AFRIINSTRUCT: Instruction Tuning of African Languages for Diverse Tasks

Kosei Uemura[1] Mahe Chen[1] Alex Pejovic[1] Chika Maduabuchi[2]
Yifei Sun[1] En-Shiun Annie Lee[1,3]
[1]University of Toronto [2]Massachusetts Institute of Technology [3]Ontario Tech University
{k.uemura, mahe.chen, alex.pejovic, idris.sun}@mail.utoronto.ca
chika691@mit.edu, annie.lee@ontariotechu.ca

Abstract

Large language models (LLMs) for African languages perform worse compared to their performance in high-resource languages. To address this issue, we introduce AFRIINSTRUCT, which specializes in instruction-tuning of multiple African languages covering various tasks. We trained the LLaMa-2-7B using continual pretraining and instruction fine-tuning, which demonstrates superior performance across multiple tasks. Our mixed task evaluation shows that our model outperforms GPT-3.5-Turbo and other baseline models of similar size. Our contributions fill a critical gap of LLM performance between high-resource and African languages. [1]

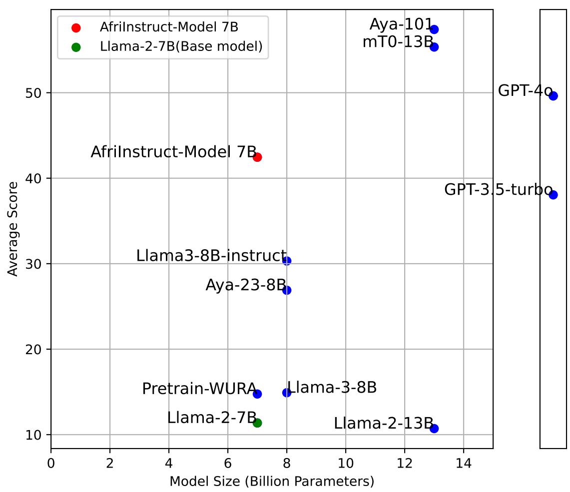

Figure 1: Average score of evaluating three tasks on by models and their sizes. Our model outperforms other baseline models of similar size.

1 Introduction

The development of large language models (LLMs) has revolutionized the field of natural language processing (NLP), enabling significant advancements in tasks such as machine translation (Arivazhagan et al., 2019; Pourkamali and Sharifi, 2024; Wang et al., 2023a; Zhu et al., 2023), sentiment analysis (Zhan et al., 2024; Zhang et al., 2023; Chandra et al., 2024; Zhan et al., 2024), and question answering (Kumar et al., 2024; Wang et al., 2024; Zhuang et al., 2023; Li et al., 2023). However, the vast majority of these breakthroughs have been concentrated on high-resource languages (HRLs), particularly English, due to the abundance of training data and resources available (Kargaran et al., 2023; Magueresse et al., 2020; Lai et al., 2024; Li et al., 2024b). In contrast, low-resource languages (LRLs), such as many African languages, have been largely left behind by the latest developments in NLP, despite their importance to millions of speakers worldwide (Nekoto et al., 2020; Adebara et al., 2024; Adebara and Abdul-Mageed, 2022; Tonja et al., 2024b; Adelani et al., 2023).

The challenges faced by LRLs in NLP are multifaceted and distinct from those of HRLs (Hedderich et al., 2021; Xu et al., 2024b; Tonja et al., 2024a; Khan et al., 2023; Krasadakis et al., 2024). These challenges include limited availability of annotated data (Khiu et al., 2024; Ding et al., 2024), complex morphology and syntax (Ghosh et al., 2024; Nzeyimana, 2024; Lopo and Tanone, 2024; Ghosh et al., 2023), and a lack of standardized orthography and terminology (Issaka et al., 2024; Lusito et al., 2023; Lin et al., 2024b; Downey et al., 2024). Existing approaches to address these challenges, such as adapting multilingual models (Ogueji et al., 2021; Wu et al., 2024; Csaki et al., 2023; Lin et al., 2024a) or creating targeted datasets (Muhammad et al., 2022; Lopo and Tanone, 2024; Yong et al., 2024; Bala et al., 2024), often face limitations in terms of scalability, generalizability, and performance (Urbizu et al., 2023; Cahyawijaya et al., 2024; Wang et al., 2023b; Ghosh et al., 2024). Multilingual models, while capable of handling a wide range of languages, of-

[1]Instructions to use models and datasets are available on https://github.com/AfricanLlama/AfriInstruct

Findings of the Association for Computational Linguistics: EACL 2024, pages 13571–13585
November 12-16, 2024 ©2024 Association for Computational Linguistics

Source Data	Task	of Tokens	of Prompts	of Languages
MasakhaNEWS	News Topic Classification	6,154,176	90,890	eng, fra, amh, hau, ibo, orm, sna, som, swa, tir, xho, yor
MasakhaPOS	Part-of-Speech Tagging	1,780,578	6,879	hau, ibo, kin, nya, sna, swa, xho, yor, zul
AfriSenti	Sentiment Analysis	19,201,035	235,225	amh, hau, ibo, yor, por, kin, swa
NollySenti	Sentiment Analysis	1,213,691	15,100	hau, ibo, eng, yor
xP3	xP3 - Multitask	640,745,532	7,773,312	eng, ara, ibo, hau, kin, nya, sna, sot, swa, xho, yor, zul
xP3	xP3 - Question Answering	146,758,736	541,630	eng, ara, ibo, hau, kin, nya, sna, sot, swa, xho, yor , zul
FLORES	Translation	5,692,402	72,324	eng, fra, afr, amh, ara, hau, ibo, kin, nya, por, som, sna, sot, swa, tir, xho, yor, zul
MAFAND	Translation	4,467,767	66,234	eng, amh, hau, ibo, kin, nya, sna, swa, xho, yor, zul
MasakhaNER2.0	Named Entity Recognition	12,935,191	58,667	hau, ibo, kin, nya, sna, swa, xho, yor, zul
MENYO	Translation	1,225,883	16,703	eng, yor
XL-Sum	Summarization	32,814,291	72,124	eng, amh, ara, hau, ibo, orm, por, swa, tir, yor

Table 1: Token and prompt counts by source and task in AFRIINSTRUCT-Data. A total of 19 languages are included in the data. All token counts have been computed with the Llama-2 tokenizer.

		eng	fra	afr	amh	ara	hau	igb	kin	mlg	nya	orm	por	som	sna	sot	saw	tir	xho	yor	zul
Train	Texts		2220759	2390884	291026	1116034	565471	121421	61485	355390	150016	37280	1548167	1235959	141559	124082	1801101	9807	69713	141321	166370
	Tokens	797885070	759908071	1413686044	1022590005	84210435	273874667	79459903	30727804	191150585	109037755	18421151	512594713	578216725	87883070	79440367	1131951011	29516647	48519904	81704390	112352151
Eval	Texts	260463	246611	265117	32307	124808	63067	13899	6902	39314	16880	4005	173578	137938	16126	13954	200345	1084	7846	15612	18289
	Tokens	89210685	83868231	156702768	113706484	9593837	30453342	9201202	3445257	21124508	12192695	1938844	57512714	64553254	10077670	8903982	125345094	3304872	5544281	8953997	12277472

Table 2: Number of Prompts per Task and Language in WURA text corpus. A total of 20 languages are included in the data. All token counts have been computed with the Llama-2 tokenizer.

Task	Prompt
Machine Translation	Translate the following text from {source language} to {target language}. {source language}:{source texts}. {target language}:
Named Entity Recognition	Study this taxonomy for classifying named entities:- LOC (Location or physical facilities)- ORG (Organizations, corporations or other entities)- PER (Names of people)- DATE (Date or time)Identify all named entities in the following tokens:{split tokens} Additionally, you should add B- to the first token of a given entity and I- to subsequent ones if they exist. For tokens that are not named entities, mark them as O.Answer:
News Topic Classification	Which of these labels best describes this news article:{topic candidates}{target sentence} Label:
Part-of-Speech Tagging	Study this taxonomy for classifying parts of speech:- X: Other- ADJ: Adjective- ADP: Adposition- ADV: Adverb- AUX: Auxiliary verb- CCONJ: Coordinating conjunction- DET: Determiner- INTJ: Interjection- NOUN: Noun- NUM: Numeral- PART: Particle- PRON: Pronoun- PROPN: Proper noun- PUNCT: Punctuation- SCONJ: Subordinating conjunction- SYM: Symbol- VERB: VerbPerform Part-of-Speech (POS) tagging on the following tokens: {split tokens} Answer:
Sentiment Analysis	Analyze the sentiment expressed in the following tweet'{ text }'Options: positive, negative, neutral
Summarization	{ passage } Write a summary of the text above in { target language}:

Table 3: Prompt templates used for different tasks and datasets. We referred to (Sanh et al., 2022) for prompt templates.

ten underperform compared to monolingual models and struggle with the unique characteristics of LRLs (Yoon et al., 2024; Huang et al., 2024; Xu et al., 2024b; Blevins et al., 2024). Targeted datasets, while valuable for specific tasks and languages, may lack the comprehensiveness and diversity needed to train robust and versatile NLP models (Li et al., 2024a; Du et al., 2024; Kesgin et al., 2024).

To address the critical shortage of resources for African languages, we propose AFRIINSTRUCT, which contains the following contributions: The main contributions of this work are as follows:

1. AFRIINSTRUCT-Data: a comprehensive Africentric instruction tuning dataset covering diverse tasks; and

2. AFRIINSTRUCT-Model: a high-performing language model for multiple African languages that demonstrates the effectiveness of targeted pretraining and fine-tuning strategies in low-resource settings.

2 Materials and Methods

Dataset Corpora AFRIINSTRUCT-Data is compiled from ten publicly available multilingual datasets, including FLORES (Goyal et al., 2021), MAFAND-MT (Adelani et al., 2022a), MENYO (Adelani et al., 2021a), MasakhaNER2.0 (Adelani et al., 2022b), MasakhaNEWS (Adelani et al., 2023), MasakhaPOS (Dione et al., 2023b), xP3 (Muennighoff et al., 2023a), AfriSenti (Muhammad et al., 2023b), NollySenti (Shode et al., 2023), and XL-Sum (Hasan et al., 2021). In total, AFRIINSTRUCT-Data comprises approximately 17 million prompts and 870 million tokens counted by Llama-2 tokenizer (Touvron et al., 2023), covering a wide range of African languages, with FLORES and MAFAND offering the broadest language coverage (Table 1) [2]. The dataset is preprocessed for instruction tuning by creating prompts in a zero-shot cross-lingual manner, where the context and

[2]We acknowledge that FLORES was originally designed as an evaluation dataset. However, due to its high quality and coverage of African languages, we opted to use FLORES for training, while utilizing other datasets, such as NTREX, for evaluation purposes.

LoRA Rank	CPT	Hau			Ibo			Kin			Swa			Yor			Zul			General			Avg		
		QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC
0	F	1.13	12.54	17.10	2.11	11.98	15.11	3.14	15.99	17.55	0.49	21.35	18.16	0.23	14.05	19.35	2.07	13.89	17.77	1.51	14.62	18.56	1.53	14.92	17.66
0	T	7.37	14.97	26.12	10.13	15.14	35.06	11.80	16.15	5.49	4.50	15.95	46.00	3.44	11.96	9.96	8.02	16.15	8.84	6.67	14.18	21.91	7.42	14.93	21.91
32	F	3.10	12.79	8.10	3.93	12.63	8.31	5.15	12.78	4.01	1.60	14.32	7.14	0.70	12.60	10.74	3.50	12.93	6.83	2.94	12.69	7.12	2.99	12.96	7.46
32	T	25.33	32.51	18.38	34.76	28.98	16.66	28.60	31.78	21.73	8.13	37.03	21.96	10.07	20.74	13.62	30.33	33.56	18.39	14.95	28.79	20.67	21.74	30.48	18.77
64	T	24.05	31.38	16.25	32.24	28.85	14.59	27.28	30.44	19.28	14.83	35.16	21.64	9.23	20.31	13.51	29.24	32.41	20.70	15.21	27.63	21.74	21.73	29.45	18.24
128	T	24.02	32.41	18.16	32.08	28.44	20.52	29.37	29.92	19.08	11.72	37.35	20.48	12.14	20.76	16.09	27.14	33.12	13.82	19.17	29.04	20.88	22.23	30.15	18.43
256	T	26.39	32.60	18.36	39.60	28.35	18.96	33.91	31.50	27.41	9.68	37.99	23.90	8.84	21.26	17.33	34.95	33.17	18.19	19.80	29.51	24.14	24.74	30.63	21.18
512	T	31.32	32.88	22.86	42.45	29.81	28.77	32.26	31.62	26.47	12.39	38.50	30.98	8.82	21.37	18.12	37.24	33.64	25.27	17.21	29.16	28.67	25.96	31.00	25.46

Table 4: Comparison of question answering(QA), machine translation(MT), and topic classification scores(TC) across different models, rank and whether we conduct continual pre-training. We used ChrF for machine translation, and F1 score for question answering and topic classification. For coloring, 0-10: low, 11-20: medium-low, 21-30: medium, 31-40: medium-high, 40-: high

query are provided in English, while the text to be analyzed is in the target African language (Table 3). This approach leverages English prompts to facilitate cross-lingual transfer and improve performance on African languages (Philippy et al., 2024; Ogundepo et al., 2023; Qiu et al., 2024; Lin et al., 2019; Chai et al., 2024; Adewumi et al., 2022).

Language Model Given the created AFRIINSTRUCT-Data for African languages, we developed the AFRIINSTRUCT-Model as an instruction-tuned LLM. The base model we used is LLaMa-2 (Touvron et al., 2023), one of the leading LLMs on many benchmarks. The training of AFRIINSTRUCT-Model involves two stages. First, we performed language adaptation using continued pretraining (Gururangan et al., 2020) on African language corpora. This extends the capabilities of LLaMa-2, which is originally English-centric, to African languages. Second, we further conducted instruction tuning on the model to improve the model's instruction following ability on diverse African tasks.

To adapt existing English-centric LLMs to other languages, continual pretraining is a popular intermediate training strategy that has been applied in previous studies (Cui et al., 2023; Xu et al., 2024a; Zhao et al., 2024). In this work, we use the African corpus WURA (Oladipo et al., 2023) for pretraining. It covers 16 African languages in total, and detailed information is provided in Table 2. After continual pretraining, we name the resulting model "Pretrain-WURA".

We fine-tune the continual pretrained model on our AFRIINSTRUCT-Data to enhance the model's general capabilities. At this stage, we use Low-Rank Adaptation (LoRA) (Hu et al., 2022), which is an effective yet lightweight fine-tuning strategy.

3 Experimental Settings

Abalation Study To determine the importance of hyperparameters, we evaluated the effectiveness of continual pretraining and LoRA (Hu et al., 2021) fine-tuning with different ranks using LLaMa2-7B[3]. The continued pretraining on LLaMa-2-7B using the WURA dataset was done using the run_llmmt.py script provided in the ALMA codebase (Xu et al., 2024a). We used eight NVIDIA-A100 40GB GPUs, and we ensured that we trained 1B tokens by training 8000 steps, with a per-device batch size of 2 and 16 gradient accumulation steps. Since the sequence length is 512, we get $8*8000*2*16*512 \approx 1B$ tokens. The LoRA instruct-tuning process was limited to 500 steps, which is significantly less than 1 epoch, but sufficient to observe the convergence of model training. We used the Unsloth repository to fine-tune the model during this experiment.

Comparative Study In comparison with other base models, we conducted one epoch fine-tuning with LoRA using Axolotl. We used 2 x A10 24 GPUs to fine-tune the continual pre-trained model based on LLaMa2-7B. For important parameters, we employed LoRA rank: 32, LoRA alpha: 16, and LoRA dropout: 0.05, and learning rate: 0.00002. Note that based on the result of experiment one,

[3]We chose LoRA rank, which the most effective hyperparameter in LoRA fine-tuning. Also, LoRA is applied to key, query, value, output, gate, up, and down projections.

Models	Rank	CPT	Hau			Ibo			Kin			Swa			Yor			Zul			General		
			QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC
Llama3 8b	32	F	6.94	17.41	38.07	11.19	14.82	24.45	7.44	13.65	27.75	4.70	26.97	32.85	1.19	12.42	15.79	7.53	14.88	25.16	4.17	15.62	39.16
Llama3 8b	64	F	8.16	19.59	28.90	13.74	16.10	27.32	8.23	16.03	33.45	4.22	30.02	39.20	1.24	13.98	26.76	6.20	14.83	35.13	3.50	16.51	33.98
Llama3 8b	128	F	6.95	22.50	33.11	12.36	18.58	30.41	7.80	17.32	38.57	4.41	32.30	33.59	1.06	15.62	24.39	8.95	18.14	28.51	4.80	19.55	34.07
Llama3 8b	256	F	7.91	25.79	23.39	13.12	20.18	26.29	7.93	20.12	36.41	5.98	35.98	32.11	1.32	17.50	20.25	8.59	19.77	20.44	5.53	20.10	42.11
Llama3 8b	512	F	27.71	30.07	40.46	45.81	24.82	43.44	21.56	23.57	42.15	20.42	41.55	47.73	6.57	19.08	28.55	29.16	24.67	23.93	17.97	24.99	41.27

Table 5: Comparison of question answering(QA), machine translation(MT), and topic classification scores(TC) across different models, rank and whether we conduct continual pre-training. We used ChrF for machine translation, and F1 score for question answering and topic classification.

we decided to use LoRA rank as 32. See further discussion in Result and Analysis. We established baseline results by conducting inference across a diverse range of language models (mT0-xxl (Muennighoff et al., 2023b), Aya23-8B (Aryabumi et al., 2024), LLaMa2-7B, LLaMa2-13B (Touvron et al., 2023), LLaMa3-8B (AI@Meta), GPT-3.5-Turbo and GPT-4o (OpenAI et al., 2024)).

We assessed the model on three evaluation tasks: Translation, Topic Classification, and Question Answering (Table 6). For Translation using NTrex (Federmann et al., 2022), we calculate chr-f scores between inferred response and target response directly. For Topic Classification using SIB-200 (Adelani et al., 2024a) we calculate F1 scores between ground truth and predicted labels. In detail, we prompt the model to choose from science/technology, travel, politics, sports, health, entertainment, and geography. The extracted output is then matched to the closest topic using fuzzy logic, ensuring a label is assigned only if the similarity ratio exceeds 80%. For Question Answering using AfriQA (Ogundepo et al., 2023), we adopt a token-based F1 score to evaluate the precision and recall of the predicted answer. We calculate the number of tokens that accurately appear in both the predicted response and the true answer. Precision is computed as the proportion of correct tokens within the prediction, while recall measures the proportion of correct tokens relative to the total in the true answer.

Task	hau	ibo	kin	swa	yor	zul	general
Question-answering	226	295	273	184	166	194	1338
Topic classification	161	154	168	118	132	125	858
Machine Translation	481	440	438	384	374	360	2477

Table 6: Number of Prompts per Task and Language

4 Result and Analysis

Our ablation study demonstrates the effectiveness of LoRA rank and continual pretraining (Table 4). Continual pretraining with a corpus in African languages contributed significantly to improved accuracy, suggesting effective knowledge injection. In the comparative verification by rank, no significant improvement in accuracy was observed from Rank 32 to 256, indicating that knowledge transfer does not vary significantly within this rank range. However, an improvement of about 5 points was seen at Rank 512 compared to Rank 32, suggesting that higher-rank LoRA training can be expected to facilitate certain levels of knowledge injection. Due to the dataset and computer resources available, training at Rank 512 for one epoch was not feasible, so we conducted LoRA fine-tune with one epoch of training at Rank 32.

Additionally, we conduct this comparative evaluation on Llama3-8b (Table 5). Llama3 has demonstrated gradual improvement in performance across all languages and tasks as the rank increases. The performance improvements indicate that the model benefits from higher ranks, which allow for more effective handling of the linguistic diversity present in these tasks and languages. [4]

Next, when compared with baseline models, our AfriInstruct-Model-7B outperforms language models of similar size, such as Aya23-8B (Aryabumi et al., 2024), LLaMa-3-8B (AI@Meta), and GPT-3.5-Turbo (Brown et al., 2020) (Table 7, described in Appendix B). However, overall, Aya101 (Üstün et al., 2024) and mT0-xxl (Muennighoff et al., 2023b) achieve the best performance

[4] Llama2 was chosen as the base model for the later experiment because, at the time of conducting continual pretraining for our experiments, Llama3 had not yet been released.

Models	Hau			Ibo			Kin			Swa			Yor			Zul			General			Avg		
	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC	QA	MT	TC
AFRIINSTRUCT-Model-7B	58.63	25.82	53.12	71.77	23.89	60.13	54.79	26.36	54.84	24.01	30.78	54.53	14.17	17.47	60.01	60.22	26.84	57.76	33.83	24.04	58.54	45.35	25.03	57.00
LLaMa-2-7B	1.13	12.54	17.10	2.11	11.98	15.11	3.14	15.99	17.55	0.49	21.35	18.16	0.23	14.05	19.35	2.07	13.89	17.77	1.51	14.62	18.56	1.53	14.92	17.66
Pretrain-WURA	7.37	14.97	26.12	10.13	15.14	35.06	11.80	16.15	5.49	4.50	15.95	46.00	3.44	11.96	9.96	8.02	16.15	8.84	6.67	14.18	21.91	7.42	14.93	21.91
LLaMa-3-8B	3.14	13.49	32.86	3.09	12.11	34.66	4.38	12.18	22.99	0.78	13.83	34.71	0.22	10.19	17.89	3.07	13.60	30.89	2.63	11.50	34.75	2.47	12.41	29.82
LLaMa-3-8B-Instruct	27.36	35.60	37.44	48.79	27.71	42.38	28.61	26.29	34.20	10.63	46.45	42.39	1.15	20.54	29.35	31.09	25.92	28.00	19.58	27.48	45.64	23.88	30.00	37.05
Aya23-8B	27.25	18.72	33.38	34.62	14.91	40.85	19.83	17.51	40.32	12.44	21.39	54.14	0.31	14.19	40.59	27.13	18.24	42.27	16.36	18.05	52.20	19.70	17.57	43.39
LLaMa2-13B	1.63	11.04	15.35	2.10	9.47	13.49	2.73	12.23	15.04	0.23	6.10	3.70	0.09	5.06	6.35	27.13	18.24	42.27	1.65	10.97	20.02	7.18	11.46	21.20
Aya101	68.85	46.80	79.37	83.10	40.14	78.27	64.82	39.10	76.79	22.07	52.23	82.61	8.81	25.45	71.53	79.36	43.46	80.49	45.14	39.00	77.92	53.16	40.88	78.14
mT0-xxl	62.95	38.94	73.99	80.46	40.16	71.71	64.17	41.51	71.38	21.18	53.16	81.66	6.23	25.08	73.04	72.89	45.24	79.60	46.18	38.03	74.59	50.58	40.30	75.14
GPT-3.5-Turbo	10.77	35.62	61.79	18.46	25.88	63.42	19.74	31.85	67.39	10.56	58.54	81.39	3.06	22.57	49.12	19.99	35.27	68.32	10.69	32.66	71.92	13.32	34.62	66.19
GPT-4o	18.18	52.96	83.08	26.10	45.49	86.09	29.93	48.53	81.98	6.61	60.08	84.69	1.54	27.73	82.53	25.89	49.51	84.72	16.11	45.11	85.30	17.77	47.06	84.06

Table 7: Comparison of question answering (QA), machine translation (MT), and topic classification (TC) scores across different models. We used ChrF for machine translation and F1 score for question answering and topic classification. For coloring, 0-10: low, 11-20: medium-low, 21-30: medium, 31-50: medium-high, 50-: high. The rows are divided according to the model size.

on all tasks, as they have been trained on a massive amount of multilingual and multitask instruction tuning datasets. This can be primarily attributed to the smaller size of our AFRIINSTRUCT-Model-7B model. Particularly noteworthy is the model's performance on QA tasks, which require an understanding of both low-resource and high-resource languages as they involve answering questions in low-resource languages based on English references. AFRIINSTRUCT-Model-7B surpasses GPT-4o in this area. It is also important to note the similar distribution of scores among Aya101, AFRIINSTRUCT-Model-7B, and mT0-xxl, especially in QA and TC tasks, with a slightly lower tendency in MT tasks. This similarity can be attributed to the significant proportion of the xP3 dataset they all share. The comparison between AFRIINSTRUCT-Model-7B and the pretrained model shows the substantial utility of the instruct dataset. Achieving results close to Aya101 and mT0-xxl within about 10 points, using only a 7B model with LoRA Rank 32, indicates that our training strategy is effective.

5 Conclusion

This paper showcases our advancement for African languages through the development of AFRIINSTRUCT via AFRIINSTRUCT-Model-7B and AFRIINSTRUCT-Data. The AFRIINSTRUCT-Data dataset supports instruction tuning of diverse tasks such as machine translation, topic classification, and more. AFRIINSTRUCT-Model-7B, enhanced by continual pretraining with the WURA dataset and fine-tuning with the LoRA technique, excels particularly in question-answering, outperforming prominent models like LLaMa2-7B, LLaMa3-8B, and GPT-3.5-Turbo. This implies the effectiveness of targeted instruction tuning datasets for pretraining and fine-tuning of African languages, addressing the critical need for comprehensive datasets and models for low-resource languages.

6 Limitation

Despite the progress made in this study, several limitations should be acknowledged:

AFRIINSTRUCT-Data offers coverage across multiple African languages and NLP tasks but is not exhaustive. Many African languages remain underrepresented, and several NLP tasks are not included.

AFRIINSTRUCT-Data may be culturally biased. The public datasets we compile often favor accessible, well-documented cultures, leading to biases in expressions and idioms towards dominant cultural narratives. Expanding and diversifying data sources would help better represent all low-resource languages.

AFRIINSTRUCT-Model is based on LLaMa2, which primarily benefits high-resource languages. This limits its effectiveness for low-resource languages, sometimes resulting in meaning distortions or incoherent text. Further research into better adaptation strategies and broader linguistic inputs is recommended to enhance the model's capability across all languages.

While our benchmark demonstrated the potential of AFRIINSTRUCT-Model-7B, its generalizability to other NLP tasks or domains in African languages remains uncertain. Testing in more varied contexts is needed.

Finally, our evaluation used metrics like F1 scores and ChrF, which may not fully capture culturally specific nuances in low-resource languages. Developing more culturally sensitive evaluation methods could provide a more accurate assessment of model performance.

Addressing these issues in future work will help improve the inclusivity and robustness of NLP models for African languages, fostering greater equity in technology.

Ethics Statement

This paper presents the development of AFRIINSTRUCT-Model which is built upon the AFRIINSTRUCT-Data. In conducting this research, we adhered to the following ethical guidelines and considerations:

1. Dataset Usage and Permissions: AFRIINSTRUCT-Data is compiled from publicly-available datasets. It contains no personally identifiable information or sensitive data, ensuring compliance with privacy standards.

2. Model Development and Integrity: We have ensured that the development of AFRIINSTRUCT-Model does not amplify biases inherent from the source datasets. Our approach to building and testing the model was transparent and can be independently verified through the benchmarks we introduced.

3. Adherence to Ethical Guidelines: Our research complies with international guidelines for ethical research in computational linguistics and artificial intelligence.

Acknowledgements

We would like to express our sincere gratitude to Miaoran Zhang and Jesujoba Oluwadara Alabi for their invaluable advice on writing and structuring this paper. Their guidance greatly improved the clarity and quality of our work. We also extend our sincere thanks to David Ifeoluwa Adelani for his insightful theoretical and experimental advice, which was instrumental in shaping the direction of our research.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *Preprint*, arXiv:2203.08351.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *Preprint*, arXiv:2401.01053.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen 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. 2022a. 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 Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David I. Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, and et al. Adesina Ayeni. 2021b. The effect of domain and diacritics in yorùbá-english neural machine translation. *Preprint*, arXiv:2103.08647.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, and et al. Xiaoyu Shen. 2022c. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *Preprint*, arXiv:2205.02022.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, and et al. Jesujoba O. Alabi. 2024a. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, and et al. Atnafu Lambebo Tonja. 2023. Masakhanews: News topic classification for african languages. *Preprint*, arXiv:2304.09972.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, and et al. Michael Beukman. 2022d. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *Preprint*, arXiv:2210.12391.

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, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024b. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Tosin Adewumi, Mofetoluwa Adeyemi, Aremu Anuoluwapo, Bukola Peters, Happy Buzaaba, Oyerinde Samuel, Amina Mardiyyah Rufai, Benjamin Ajibade, Tajudeen Gwadabe, Mory Moussou Koulibaly Traore, Tunde Ajayi, Shamsuddeen Muhammad, Ahmed Baruwa, Paul Owoicho, Tolulope Ogunremi, Phylis Ngigi, Orevaoghene Ahia, Ruqayya Nasir, Foteini Liwicki, and Marcus Liwicki. 2022. Afriwoz: Corpus for exploiting cross-lingual transferability for generation of dialogues in low-resource, african languages. *Preprint*, arXiv:2204.08083.

AI@Meta. Llama 3 model card.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, and et al. David Cairuz. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Abhinaba Bala, Ashok Urlana, Rahul Mishra, and Parameswari Krishnamurthy. 2024. Exploring news summarization and enrichment in a highly resource-scarce indian language: A case study of mizo. *Preprint*, arXiv:2405.00717.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *Preprint*, arXiv:2401.10440.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *Preprint*, arXiv:2403.16512.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Preprint*, arXiv:2401.07037.

Rohitash Chandra, Baicheng Zhu, Qingying Fang, and Eka Shinjikashvili. 2024. Large language models for sentiment analysis of newspaper articles during covid-19: The guardian. *Preprint*, arXiv:2405.13056.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *Preprint*, arXiv:2311.05741.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *Preprint*, arXiv:2403.02990.

Cheikh M. Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, and et al. Thapelo Sindane. 2023a. Masakhapos: Part-of-speech tagging for typologically diverse african languages. *Preprint*, arXiv:2305.13989.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023b. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted multilingual adaptation for low-resource language families. *Preprint*, arXiv:2405.12413.

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Binhang Yuan, Wenhu Chen, Jie Fu, and Ge Zhang. 2024. Chinese tiny llm: Pretraining a chinese-centric large language model. *Preprint*, arXiv:2404.04167.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, and et al. Ahmed El-Kishky. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Poulami Ghosh, Shikhar Vashishth, Raj Dabre, and Pushpak Bhattacharyya. 2024. A morphology-based investigation of positional encodings. *Preprint*, arXiv:2404.04530.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. Dale: Generative data augmentation for low-resource legal nlp. *Preprint*, arXiv:2310.15799.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, and et al. Guillaume Wenzek. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. 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.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, and et al. Yuan-Fang Li. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *Preprint*, arXiv:2106.13822.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Preprint*, arXiv:2010.12309.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen 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. *Preprint*, arXiv:2405.10936.

Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, and Xuefeng Du. 2024. The ghanaian nlp landscape: A first look. *Preprint*, arXiv:2405.06818.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosgpt: Monolingual training for turkish language models. *Preprint*, arXiv:2404.17336.

Muzammil Khan, Kifayat Ullah, Yasser Alharbi, Ali Alferaidi, Talal Saad Alharbi, Kusum Yadav, Naif Alsharabi, and Aakash Ahmad. 2023. Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Applied Sciences*, 13(15).

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Armando Parra Flores, Leandro Acros Roman, A. Seza Doğruöz, and En-Shiun Annie Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. *Preprint*, arXiv:2402.02633.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3).

Rohan Kumar, Youngmin Kim, Sunitha Ravi, Haitian Sun, Christos Faloutsos, Ruslan Salakhutdinov, and Minji Yoon. 2024. Automatic question-answer generation for long-tail knowledge. *Preprint*, arXiv:2403.01382.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *Preprint*, arXiv:2406.01771.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024a. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models using code execution. *Preprint*, arXiv:2305.15014.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024b. Quantifying multilingual performance of large language models across languages. *Preprint*, arXiv:2404.11553.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024a. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024b. Modeling orthographic variation improves nlp performance for nigerian pidgin. *Preprint*, arXiv:2404.18264.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. *Preprint*, arXiv:1905.12688.

Joanito Agili Lopo and Radius Tanone. 2024. Constructing and expanding low-resource and underrepresented parallel datasets for indonesian local languages. *Preprint*, arXiv:2404.01009.

Stefano Lusito, Edoardo Ferrante, and Jean Maillard. 2023. Text normalization for low-resource languages: the case of ligurian. *Preprint*, arXiv:2206.07861.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. *Preprint*, arXiv:2210.11621.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023a. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, and et al. Stella Biderman. 2023b. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, and et al. David Ifeoluwa Adelani. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages. *Preprint*, arXiv:2302.08956.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *Preprint*, arXiv:2201.08277.

Wilhelmina Nekoto, Vukosi Marivate, Yoshihiko Matsumato, Kadima Tshibanda, and et al. Auer, Benjamin. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1710–1723.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. *Preprint*, arXiv:2404.02392.

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, and et al. Sebastian Ruder. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *Preprint*, arXiv:2305.06897.

Akintunde Oladipo, Mofetoluwa Adeyemi, and et al. Ahia. 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. Lama Ahmad. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fred Philippy, Siwen Guo, Shohreh Haddadan, Cedric Lothritz, Jacques Klein, and Tegawendé F. Bissyandé. 2024. Soft prompt tuning for cross-lingual transfer: When less is more. *Preprint*, arXiv:2402.03782.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *Preprint*, arXiv:2401.08429.

Xiaoyu Qiu, Yuechen Wang, Jiaxin Shi, Wengang Zhou, and Houqiang Li. 2024. Cross-lingual transfer for natural language inference via multilingual prompt translator. *Preprint*, arXiv:2403.12407.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, and et al. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. *Preprint*, arXiv:2305.10971.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, and et al. Maha Elbayad. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024a. Nlp progress in indigenous latin american languages. *Preprint*, arXiv:2404.05365.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Jugal Kalita. 2024b. Ethiomt: Parallel corpus for low-resource ethiopian languages. *Preprint*, arXiv:2403.19365.

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

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2023. Scaling laws for BERT in low-resource settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7771–7789, Toronto, Canada. Association for Computational Linguistics.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *Preprint*, arXiv:2402.01349.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Yudong Wang, Chang Ma, Qingxiu Dong, Lingpeng Kong, and Jingjing Xu. 2023b. A challenging benchmark for low-resource learning. *Preprint*, arXiv:2303.03840.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024b. A survey on multilingual large language models: Corpora, alignment, and bias. *Preprint*, arXiv:2404.00929.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, and et al. Zekai Shao. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. *Preprint*, arXiv:2402.14086.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. *Preprint*, arXiv:2401.10695.

Tong Zhan, Chenxi Shi, Yadong Shi, Huixiang Li, and Yiyu Lin. 2024. Optimization techniques for sentiment analysis based on llm (gpt-3). *Preprint*, arXiv:2405.09770.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Preprint*, arXiv:2306.13304.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, and et al. Daniel D'souza. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A NLP Tasks

Machine Translation

For machine translation, AFRIINSTRUCT includes FLORES (Goyal et al., 2021), MAFAND-MT (Adelani et al., 2022c), and MENYO (Adelani et al., 2021b). These datasets cover a variety of African languages and provide parallel sentences for training and evaluating machine translation models. FLORES is a benchmark dataset designed to evaluate machine translation systems on 101 languages including 13 African languages. MAFAND-MT contains professionally translated news articles across 16 African languages. MENYO focuses on the effects of various strategies on machine translation for African languages, including texts in English- Yoruba from various domains such as news articles, TED talks, movie and radio transcripts, science and technology texts, and other short articles.

Named Entity Recognition

AFRIINSTRUCT incorporates MasakhaNER2.0 (Adelani et al., 2022d), an extension of the original MasakhaNER dataset. This dataset covers 20 African languages and has been adapted for evaluating generative models. The task involves identifying and classifying named entities such as persons, organizations, and locations within text.

News Topic Classification

AFRIINSTRUCT utilizes MasakhaNEWS (Adelani et al., 2023), a multilingual news classification dataset covering 16 typologically diverse languages spoken in Africa, including English and French. The task is to classify a news article into one of seven categories: business, entertainment, health, politics, religion, sports, or technology.

Part-of-Speech Tagging

The dataset for part-of-speech tagging is MasakhaPOS (Dione et al., 2023a). This dataset includes tagged sentences in 9 African languages and is used to train models to identify the grammatical category of each word in a sentence, such as noun, verb, adjective, etc.

Question-Answering

In the realm of question-answering, AFRIINSTRUCT incorporates xP3 (Muennighoff et al., 2023b), a cross-lingual, open-retrieval question-answering dataset. It consists of a variety of examples across different African languages. The dataset is designed to evaluate models on their ability to retrieve and generate accurate answers from a given context.

Sentiment Analysis

For sentiment analysis, AFRIINSTRUCT includes AfriSenti (Muhammad et al., 2023a) and NollySenti (Shode et al., 2023). AfriSenti is a multilingual sentiment classification dataset for 14 African languages, designed to classify tweets as positive, negative, or neutral. NollySenti focuses on sentiment analysis(positive/negative) for Nollywood movie reviews, providing sentiment labels in five widely spoken Nigerian languages, covering a range of sentiment annotations in these languages.

Summarization

For summarization tasks, AFRIINSTRUCT includes XL-Sum (Hasan et al., 2021). XL-Sum is a multilingual summarization dataset curated from BBC news articles. It covers 10 African languages and aims to generate short summaries, typically one to two sentences, from given articles.

Benchmarking

AFRIINSTRUCT-Bench is introduced as a benchmark to evaluate the performance of language models on African languages in this paper. It includes:

Machine Translation

NTREX (Federmann et al., 2022) is designed for machine translation (MT) evaluation from English into 128 target languages. It was created by translating the WMT19 'newstest2019' test set into these languages, ensuring high-quality translations by professional native speakers.

Topic Classification

SIB-200 (Adelani et al., 2024a) is a large-scale multilingual topic classification dataset covering 205 languages and dialects, with a focus on African languages.

Question-Answering

AFRIQA (Ogundepo et al., 2023) is designed for cross-lingual open-retrieval question-answering tasks in African languages across 10 African languages. Due to the different languages supported by each dataset, the languages included in these tasks became Hausa, Igbo, Kinyarwanda, Swahili, Yoruba, and Zulu. For each supported languages, benchmarks were created, and a general benchmark was also created that includes various African languages.

B Complete Data Analysis on AFRIINSTRUCT-Data

After completing data preprocessing on AFRIINSTRUCT, we analyzed the dataset with the number of prompts and tokens using LlamaTokenizer. Figure 1 shows the distribution of tokens and prompts in African languages across each dataset.

Figure 2 and 3 provide an overview of the token and prompt counts across the various datasets included in AFRIINSTRUCT. These datasets span a wide range of NLP tasks, such as machine translation, news topic classification, part-of-speech tagging, sentiment analysis, and named entity recognition. The token and prompt counts exhibit significant variation among the datasets. The xP3 dataset stands out with the highest counts, boasting over 640 million tokens and 8 million prompts. AfriSenti also contributes a substantial amount of data, with 19 million tokens and 235,000 prompts. On the other hand, datasets like MasakhaPOS, NollySenti, and MENYO have considerably lower token and prompt counts, ranging from 1-2 million tokens and 6,000-16,000 prompts. AFRIINSTRUCT covers a diverse set of African languages, with FLORES and MAFAND offering the broadest coverage.

The AFRIINSTRUCT training set, compiled from ten source datasets, contains a substantial amount of data totaling over 870 million tokens and 17 million prompts. This dataset covers a diverse range of African languages and NLP tasks, providing a comprehensive resource for training African language models. The AfriBench evaluation set complements AFRIINSTRUCT by offering a balanced test set focused on six languages and general prompts for machine translation, question-answering, and topic classification. The combination of AFRIINSTRUCT and AfriBench supports the goal of training and evaluating African language models on a diverse set of tasks, advancing the state-of-the-art in NLP for low-resource languages.

C Prompt Template

For each task provided, we employed promoting format to adapt original dataset to fine-tuning. (Table 3)

D Baseline model description

mT0-xxl, derived by fine-tuning mT5-XXL (Muennighoff et al., 2023a), demonstrating strong cross-lingual instruction-following capabilities, even for languages not explicitly included in its training data, due to its multitask prompted dataset (xP3).

Aya-101 specializes in multilingual capabilities, supporting a diverse range of global languages. It interprets instructions from 101 languages, over half of which are categorized as lower-resourced. (Singh et al., 2024). Aya 23 extends Aya-101, sacrificing breadth in exchange for depth. Though Aya 23 supports only 23 languages, it brings the model capacity to a state-of-the-art level, benefiting approximately half of the world's population. (Aryabumi et al., 2024)

LLaMa2 is an open-source, decoder-only LLM trained on a massive dataset of text and code (Touvron et al., 2023). Its successor, LLaMa 3, benefits from roughly double the size of LLaMa 2's training dataset, resulting in enhanced capabilities for various natural language processing tasks.

GPT-3.5 Turbo and GPT-4o are both transformer-style LLMs from OpenAI, opti-

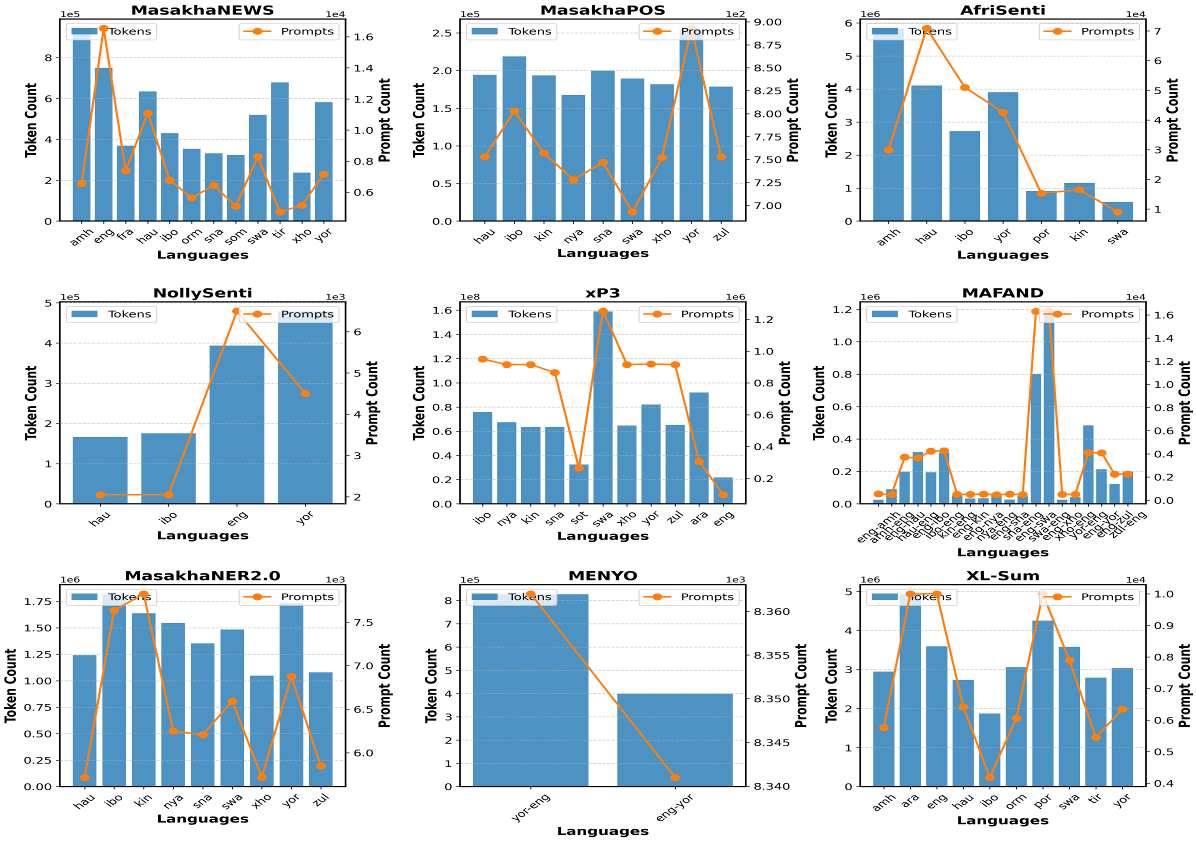

Figure 2: Data Statistics for Each of the Datasets

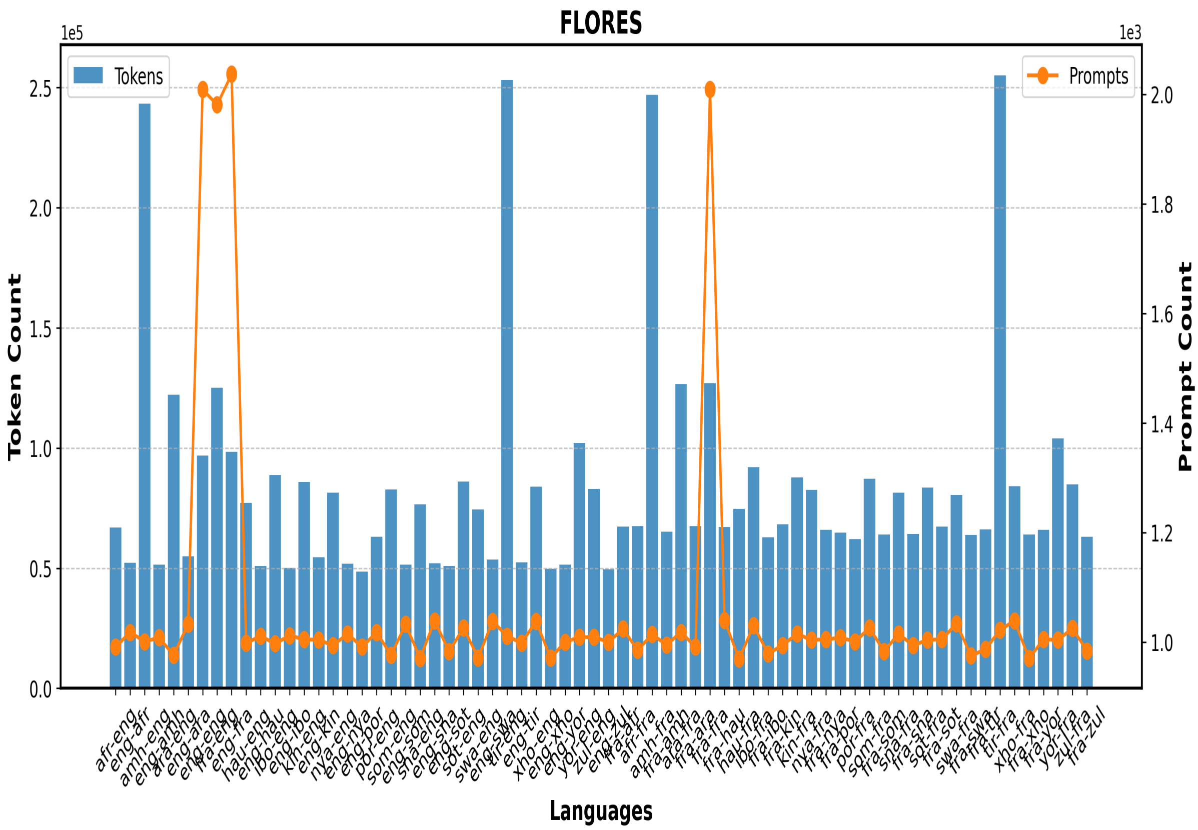

Figure 3: Data Statistics for the Entire Dataset

mized for faster response times and reduced costs (Ye et al., 2023). GPT-4 represents a more advanced iteration,

M2M100 (Fan et al., 2020) is an encoder-decoder model with a Seq2Seq transformer architecture designed for translation tasks. SMaLL-100 (Mohammadshahi et al., 2022) is a more efficient version of M2M100. NLLB-200 (Team et al., 2022) significantly expands the number of supported languages with an emphasis on low-resource languages.

Table 7 summarizes the performance of baseline models across our datasets.

E Win-Rate Evaluation

For further evaluation of better performance models, we used Win-Rate on mT0-xxl, Aya-101, and AFRIINSTRUCT-Model. Table 8, 9 show mT0-xxl and Aya-101 outperforms AFRIINSTRUCT-Model by 20 to 30 points on average.

File	our model	Aya-101	Tie
hau	22.21	51.12	26.67
ibo	24.07	43.55	32.38
kin	22.33	49.50	28.16
swa	17.54	49.12	33.33
yor	26.74	42.86	30.40
zul	22.43	48.17	29.40
general	23.1	50.9	26.0

Table 8: Win rate (%) comparison between AFRIINSTRUCT-Model(our model) and Aya-101

File	our model	mT0-xxl	Tie
hau	32.26	43.18	24.57
ibo	24.32	44.42	31.27
kin	19.85	54.34	25.81
swa	20.60	50.50	28.90
yor	24.25	45.35	30.40
zul	23.75	49.17	27.08
general	27.40	48.5	24.10

Table 9: Win-ate (%) comparison between AFRIINSTRUCT-model(our model) and mT0

F IrokoBench Evaluation

IrokoBench is a comprehensive evaluation suite specifically designed for benchmarking language models on African languages. (Adelani et al., 2024b). The dataset covers various tasks such as natural language inference (AfriXNLI), mathematical reasoning (AfriMGSM), and multi-choice knowledge-based QA (AfriMMLU) in 16 African languages.

We additionally evaluated our model, mT0-xxl, and Aya-101 with IrokoBench. The results(Table 10, 12, 13) indicate that the AfriInstruct-Model underperforms compared to mT0-xxl and Aya across various tasks. While the AfriInstruct-Model demonstrates some potential, it generally lags in option prediction accuracy and flexible match scores. However, it shows competitive performance in specific languages from the Afri-MMLU option prediction. These findings highlight the model's need for further improvement and more comprehensive training to better support African languages.

G Broader Evaluation

Although our goal is to develop a model specialized in African languages, evaluating AfriInstruct-Model-7B on English-centric benchmarks helps to assess its broader capabilities. We tested the model using MMLU (Table 14), MGSM (Table 15), and XNLI (Table 16), comparing it with Llama-2-7B. And we found that in MMLU, Llama-2-7B generally performed better. However, in XNLI, the AfriInstruct-Model-7B outperforms Llama-2-7B, which indicates that our model has gained more cross-lingual capabilities during the training.

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	avg
AFRIINSTRUCT-Model	43.83	36.33	34.33	33.33	34.00	35.50	34.50	33.00	33.00	34.83	33.00	34.17	34.50	33.67	34.17	34.17	34.17	33.83	34.68
mT0-xxl	62.50	60.33	58.17	39.50	56.83	56.67	50.83	33.50	53.33	49.17	54.50	55.33	57.67	49.67	40.50	54.83	51.33	54.50	52.18
Aya	61.50	60.17	57.83	43.00	56.33	53.83	46.50	33.17	44.33	52.17	56.00	54.50	54.50	47.50	35.33	53.33	48.67	54.83	50.75

Table 10: Afri-XLNI results in in-language: Option prediction accuracy per language

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	avg
AFRIINSTRUCT-Model	4.8	3.2	1.6	0.4	2.0	0.8	2.4	1.2	2.8	2.0	1.6	0.0	1.2	0.4	0.8	2.4	1.2	2.0	1.71
mT0-xxl	4.0	3.6	3.6	1.2	3.2	1.2	2.0	2.0	2.8	0.8	3.6	3.2	4.4	0.8	1.2	3.2	2.0	2.0	2.49
Aya	3.2	6.4	4.0	2.4	6.4	2.8	2.8	3.2	0.4	2.4	4.8	4.0	5.2	2.0	2.0	4.0	2.4	2.4	3.38

Table 11: Afri-MGSM results in in-language: flexible Match score per language

Model	amh	fra	eng		ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul
avg																			
AfriInstruct-Model	1.6	3.2	4.8	0.4	2.0	0.8	2.4	1.2	2.8	2.0	1.6	0.0	1.2	0.4	0.8	2.4	1.2	2.0	1.71
mT0-xxl	3.6	3.6	4.0	1.2	3.2	1.2	2.0	2.0	2.8	0.8	3.6	3.2	4.4	0.8	1.2	3.2	2.0	2.0	2.49
Aya	4.0	6.4	3.2	2.4	6.4	2.8	2.8	3.2	0.4	2.4	4.8	4.0	5.2	2.0	2.0	4.0	2.4	2.4	3.38

Table 12: Afri-MGSM results in in-language: flexible Match score per language

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	avg
AfriIT-Model	30.8	30.6	23.2	24.2	25.0	24.6	22.6	27.6	25.8	22.8	23.6	24.8	-	23.8	20.6	23.0	26.4	26.8	25.07
mT0-xxl	37.6	34.8	31.0	25.4	30.2	32.0	28.0	27.6	28.0	27.6	29.0	31.0	-	30.8	23.8	31.2	31.2	28.2	29.85
Aya	40.2	37.8	31.2	25.4	32.2	33.8	29.8	27.8	26.4	25.6	26.6	32.0	-	25.8	24.6	30.8	29.4	29.4	29.93

Table 13: Afri-MMLU results in in-language: Option prediction accuracy per language

Model	Humanities	Social Sciences	STEM	Other
Llama-2-7B	**0.3889 ± 0.0069**	**0.4605 ± 0.0089**	**0.3422 ± 0.0084**	**0.4699 ± 0.0089**
AfriInstruct-Model-7B	0.3107 ± 0.0067	0.3370 ± 0.0085	0.2915 ± 0.0081	0.3457 ± 0.0085

Table 14: MMLU Evaluation Results in Llama-2-7B and AfriInstruct-Model-7B

Model	Flexible Extract	Remove Whitespace
Llama-2-7B	**0.0720 ± 0.0164**	**0.0000 ± 0.0000**
AfriInstruct-Model-7B	0.0520 ± 0.0141	0.0360 ± 0.0118

Table 15: MGSM Evaluation Results in Llama-2-7B and AfriInstruct-Model-7B

Model	XNLI (Accuracy)
Llama-2-7B	0.5526 ± 0.0100
AfriInstruct-Model-7B	**0.5631 ± 0.0099**

Table 16: XNLI Evaluation Results in Llama-2-7B and AfriInstruct-Model-7B