64	35.54	42.87	41.95	53.84	59.54	19.80	46.14	13.89	8.93	47.37	27.36	49.67
Gemma3-4B	17.89	39.24	12.07	12.10	16.98	7.73	26.01	20.91	12.34	12.00	7.32	14.98	7.94	14.51	15.91	14.90	31.09	3.11	11.63	11.25	7.67	14.01	7.62	18.15
AfriqueGemma-4B	5.49	66.53	19.88	6.10	12.10	8.74	44.07	39.29	44.72	11.91	18.34	42.05	27.29	40.88	37.94	16.82	51.14	14.25	34.62	10.28	8.11	46.21	23.04	47.49
Gemma3-12B	25.47	43.57	20.81	21.54	23.40	7.45	24.79	32.27	22.03	14.16	10.18	17.14	11.86	22.58	20.58	29.82	28.01	6.38	20.09	15.11	7.60	9.21	13.91	17.76
AfriqueGemma-12B	39.54	68.97	36.09	36.24	41.66	8.47	51.36	41.95	49.58	19.64	26.60	46.87	36.98	43.94	42.14	55.54	62.13	20.90	46.61	15.58	7.11	49.21	28.59	51.82
Qwen3-8B	35.19	59.93	6.46	29.63	36.23	5.67	8.72	9.08	7.34	10.53	6.90	8.41	8.80	6.87	11.19	10.08	20.56	2.49	9.44	9.00	7.74	10.37	7.65	9.73
AfriqueQwen-8B	35.86	66.94	31.84	31.29	39.02	6.71	48.40	39.82	43.70	12.04	16.75	44.42	34.11	41.43	40.79	51.64	57.72	17.88	44.93	9.58	8.71	44.66	27.21	46.92
Qwen3-14B	36.03	61.78	9.03	30.45	37.88	6.26	13.45	11.98	10.51	15.09	8.83	10.67	11.82	9.82	15.17	13.97	31.17	3.26	11.68	13.49	10.23	13.91	10.25	14.08
AfriqueQwen-14B	36.80	67.58	33.52	32.50	40.73	6.22	49.78	41.14	46.52	14.67	21.20	45.92	36.39	42.90	41.62	53.50	59.55	18.38	46.15	12.83	8.41	46.50	27.45	49.82
Gemma3-27B	34.85	59.41	27.52	29.65	34.04	4.71	21.76	35.10	17.54	19.80	15.35	6.81	18.16	10.04	19.42	27.30	23.05	10.42	15.68	17.15	6.98	25.95	18.31	21.10
African Languages → English (xx2eng)																								
model	aeb	afr	amh	ary	arz	ewe	hau	ibo	kin	lin	lug	nya	orm	sna	som	sot	swa	tir	tsn	twi	wol	xho	yor	zul
Llama3.1-8B	52.59	74.19	31.70	50.96	55.37	21.93	45.81	40.79	34.24	27.55	28.07	30.36	22.14	30.05	31.31	30.80	56.85	20.68	29.32	30.82	25.20	32.93	29.78	33.25
Lugha-Llama-8B-wura	51.29	75.17	46.43	49.15	54.18	20.63	52.42	46.43	48.15	26.64	31.79	45.71	32.89	44.73	47.66	50.29	62.01	35.20	42.72	26.17	24.09	50.46	38.05	53.54
AfriqueLlama-8B	54.93	75.07	52.68	52.99	56.31	20.78	54.25	49.67	53.11	26.94	34.86	48.95	43.42	48.09	50.83	56.86	63.08	42.92	48.56	27.20	22.37	55.89	42.43	57.62
Gemma3-4B	21.81	20.65	19.13	22.48	22.79	10.85	25.36	18.63	17.60	13.16	11.93	16.19	13.75	20.11	24.62	17.51	30.95	14.05	14.31	16.89	12.89	21.26	14.77	23.61
AfriqueGemma-4B	54.91	74.89	53.81	51.94	55.94	18.83	53.22	48.91	52.44	25.84	32.11	48.56	41.75	48.00	50.04	55.96	62.70	43.61	47.74	25.90	21.38	55.84	41.40	57.12
Gemma3-12B	32.42	11.50	21.56	21.58	35.81	13.92	15.27	15.07	11.51	10.96	15.10	9.08	18.78	7.31	28.42	11.70	15.80	18.23	11.60	12.17	12.38	15.20	17.14	18.70
AfriqueGemma-12B	56.92	76.62	59.33	55.70	58.55	19.71	56.79	52.73	55.71	29.77	38.41	51.07	47.54	50.43	53.50	59.43	66.01	48.34	50.27	30.86	22.53	59.21	44.72	59.87
Qwen3-8B	53.40	74.07	34.64	50.49	55.80	20.99	24.37	25.49	24.85	25.68	23.93	25.69	22.22	25.49	24.11	27.26	46.72	22.91	24.96	25.39	23.91	29.75	24.25	28.01
AfriqueQwen-8B	56.43	75.40	55.90	53.75	57.35	20.94	54.03	49.50	52.29	26.38	32.25	48.86	43.80	48.57	50.16	56.41	62.37	46.42	49.32	26.41	23.24	56.69	43.08	57.71
Qwen3-14B	55.39	75.15	36.91	52.67	57.75	22.35	28.23	29.80	28.04	27.63	25.79	28.43	26.00	27.76	26.33	32.74	53.21	24.76	28.22	27.33	24.99	34.89	27.94	33.13
AfriqueQwen-14B	57.60	75.98	58.03	55.60	58.66	21.97	56.27	52.01	54.58	28.44	35.96	50.44	46.87	50.00	52.21	59.06	64.77	48.20	50.44	27.83	24.15	58.34	45.35	59.65
Gemma3-27B	57.48	68.30	51.53	55.82	57.53	21.34	46.19	44.61	49.91	35.33	38.58	45.97	38.98	45.22	44.41	50.55	61.07	36.94	41.90	35.85	26.76	52.38	38.03	51.82

Table 20: Translation performance (chrF++ score) on the FLORES-200 benchmark. Upper section: English to African languages (eng2xx). Lower section: African languages to English (xx2eng). (MT chrF++)

model	amh	eng	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul
Llama3.1-8B	42.72	83.18	12.47	60.94	53.63	31.78	40.53	32.56	18.12	28.97	28.41	75.00	32.31	27.27	38.81	42.44	34.22
Lugha-Llama-8B-wura	66.84	82.15	11.91	78.78	64.94	55.28	36.25	38.44	37.84	62.09	52.87	80.19	21.94	22.38	64.53	63.66	58.31
AfriqueLlama-8B	78.78	79.68	9.66	80.19	70.78	62.19	38.78	50.16	59.72	72.47	63.38	81.19	22.22	19.81	75.22	73.38	65.59
Gemma3-4B	73.72	79.36	11.47	72.97	61.78	50.19	40.00	31.87	19.97	53.31	35.38	84.44	25.53	25.83	62.59	38.81	57.53
AfriqueGemma-4B	75.69	79.74	9.88	77.94	70.09	62.09	38.75	39.91	55.72	71.72	58.56	82.03	16.50	21.57	76.31	68.03	63.84
Gemma3-12B	83.03	85.76	15.97	84.97	72.59	68.28	53.41	64.53	48.91	80.03	59.72	90.47	47.81	36.24	80.09	67.12	73.56
AfriqueGemma-12B	82.81	82.64	13.84	84.22	77.94	68.44	44.47	62.19	67.28	79.69	66.97	86.34	34.69	29.31	81.53	76.09	73.41
Qwen3-8B	50.37	85.27	12.91	18.69	25.84	21.12	42.34	19.78	18.47	18.94	26.28	70.12	21.56	26.43	27.25	23.88	23.25
AfriqueQwen-8B	80.47	83.67	9.78	82.88	75.59	62.34	38.31	46.97	60.53	74.72	63.09	86.84	19.47	22.98	78.19	73.19	69.12
Qwen3-14B	59.34	88.52	15.38	29.41	39.66	27.59	46.44	31.53	30.28	31.84	33.38	81.66	30.84	32.82	47.06	38.28	41.28
AfriqueQwen-14B	83.22	86.30	14.88	85.88	78.94	67.81	41.38	58.38	69.62	81.25	64.75	86.72	25.69	31.10	82.53	81.88	73.25

Table 21: 5-shot performance on the Injongo Intent Classification benchmark. (Intent)

model	aeb	afr	amh	ary	arz	eng	ewe	hau	ibo	kin	lin	lug	nya	orm	plt	por	sna	som	sot	swa	tir	twi	wol	xho	yor	zul
Llama3.1-8B	76.19	81.24	49.91	80.06	80.90	80.78	44.72	67.67	67.39	54.60	47.94	45.20	62.61	42.00	60.97	82.43	47.57	58.33	50.27	66.50	41.63	62.87	46.26	52.36	47.44	52.28
Lugha-Llama-8B-wura	75.39	80.41	74.83	79.10	78.73	80.45	43.06	80.21	76.48	78.99	59.51	63.74	76.90	67.40	77.95	81.63	74.98	76.04	72.64	79.55	64.44	58.12	55.42	76.78	73.10	76.55
AfriqueLlama-8B	76.27	78.90	62.86	79.06	77.80	66.32	38.32	64.32	65.83	64.31	41.85	51.13	75.70	61.90	75.80	78.97	66.21	75.76	61.98	65.36	73.21	53.28	38.75	64.55	61.91	61.87
Gemma3-4B	76.81	79.84	63.38	78.55	81.36	69.21	44.00	59.88	56.37	60.65	46.20	38.38	74.48	38.98	71.44	80.36	56.59	74.16	51.54	66.70	63.75	57.82	45.28	55.46	38.68	59.53
AfriqueGemma-4B	77.69	79.44	61.15	79.10	79.84	70.76	42.81	65.28	63.27	62.74	40.04	42.59	78.33	64.57	77.17	79.15	63.88	77.15	63.00	65.80	77.26	55.27	41.41	62.31	55.22	61.84
Gemma3-12B	81.49	83.65	81.56	82.89	82.26	83.21	47.47	80.64	78.58	81.36	64.38	68.54	82.31	67.65	81.13	84.28	78.95	80.73	76.04	85.33	73.51	67.01	58.80	79.74	66.67	79.70
AfriqueGemma-12B	78.85	82.43	74.02	80.16	79.65	73.64	43.23	74.27	69.80	73.11	51.73	57.45	79.56	66.92	79.60	82.23	70.86	80.12	71.26	76.32	79.67	60.74	48.49	70.31	66.51	73.12
Qwen3-8B	83.11	83.89	49.98	83.15	86.49	72.29	43.43	37.49	36.48	37.86	43.81	35.60	60.53	39.48	57.25	83.32	37.99	50.54	41.42	55.91	49.97	55.08	47.76	42.00	36.67	37.92
AfriqueQwen-8B	80.79	83.84	72.40	81.04	83.39	77.58	42.41	74.45	70.43	78.06	53.30	58.53	81.38	69.51	81.54	82.92	74.41	81.65	71.87	78.45	81.42	51.36	52.12	71.78	72.42	74.23
Qwen3-14B	82.30	85.42	74.12	85.60	84.43	85.71	47.07	55.71	60.47	54.38	63.74	55.05	57.58	56.11	65.90	84.98	55.42	56.12	63.37	80.75	61.41	56.68	61.66	63.64	54.86	61.82
AfriqueQwen-14B	81.81	83.60	83.97	85.12	83.04	82.62	46.52	82.66	81.39	83.19	61.41	65.83	82.54	84.15	82.77	82.85	80.73	82.39	80.62	85.43	81.93	58.06	58.13	80.42	81.14	82.99

Table 22: 5-shot performance on the SIB-200 topic classification benchmark. (Topic)