

Lugha-Llama: Adapting Large Language Models for African Languages

Happy Buzaaba^{†*} Alexander Wettig^{†*} David Ifeoluwa Adelani[‡] Christiane Fellbaum[†]

[†]Princeton University [‡]Mila, McGill University & Canada CIFAR AI Chair

Abstract

Large language models (LLMs) have achieved impressive results in a wide range of natural language applications. However, they often struggle to recognize low-resource languages, in particular African languages, which are not well represented in large training corpora. In this paper, we consider how to adapt LLMs to low-resource African languages. We find that combining curated data from African languages with high-quality English educational texts results in a training mix that substantially improves the model’s performance on these languages. On the challenging IrokoBench dataset, our models consistently achieve the best performance amongst similarly sized baselines, particularly on knowledge-intensive multiple-choice questions (AfriMMLU). Additionally, on the cross-lingual question answering benchmark AfriQA, our models outperform the base model by over 10%. To better understand the role of English data during training, we translate a subset of 200M tokens into Swahili language and perform an analysis which reveals that the content of these data is primarily responsible for the strong performance. We release our models¹ and data² to encourage future research on African languages.

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress in Natural Language Processing and beyond (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023a; Chung et al., 2024; Zhang et al., 2023). Although LLMs perform well on a wide range of tasks in high-resource languages (Minaee et al., 2024; Huang et al., 2024), their performance in low-resource languages, especially African languages, continues to lag behind

(Ojo and Ogueji, 2023; Hendy et al., 2023). Africa is home to approximately one third of the world’s languages (Eberhard et al., 2024), but these languages are underrepresented in commonly used pre-training datasets (Conneau et al., 2020). As a result, LLMs are pre-trained on multilingual imbalanced datasets (Chung et al., 2023a; Touvron et al., 2023b), posing a significant challenge for multilingual models to recognize low-resource African languages.

In a recent line of research, LLMs that were trained primarily on English data are adapted to African languages in an additional training phase on African language corpora (Uemura et al., 2024). However, on challenging evaluations that require reasoning and expert knowledge, such as AfriMMLU (Adelani et al., 2024), these models lag far behind closed-source frontier models like GPT-4 (OpenAI, 2023).

In this paper, we produce the African-centric Lugha-Llama³ by continuing to pre-train Llama-3.1-8B (Dubey et al., 2024) on 10B multilingual tokens. Besides using a strong base model, we demonstrate the importance of curating the training data. Surprisingly, we find that adding high-quality English educational documents to the training data can further increase model performance in African language evaluations. Our model’s overall performance on IrokoBench (Adelani et al., 2024) is the best of any current open-weight model.

Why is English data beneficial? We analyze this by translating 200M tokens of educational English documents to Swahili using GPT-4o (OpenAI, 2023). The translated data perform substantially better than the English source data, suggesting that adding English is not necessary. However, it also outperforms existing high-quality Swahili corpora (Oladipo et al., 2023). This suggests that a consequence of data scarcity in low-resource languages

*The first two authors contributed equally. Correspondence to {happy.buzaaba@, awettig@cs.}princeton.edu.

¹<https://huggingface.co/Lugha-Llama>

²https://huggingface.co/datasets/princeton-nlp/fineweb_edu-swahili-translated

³Lugha is the Kiswahili word for “language”

is a gap in data quality compared to high-resource languages. It also raises the possibility of reducing this gap via large-scale machine translation.

We open-source the Lugha-Llama models and the 200M token corpus of educational documents translated to Swahili.

2 Background

Low-resource languages. According to (Joshi et al., 2020) many African languages are categorized as low-resource, and there is a growing interest in NLP research to address challenges faced by such languages. A number of studies have focused on creating task-specific benchmark datasets like Named Entity Recognition (Adelani et al., 2021), Part-of-Speech tagging (Dione et al., 2023) and Machine Translation (Adelani et al., 2022) to enable research on low-resource African languages. The most challenging of the African benchmarks is IrokoBench (Adelani et al., 2024)—that focuses on mathematics reasoning, knowledge QA and natural language inference. In our work, we evaluate on AfriQA and IrokoBench showing progress on challenging benchmarks.

Multi-lingual language models. Multilingual models are a promising approach to work with low-resource languages. Models such as mBERT (Devlin et al., 2019), XLM-R (Conneau, 2019) and mT5 (Xue, 2020) jointly pre-train large models on over 100 languages, but only a handful of African languages are included. Another approach is to use a small dataset to pre-train multilingual models from scratch as in (Ogueji et al., 2021; Adebara et al., 2022; Tonja et al., 2024), who release comparatively smaller models that in some cases match larger models pre-trained on much more data. Several other studies have adapted multi-lingual models through continued pre-training (Liu et al., 2021; Lu et al., 2024; Wu et al., 2024; Lin et al., 2024; Alabi et al., 2022; Adelani et al., 2022) and instruction tuning (Muennighoff et al., 2022; Huang et al., 2023; Singh et al., 2024b). The promise of continual pre-training is that some of the base model’s capabilities seen with a high-resource language will transfer to a low-resource setting with comparatively little training compared to training from scratch. Therefore, we build on the existing studies and focus on continual pre-training to develop an Africa-centric language model.

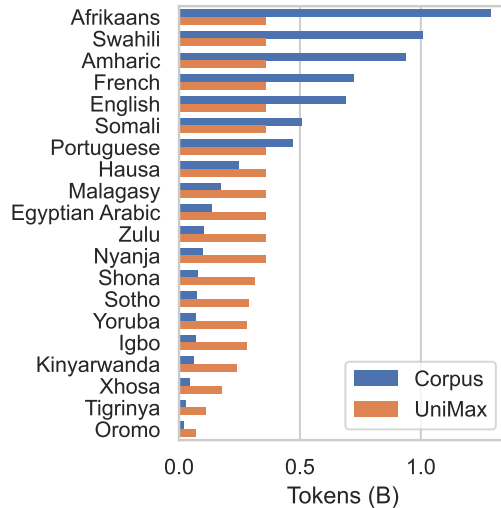


Figure 1: Tokens per language in the WURA corpus (Oladipo et al., 2023) and the 6B-token training data for training using UniMax sampling (Chung et al., 2023b).

Data Quality. Data curation has emerged as an important part of language model development, as it can have a substantial impact of downstream task performance (Li et al., 2024; Albalak et al., 2024; Oladipo et al., 2023). Techniques include heuristic filters to remove internet artifacts (Rafael et al., 2020; Rae et al., 2021; Penedo et al., 2023), identifying duplicated documents (Lee et al., 2022; Abbas et al., 2023), and training classifiers to identify documents based on some attribute, such as educational value (Wettig et al., 2024; Penedo et al., 2024). While this focus on data quality has achieved strong results on English-language corpora, little has been explored about the relationship of data quality and data availability across languages, as well as the implications of this on mixing multi-lingual data sources.

3 Experimental Setup

In our main experiments, we adapt a Llama-3.1-8B model to African Languages by training on 10B tokens with three different data mixtures. Note that the base model was pre-trained on 15T tokens, of which 8% are non-English (Dubey et al., 2024).

- **(Lugha-Llama)** We sample 10B tokens from the WURA corpus (Oladipo et al., 2023), which is comprised of sixteen African languages and four high-resource languages commonly spoken on the African continent, namely English, French, Arabic, and Portuguese. The corpus was collected by inspecting and cleaning mC4 (Xue, 2020) and crawling African websites. The cor-

Model	Size	eng	fra	amh	hau	ibo	kin	orm	sna	sot	swa	xho	yor	zul	ewe	lin	lug	twi	wol	Avg [†]
AfrimMLU																				
Llama-3.1	8B	61.4	47.8	30.6	31.4	30.4	31.0	29.2	30.2	31.0	34.6	27.0	30.8	30.4	26.0	33.0	29.6	26.8	28.4	30.0
Lugha-Llama	8B	65.4	51.2	37.2	35.4	35.6	35.2	33.6	36.4	37.8	38.4	33.4	32.2	34.6	25.8	33.8	29.0	26.2	26.0	33.2
-edu	8B	66.6	51.4	37.0	38.0	38.4	34.0	34.6	37.4	38.0	39.2	34.0	34.4	37.2	26.8	34.4	30.8	28.0	27.2	34.3
-math	8B	65.8	51.4	37.0	36.8	36.0	33.8	36.4	34.8	35.0	39.0	30.0	36.2	33.8	24.2	33.6	28.2	27.0	25.6	33.0
aya-101	13B	42.6	40.2	35.4	33.6	36.4	32.6	30.2	31.2	32.8	35.0	32.4	33.6	34.6	26.6	30.0	29.8	27.0	24.8	31.6
LLaMAX3	8B	55.4	44.0	30.0	33.2	32.8	28.8	30.0	32.4	28.8	35.2	29.4	29.8	31.0	22.4	32.8	27.0	25.4	25.6	29.7
AfroLlama-V1	8B	36.4	23.0	24.6	30.4	26.4	24.0	23.4	25.0	25.2	28.4	26.4	26.0	27.0	25.4	24.0	21.8	25.2	22.4	25.3
AfriInstruct	7B	43.2	33.6	21.4	26.0	24.0	26.8	27.2	28.2	27.8	30.2	30.8	26.2	28.2	26.0	27.0	24.0	27.8	25.6	26.7
InkubaLM	0.4B	24.0	23.2	25.4	25.8	25.2	23.6	27.0	21.8	26.0	25.4	25.0	23.6	24.2	24.2	23.6	23.8	25.6	26.8	24.8
AfrimGSM																				
Llama-3.1	8B	16.0	12.4	2.8	5.2	3.2	4.8	2.8	5.6	3.6	8.8	6.8	5.2	3.2	3.2	4.4	4.8	2.8	4.0	4.5
Lugha-Llama	8B	11.2	11.2	5.6	7.2	5.6	5.2	2.4	4.8	4.4	9.6	5.2	7.2	5.6	3.2	4.4	6.0	2.4	2.4	5.1
-edu	8B	13.2	9.2	4.4	7.6	3.6	4.8	4.8	4.4	5.2	9.2	4.4	6.0	7.2	2.4	4.8	4.4	2.8	3.2	5.0
-math	8B	15.2	11.6	5.6	10.8	5.2	6.0	4.0	5.6	5.2	11.2	6.4	8.0	6.0	2.8	4.8	6.4	3.6	3.2	5.9
aya-101	13B	7.6	6.8	5.2	6.0	2.4	2.8	2.0	3.2	2.8	4.4	2.8	3.6	3.2	1.6	1.6	2.4	2.4	1.2	3.0
LLaMAX3	8B	4.0	3.2	1.2	1.2	1.6	1.6	0.8	1.6	2.0	2.8	0.8	1.2	2.4	0.8	2.0	2.0	0.8	3.6	1.7
AfroLlama-V1	8B	3.2	4.0	0.8	3.6	0.8	1.6	0.0	2.4	0.8	4.0	2.4	3.2	2.8	0.8	2.4	2.4	0.4	0.8	1.8
AfriInstruct	7B	4.8	3.2	1.6	2.8	1.6	1.6	1.2	0.8	1.6	3.6	2.8	1.6	1.2	1.2	1.2	0.4	0.4	1.2	1.6
InkubaLM	0.4B	1.6	0.4	0.0	2.4	0.4	1.6	0.8	0.8	0.4	0.8	0.8	0.0	0.8	0.4	0.0	2.0	0.4	0.0	0.7
AfrimNLI																				
Llama-3.1	8B	50.8	51.7	35.0	35.8	33.7	35.8	37.2	35.7	33.5	38.2	32.0	35.3	33.3	34.5	32.0	33.2	35.3	33.8	34.6
Lugha-Llama	8B	50.5	50.8	40.3	41.2	38.0	39.0	38.7	39.8	41.3	43.5	41.5	39.7	38.5	33.2	33.7	33.3	33.0	32.8	38.0
-edu	8B	51.3	51.7	37.0	41.5	39.0	37.5	39.2	39.0	40.5	43.8	41.0	39.3	40.2	33.5	34.2	33.2	33.8	33.3	37.9
-math	8B	50.2	51.5	39.7	41.5	40.3	38.7	38.8	36.8	40.0	43.2	41.7	40.3	41.3	33.5	33.5	33.3	33.3	33.2	38.1
aya-101	13B	41.3	37.5	38.5	35.7	36.2	35.0	34.0	39.8	35.0	38.7	36.3	37.2	36.0	35.8	32.7	36.5	33.5	33.7	35.9
LLaMAX3	8B	47.0	48.7	34.8	41.7	33.5	35.0	37.3	40.3	34.7	43.8	37.8	36.0	35.5	33.5	33.0	34.3	33.0	34.5	36.2
AfroLlama-V1	8B	44.0	39.2	33.8	40.7	31.7	34.7	35.8	31.3	32.8	42.2	40.7	40.8	37.7	31.8	33.7	32.8	34.2	34.0	35.5
AfriInstruct	7B	51.0	49.7	34.2	40.7	37.2	35.5	35.8	38.8	36.2	37.3	39.7	40.5	38.2	36.0	32.7	31.8	34.3	33.0	36.4
InkubaLM	0.4B	31.2	32.8	33.2	35.0	35.2	33.2	34.2	33.2	32.7	32.3	34.5	34.0	33.8	33.0	33.7	32.8	34.5	33.7	33.7

Table 1: Results of Lugha-Llama models and baselines on IrokoBench (Adelani et al., 2024). Languages in *italic* are not present in the continual pre-training data. [†]: We exclude the high-resource languages English (eng) and French (fra) from the average, as they would otherwise skew the results due to strong English base models.

pus also includes 3 languages with non-latin scripts, namely Amharic, Arabic and Tigrinya.

- **(Lugha-Llama-edu)** We consider the effect of adding high-quality English educational documents to the WURA corpus. We combine 6B tokens from WURA with 4B tokens from FineWeb-Edu (Penedo et al., 2024)—a dataset obtained by prompting LLMs to score the educational content of web pages.
- **(Lugha-Llama-math)** To boost the mathematical reasoning abilities of African language models, we replace FineWeb-Edu with the OpenWebMath dataset (Paster et al., 2024), which contains documents with mathematical content identified via heuristic rules.

We sample from WURA using UniMax sampling (Chung et al., 2023b), which attempts to sample

as uniformly as possible across languages while limiting the number of times data is repeated. We upsample rare languages by at most four epochs, since this has been found to incur no discernible degradation during model training (Muennighoff et al., 2023). In Figure 1, we show the language proportions in the WURA corpus, as well as the training distribution with UniMax.

Training. We initialize the models with Llama-3.1-8B (Dubey et al., 2024) and train with a batch size of 512 sequences containing 8192 tokens each. We train for 2400 steps, totalling 10B tokens, with a learning rate of 10^{-5} with a cosine learning rate schedule with 240 steps linear warmup, and decaying to 10^{-6} . We disable attention across document boundaries within a sequence, following the strategy used by Llama-3.1-8B during pre-training.

Evaluation. We make use of the EleutherAI LM Evaluation Harness (Biderman et al., 2024) to evaluate on the IrokoBench (Adelani et al., 2024), a human translated benchmark for 16 typologically diverse low-resource African languages. This benchmark covers natural language inference (AfriXNLI), mathematical reasoning (AfriMGSM), and multi-choice knowledge-based question answering (AfriMMLU), which are derived from subsets of XNLI, MMLU, and MGSM respectively (Conneau et al., 2018; Hendrycks et al., 2021; Shi et al., 2022). We use few-shot prompting and follow Adelani et al. (2024) by setting the number of in-context demonstrations to eight for AfriMGSM and GSM8K, and 5 for all other tasks. We report the results with the language-specific prompt template for AfriXNLI, and the English templates for AfriMMLU and AfriMGSM, as these tasks do not provide a native language prompt template. Separately, we evaluate on AfriQA (Ogundepo et al., 2023), a cross-lingual open-retrieval question answering benchmark for African languages.

Baseline models. We compare Lugha-Llama models to six open-weight LLMs: Aya-101 (Üstün et al., 2024), InkubaLM-0.4B (Tonja et al., 2024), Llama-3.1-8B (Dubey et al., 2024), AfroLlama-v1 (8B)⁴, LLaMaX3-8B (Lu et al., 2024), and AfriInstruct (Uemura et al., 2024). We report more details about the provenance of these models in §A.1.

Model	Size	Average Score
Llama-3.1	8B	20.1
Lugha-Llama	8B	34.2
Lugha-Llama-Edu	8B	30.3
Lugha-Llama-Math	8B	37.7
AfroLlama-V1	8B	19.0
AfriInstruct	7B	21.9

Table 2: Results of Lugha-Llama models compared to base Llama and similarly sized Africa centric language models on AfriQA (Ogundepo et al., 2023).

4 Results

Our main results are shown in Table 1. Additional results and ablations can be found in §A.2.

⁴https://huggingface.co/Jacaranda/AfroLlama_V1

Training Data	AfriMMLU		
	eng	swa	Avg [†]
100% WURA _{swa}	64.4	41.0	30.6
60% WURA _{swa} + 40% FW-Edu	66.6	42.6	31.6
60% WURA _{swa} + 40% FW-Edu _{swa}	65.4	46.0	31.5
100% FW-Edu _{swa}	66.2	43.8	31.5

Table 3: We translate 200M-tokens from FineWeb-Edu to Swahili to disentangle the effects of semantic content from source language during continued pre-training.

Lugha-Llamas achieve strong results. In Table 1 we compare the three Lugha-llama models to the baselines on three tasks in IrokoBench. When considering the average scores over low-resource African languages, all three models perform better than all baselines across AfriMMLU, AfriMGSM, and AfriXNLI. The adaptation to African languages increases AfriMMLU scores by up to 8 percentage points compared to Llama-3.1-8B, with the largest improvements in Igbo (ibo). However, the performance is similar to the Llama-3.1-8B baseline on languages not present in WURA. The results in Table 2 show that our Lugha-Llama models significantly outperform the base model by more than 10% and consistently outperforms similarly sized Africa-centric language models on the cross-lingual question answering benchmark (AfriQA).

Addition of English data. We observe that including FineWeb-Edu data boosts the performance in AfriMMLU, and including data from OpenWeb-Math improves performance in AfriMGSM. This is surprising, since these models are trained on 40% less tokens from African languages. However, it means that performance may still be improved by training on the leftover data, which is especially appealing in a low-resource settings.

5 Case Study on Swahili Data Quality

We have adapted a language model that was primarily pre-trained on English data and saw improved performance on AfriMMLU when FineWeb-edu is part of the training data. Such cross-lingual generalization in language models has been observed and studied by many works (Artetxe et al., 2020; Deshpande et al., 2022; Ye et al., 2023, inter alia).

We study the question whether it is *necessary* for the FineWeb-Edu data to be in English to achieve these performance gains. For example, given the English nature of the base model, it may prevent catastrophic forgetting, a common problem in con-

tinual learning (Wang et al., 2023), or better help the model integrate African languages better with its pre-trained representations.

We take a random sample of FineWeb-Edu and translate 130M English tokens into 190M Swahili tokens by prompting GPT-4o with individual documents. We choose Swahili since we find that it is well supported by GPT-4o and is reported to have a better translation quality among many African languages (Costa-jussà et al., 2022; Robinson et al., 2023).

We repeat experiments similar to our main results but train only for 1B tokens on Swahili or English data. Table 3 shows that combining WURA with the translated educational data substantially outperforms the other data mixes. This suggests that the content of the FineWeb-Edu data is more critical to the performance than its English nature, and that a gap remains in terms of the pre-training data quality between English and low-resource languages such as Swahili.

6 Conclusion

We introduce three Lugha-Llama models based on continued pre-training, which achieve best-in-class performance on the challenging IrokoBench tasks. We demonstrate that combining African language pre-training data with educational English documents can improve downstream performance. Our case study suggests that there is still a gap in data quality between English and low-resource languages, but raises the possibility of closing this gap via large-scale machine translation.

Limitations

We discuss the limitations of our study of adapting language models to African languages.

Narrow evaluation. Our conclusions about the data mixture are sensitive to the choice of evaluations. We rely heavily on AfriMMLU, as MMLU is known to be an excellent signal for model quality. However, in a multilingual setting, a translated dataset like MMLU suffer from cultural bias which introduces evaluation challenges and limit its practical effectiveness as a global benchmark (Singh et al., 2024a). In addition, we notice that many questions look similar across languages as they use mathematical symbols and entity names that are largely unchanged in the languages. It is possible that the English FineWeb-edu data helps model make educated guesses based on surface cues in

these cases. Furthermore, some MMLU subjects are skewed towards Western and US-centric knowledge (Gema et al., 2024). African-centric LLMs should also be developed and evaluated based on African-centric knowledge benchmarks.

English prompts. IrokoBench only provides English task prompts for AfriMMLU and AfriMGSM, meaning that part of the evaluation input is always in English. While this might also contribute towards the usefulness of additional English data, our case study in Swahili suggests that this should not be a major factor, since continued pre-training with Swahili translations performs the best.

Limited coverage of African languages. We only train on the WURA corpus, which covers 16 low-resource African languages. This ignores the great linguistic diversity on the African continent, and future work is necessary to extend the coverage of language models.

Computational cost. Each model training run uses a considerable pre-training budget, requiring a total of 355 hours on Nvidia H100 GPUs. The resulting 8B-parameter model is also too expensive to run on most current consumer hardware.

Acknowledgment

We thank Tianyu Gao, Sewon Min, and Sanjeev Arora for useful discussions. The authors gratefully acknowledges financial support from Princeton Laboratory for Artificial Intelligence and Princeton Language and Intelligence for providing compute and azure credits. Happy Buzaaba is supported by the Center for Digital Humanities, the Africa Humanities Colloquium, and the Africa World Initiative at Princeton.

References

- Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. [arXiv preprint arXiv:2303.09540](#).
- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Serengeti: Massively multilingual language models for africa. [arXiv preprint arXiv:2212.10785](#).
- David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiters, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajudeen Gwadabe, Freshia Sackey, Bonaventure F. P.

- Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. [A few thousand translations go a long way! leveraging pre-trained models for African news translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. *Masakhaner: Named entity recognition for african languages*. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.
- Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *arXiv preprint arXiv:2204.06487*.
- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julien Etzaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Mimsa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. [Lessons from the trenches on reproducible evaluation of language models](#). *Preprint*, arXiv:2405.14782.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. *Palm: Scaling language modeling with pathways*. *Journal of Machine Learning Research*, 24(240):1–113.
- Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023a. [Unimax: Fairer and more effective language sampling for large-scale multilingual pre-training](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023b. [Unimax: Fairer and more effective language sampling for large-scale multilingual pre-training](#). In *The Eleventh International Conference on Learning Representations*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. *Scaling instruction-finetuned language models*. *Journal of Machine Learning Research*, 25(70):1–53.
- A Conneau. 2019. *Unsupervised cross-lingual representation learning at scale*. *arXiv preprint arXiv:1911.02116*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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, pages 2475–2485, Brussels, Belgium. 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](#).
- Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. [When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer](#). In [Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), pages 3610–3623, Seattle, United States. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In [North American Chapter of the Association for Computational Linguistics \(NAACL\)](#).
- Cheikh M Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, et al. 2023. Masakhapos: part-of-speech tagging for typologically diverse african languages. [arXiv preprint arXiv:2305.13989](#).
- 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](#).
- David M. Eberhard, F. Simons Gary, , and Charles D. Fennig. 2024. [\[link\]](#).
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. [Are we done with mmlu?](#) [Preprint, arXiv:2406.04127](#).
- 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](#).
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How good are GPT models at machine translation? A comprehensive evaluation](#). [CoRR](#), abs/2302.09210.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. [arXiv preprint arXiv:2305.07004](#).
- Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuan Zhang, Weijian Yi, Yulong Mao, Jincheng Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024. [A survey on large language models with multilingualism: Recent advances and new frontiers](#). [ArXiv](#), abs/2405.10936.
- 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](#). In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics](#), pages 6282–6293, Online. Association for Computational Linguistics.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In [Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruva Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishal Shankar. 2024. Datacomp-lm: In search of the next generation of training sets for language models. [arXiv preprint arXiv:2406.11794](#).
- Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, Andr’e F. T. Martins, and Hinrich Schütze. 2024. [Mala-500: Massive language adaptation of large language models](#). [ArXiv](#), abs/2401.13303.
- Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. [arXiv preprint arXiv:2105.03953](#).
- Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of

- llm by enhancing translation capabilities beyond 100 languages. [arXiv preprint arXiv:2407.05975](#).
- Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Asgari Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. [Large language models: A survey](#). [ArXiv](#), abs/2402.06196.
- Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. [Scaling data-constrained language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailley Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. [arXiv preprint arXiv:2211.01786](#).
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. [Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages](#). In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dos-sou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Roowether Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen Hassan Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Akari Asai, Tunde Oluwaseyi Ajayi, Clemencia Siro, Steven Arthur, Mofetoluwa Adeyemi, Orevaoghene Ahia, Anuoluwapo Aremu, Oyinkan-sola Awosan, Chiamaka Chukwuneke, Bernard Opoku, Awokoya Ayodele, Verrah Otiende, Christine Mwase, Boyd Sinkala, Andre Niyongabo Rubungo, Daniel A. Ajisafe, Emeka Felix Onwuegbuzia, Habib Mbow, Emile Niyomutabazi, Eunice Mukonde, Falalu Ibrahim Lawan, Ibrahim Said Ahmad, Jesujoba O. Alabi, Martin Namukombo, Mbonu Chinedu, Mofya Phiri, Neo Putini, Ndumiso Mgoma, Priscilla A. Amouk, Ruqayya Nasir Iro, and Sonia Adhiambo. 2023. [AfriQA: Cross-lingual open-retrieval question answering for African languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.
- Jessica Ojo and Kelechi Ogueji. 2023. [How good are commercial large language models on african languages?](#) In *Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda, May 1, 2023*.
- 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.
- OpenAI. 2023. [GPT-4 Technical Report](#). [Preprint](#), arXiv:2303.08774.
- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2024. Openwebmath: An open dataset of high-quality mathematical web text. In *The Twelfth International Conference on Learning Representations*.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, 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.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.
- 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](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.
- 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. 2024a. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. [arXiv preprint arXiv:2412.03304](#).

- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrman, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaei, and Sara Hooker. 2024b. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. 2024. Inkubalm: A small language model for low-resource african languages. [arXiv preprint arXiv:2408.17024](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. [LLaMA: Open and Efficient Foundation Language Models](#). [arXiv preprint arXiv:2302.13971](#).
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutika Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madsen Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rishi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). [ArXiv](#), abs/2307.09288.
- 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.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. [arXiv preprint arXiv:2402.07827](#).
- Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. 2023. A comprehensive survey of forgetting in deep learning beyond continual learning. [arXiv preprint arXiv:2307.09218](#).
- Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. QuRating: Selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52915–52971. PMLR.
- Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. [Adapting large language models for document-level machine translation](#). [ArXiv](#), abs/2401.06468.
- L. Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. [arXiv preprint arXiv:2010.11934](#).
- Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language generalists vs. specialists: An empirical revisiting on multilingual transfer ability. [arXiv preprint arXiv:2306.06688](#).
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-llama: An instruction-tuned audio-visual language model for video understanding](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations*, Singapore, December 6–10, 2023, pages 543–553. Association for Computational Linguistics.

A Appendix

A.1 Baselines

Aya-101 is based on mT5 (Xue, 2020) backbone and has been instruction tuned on diverse multilingual prompted task datasets covering 101 languages including ten African languages in IrokoBench. Llama-3-8B was pre-trained from scratch on 15 trillion tokens of which 8% are non-English, and was further adapted to **Llama-3.1-8B** (Dubey et al., 2024). **LLaMaX3-8B** (Lu et al., 2024) is based on continual pre-training of LLaMa-3-8B on massive multilingual data of 101 languages. **AfroLlama-v1** is based on LORA-based continual pre-training and instruction tuning of Llama 3 8B on six languages of Africa: Swahili, Xhosa, Zulu, Yoruba, Hausa and English.

Model	Size	eng	fra	amh	hau	ibo	kin	orm	sna	sot	swa	xho	yor	zul	ewe	lin	lug	twi	wol	Avg [†]
AfriXNLI (English Prompt)																				
Llama-3.1	8B	59.2	55.7	41.2	41.3	42.2	34.7	35.3	37.3	35.2	41.8	36.5	41.2	38.3	34.3	31.3	36.2	36.8	33.5	37.3
Lugha-Llama	8B	59.0	56.0	45.0	43.7	44.3	40.2	42.0	44.5	41.8	43.7	46.5	42.8	46.7	33.3	33.8	38.3	32.7	32.7	40.8
-edu	8B	60.0	56.7	45.7	44.0	44.3	38.7	42.3	44.8	43.7	42.3	45.2	43.0	44.3	34.0	32.5	39.0	33.5	32.5	40.6
-math	8B	60.2	58.3	44.8	46.5	47.2	41.8	44.8	47.2	45.0	47.8	47.8	45.2	44.3	35.7	33.5	39.2	33.5	32.0	42.3
aya-101	13B	63.8	61.2	56.2	55.5	53.3	56.2	52.3	57.0	54.5	54.2	57.2	50.5	53.8	45.2	33.0	53.3	49.7	36.0	51.1
LLaMAX3	8B	54.5	52.2	45.3	46.3	44.5	36.8	41.2	45.2	40.2	46.8	45.2	43.2	44.2	33.5	34.8	40.2	34.3	35.3	41.1
AfroLlama-V1	8B	48.5	44.0	33.3	44.8	34.8	35.3	34.2	34.3	34.7	45.8	44.5	42.3	44.7	33.0	33.0	35.2	36.8	34.5	37.6
AfriInstruct	7B	48.8	48.3	36.3	34.5	37.5	35.3	34.2	37.3	37.2	35.5	37.5	37.2	39.0	33.3	32.3	32.8	29.3	35.7	35.3
InkubaLM	0.4B	33.7	33.2	34.2	34.0	34.0	32.5	32.5	32.0	31.8	33.8	33.3	33.7	33.2	33.7	33.0	31.2	32.2	31.8	32.9

Table 4: Results of running AfriXNLI with an English prompt template, instead of a language-specific prompt template.

Model	Size	ibo	swa	hau	kin	zul	yor	bem	fon	twi	Avg [†]
AfriQA											
Llama-3.1	8B	44.1	42.3	19.5	15.5	12.5	13.0	11.4	12.6	9.9	20.1
Lugha-Llama	8B	57.0	47.5	49.2	44.9	37.7	37.3	16.8	10.3	6.9	34.2
-edu	8B	57.7	43.1	39.8	36.8	29.3	33.9	11.6	11.2	9.2	30.3
-math	8B	66.1	52.8	48.2	45.0	45.9	41.1	16.9	12.4	10.8	37.7
aya-101	13B	85.0	70.8	73.8	68.7	79.5	68.8	43.1	30.6	42.7	62.6
LLaMAX3	8B	2.27	1.83	1.67	2.56	1.55	1.57	1.23	3.64	1.26	1.97
AfroLlama-V1	8B	10.3	31.6	33.0	12.2	27.6	28.0	11.6	8.2	8.4	19.0
AfriInstruct	7B	35.3	32.0	29.6	25.6	28.7	30.5	5.7	5.8	3.8	21.9
InkubaLM	0.4B	1.1	1.3	1.2	2.7	2.2	1.3	1.9	2.0	1.3	1.7

Table 5: Results of Lugha-Llama models and baselines on AfriQA (Ogundepo et al., 2023).

InkubaLM-0.4B (Tonja et al., 2024) has been pre-trained from scratch on five African languages: isiXhosa, isiZulu, Swahili, Hausa and Yoruba. The closest model to our pre-training setup is **AfriInstruct** (Uemura et al., 2024)—based on continual pre-training of Llama 2 7B on WURA dataset followed by instruction-tuning on several African languages tasks.

A.2 Additional Results

Ablations. We explore some design decisions by running small-scale ablations by continuing to train Llama-3.2-3B for 1200 steps (5B tokens). Table 6 shows the results of comparing UniMax sampling vs. sampling languages according to their natural frequency in the WURA corpus. The trends of our main results still hold true at the 3B-parameter scale, e.g., adding FineWeb-edu to WURA helps on AfriMMLU. However, the magnitude of the gain is smaller compared to the 8B models. Table 6 also shows that adding 40% FineWeb-edu data performs marginally better than only adding 20%.

	AfriMMLU	AfriXNLI
Llama-3.2-3B	25.59	32.61
WURA only	26.05	33.88
w/o UniMax	26.00	33.11
80% WURA : 20% FineWeb-Edu	26.66	33.57
60% WURA : 40% FineWeb-Edu	26.70	33.55
w/o UniMax	26.44	32.82

Table 6: We ablate the importance of UniMax sampling (Chung et al., 2023b) by continuing to train Llama-3.2-3B by 5B tokens. We report the same average over non-high-resource languages as in Table 1.

Prompt formats. The AfriXNLI also has an English prompt template in the IrokoBench dataset (Adelani et al., 2024). We opt for choosing the native-language prompt template as our main evaluation, since we believe that it better reflects the actual use cases of these models. However, we feature the English prompt in Table 4. We find that

in this setting, the aya-101 model performs best. We note that it even outperforms Llama-3.1-8B on the English examples, despite Llama-3.1-8B’s strong overall benchmark performance. Similarly, in [Table 5](#) aya-101 model performs best on AfriQA (Ogundepo et al., 2023) compared to baselines.