AfroBench: How Good are Large Language Models on African Languages?

Jessica Ojo [1,3*], Odunayo Ogundepo [4,5*], Akintunde Oladipo [4,5*], Kelechi Ogueji[5*], Jimmy Lin [5], Pontus Stenetorp [6], David Ifeoluwa Adelani [1,2*]

[*]Masakhane NLP, [1]Mila - Quebec AI Institute & McGill University, [2]Canada CIFAR AI Chair, [3]Lelapa AI, [4]The African Research Collective [5]University of Waterloo, [6]University College London

Correspondence:{jessica.ojo, david.adelani}@mila.quebec

Abstract

Large-scale multilingual evaluations, such as MEGA, often include only a handful of African languages due to the scarcity of high-quality evaluation data and the limited discoverability of existing African datasets. This lack of representation hinders comprehensive LLM evaluation across a diverse range of languages and tasks. To address these challenges, we introduce AFROBENCH—a multi-task benchmark for evaluating the performance of LLMs across 64 African languages, 15 tasks and 22 datasets. AFROBENCH consists of nine natural language understanding datasets, six text generation datasets, six knowledge and question answering tasks, and one mathematical reasoning task. We present results comparing the performance of prompting LLMs to fine-tuned baselines based on BERT and T5-style models. Our results suggest large gaps in performance between high-resource languages, such as English, and African languages across most tasks; but performance also varies based on the availability of monolingual data resources. Our findings confirm that performance on African languages continues to remain a hurdle for current LLMs, underscoring the need for additional efforts to close this gap.[1]

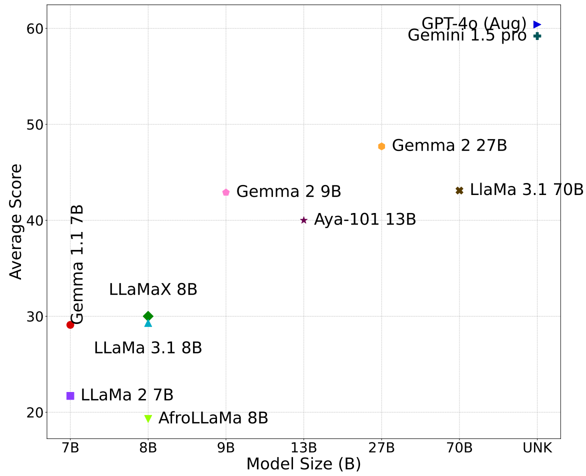

Figure 1: AFROBENCH average score on various LLMs

1 Introduction

Large language models (LLMs) have risen to the forefront of natural language processing (NLP) and have also become increasingly commercially viable. These models have empirically demonstrated strong performance across a variety of NLP tasks and languages (Brown et al., 2020; Lin et al., 2021; Chowdhery et al., 2022; Chung et al., 2022). However, their performance on low-resource languages (LRLs), such as African languages, is largely understudied. This is problematic because there is a significant disparity in the coverage of languages by NLP technologies. Joshi et al. (2020) note that over 90% of the world's 7000+ languages are under-studied by the NLP community. Ideally, approaches to enhance language understanding should be applicable to all languages.

While there have been evaluations of LLM performance in several languages (Ahuja et al., 2023a; Lai et al., 2023; Robinson et al., 2023), the evaluations have primarily focused on *closed models* like GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023). Megaverse (Ahuja et al., 2023b) extended the evaluation to more models such as PaLM 2 (Anil et al., 2023) and LLaMa 2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Gemma (Mesnard et al., 2024) and Gemini Pro (Team et al., 2023). However, previous evaluation face two main issues: (1) they cover only a few tasks for African languages, for example, Megaverse only evaluated on part-of-speech, named entity recognition, and cross-lingual question answering for African languages, due to *poor discoverability* of African languages benchmarks, *limited available evaluation data*, and *bias in the selection* of languages covered in the evaluation.[2] (2) Evaluation of LLMs needs

[1]https://mcgill-nlp.github.io/AfroBench/

[2]Belebele (Bandarkar et al., 2024) covers over 29 African languages, but Megaverse did not include any in their evaluation.

Findings of the Association for Computational Linguistics: ACL 2025, pages 19048–19095

Benchmark	# Tasks	# Datasets	# African Lang.	# LLMs	Closed LLMs evaluated	Dominant task(s)
ChatGPT-MT (Robinson et al., 2023)	1	1	57	1	GPT-3.5	MT
Mega (Ahuja et al., 2023a)	10	16	11	4	GPT-3, GPT-3.5-Turbo, GPT-4	POS, NER
Megaverse (Ahuja et al., 2024)	16	22	16	8	PaLM, GPT-3.5, GPT-4, Gemini Pro	POS, NER, XQA
SIB-200 (Adelani et al., 2024a)	1	1	57	2	GPT-3.5, GPT-4	Topic classification
Belebele (Bandarkar et al., 2024)	1	1	28	6	GPT-3.5-Turbo	QA
Uhura (Bayes et al., 2024)	1	2	6	6	Claude-3.5-Sonnet, GPT-4, 4o, o1-preview	QA
IrokoBench (Adelani et al., 2024b)	3	3	16	16	GPT-3.5,4,4o, Gemini-1.5-Pro, Claude OPUS	NLI, MMLU, Math.
AFROBENCH(Ours)	15	22	60	12	Gemini-1.5-Pro, GPT-4o	several

Table 1: **Overview of Related works that evaluated on African languages**. We included the number of tasks, datasets, African languages, LLMs evaluated, and the dominant tasks covering at least three African languages.

to be continuous, since many new LLMs have been released with improved multilingual abilities, but comprehensive evaluations for African languages remain limited.

In this paper, we address the challenges of previous large-scale LLM evaluation by introducing a new carefully curated benchmark known as **AFROBENCH which comprises 15 tasks, 22 evaluation datasets, and 64 indigenous African languages.** AFROBENCH consists of nine natural language understanding tasks, six text generation tasks, six question answering (QA) tasks, and one mathematical reasoning task. Finally, we created a **new evaluation dataset**, AFRIADR, a dataset for restoring diacritic marks and tonal accents in African language texts. Leveraging AFROBENCH, we conduct an extensive analysis of LLMs performance on African languages from different language families and geographical locations.

For our evaluation, we compute the average performance score over the 15 tasks covered in AFROBENCH. Additionally, we introduce AFROBENCH-LITE that only covers a subset of seven tasks and 14 diverse languages from AFROBENCH, which reduces the evaluation cost for a newly introduced LLM on our leaderboard. Figure 1 shows our evaluation on AFROBENCH. We find that proprietary models such as `GPT-4o` and `Gemini-1.5 pro` achieve a +13 score improvement over `Gemma 2 27B`, our best-performing open model. We also compared the performance on English to 14 African languages, finding that `GPT-4o` and `Gemma 2 27B` achieve better performance than African languages by over +25 and +40 points, respectively. This shows that the gap in the multilingual abilities of open models is wider than that of proprietary models. Finally, we compare the performance of LLMs to fine-tuned models based on AfroXLMR (Alabi et al., 2022), AfriTeVa V2 T5 model (Oladipo et al., 2023) and NLLB (NLLB Team et al., 2022) when training data is present. Results show that prompting LLMs often yields lower average performance than the fine-tuned baselines. Our findings show that greater effort is needed to close the performance gap between LLMs on high-resource and African languages

2 Related Work

Large Language Model Evaluation: Accurate and reproducible evaluation of LLMs is important as the number of released models continues to grow. As these models are integrated into various applications, developing robust evaluation frameworks becomes critical for understanding their true capabilities and limitations. As a result, the community has worked on developing evaluation frameworks (Gao et al., 2024; Fourrier et al., 2023; Liang et al., 2023), leaderboards (Chiang et al., 2024; Srivastava et al., 2023; Fourrier et al., 2024) and benchmarks (Adelani et al., 2024b; Zhou et al., 2023; Hendrycks et al., 2021). While each of these evaluation tools focuses on assessing specific aspects of language model capabilities, from basic linguistic understanding to complex reasoning tasks, the development of truly comprehensive benchmarks remains a significant challenge (Ruder, 2021; Biderman et al., 2024). These challenges stem from the complex nature of language understanding and the stochastic nature of language models.

Multilingual LLM Benchmarks: Benchmarks serve as a standard for measuring how systems have improved over time on specific tasks and metrics. In the context of LLMs, multilingual benchmarks are crucial for assessing both the quality and practical utility of these models across diverse languages and tasks. Our primary focus lies in understanding LLM performance specifically for African languages, with several notable benchmarks having emerged in recent years to address this need. ChatGPT-MT (Robinson et al., 2023) evaluated the translation capability of GPT-4 and they find that it demonstrates strong performance on high-resource languages, but performance on low-resource languages is subpar. Belebele (Bandarkar et al., 2024) is a QA task covering 122 languages, including 28

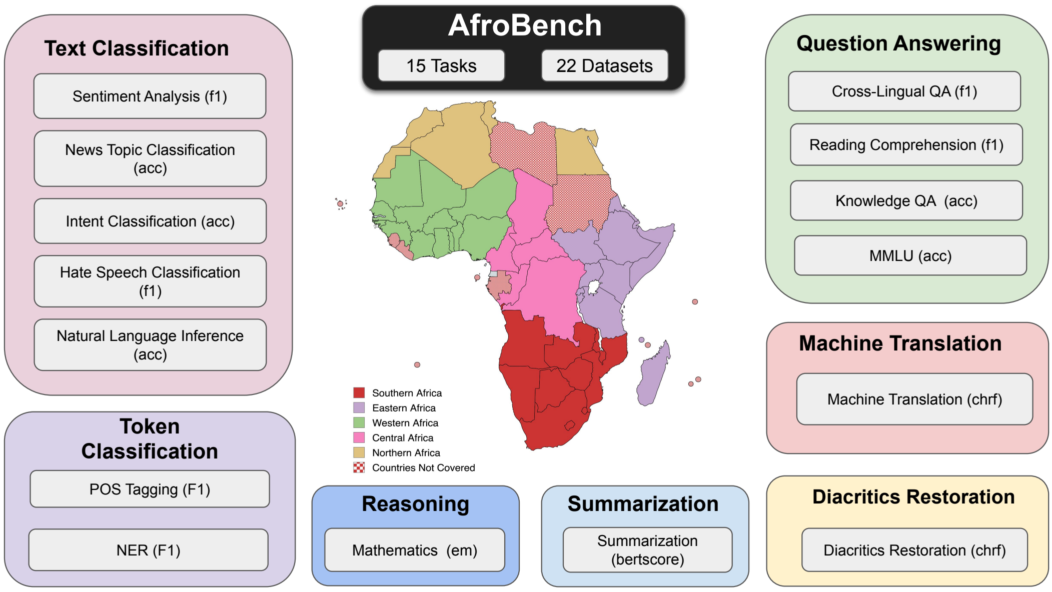

Figure 2: **AFROBENCH: A comprehensive benchmark for evaluating performance of LLMs on African Language tasks**. The benchmark features 15 distinct tasks across 22 datasets and 64 indigenous African languages. The benchmark covers diverse tasks with geographical coverage spanning different regions in Africa.

African languages, for evaluating reading comprehension abilities of LLMs. Mega (Ahuja et al., 2023a) and Megaverse (Ahuja et al., 2024) are multi-task, multilingual and multi-modal benchmarks in 83 languages including 16 African languages. Table 1 summarizes the related works.

While these existing benchmarks have provided valuable insights, they collectively highlight a pressing need for more comprehensive evaluation that encompasses a broader range of African languages and diverse tasks. Our research, through the development of AFROBENCH, addresses this gap by building upon and complementing existing work. We introduce a robust evaluation framework that assesses LLM performance across 64 African languages, evaluating capabilities across 15 distinct tasks. This broader scope enables a more nuanced and thorough understanding of LLM capabilities in African language contexts.

3 AfroBench

AFROBENCH is a comprehensive LLM evaluation benchmark designed to assess both proprietary and open LLMs across diverse Natural Language Processing (NLP) tasks in African languages. As shown in Figure 2, the benchmark encompasses 15 distinct tasks, spanning Natural Language Generation (NLG) and Natural Language Understanding (NLU), incorporating 22 curated datasets in 64 African languages. These evaluation tasks extend beyond traditional NLP benchmarks, such as text classification and named entity recognition, to include more challenging tasks such as mathematical reasoning and knowledge QA.

Each task within AFROBENCH has been carefully selected to assess different aspects of language model capabilities, from basic linguistic capabilities to more complex reasoning abilities. AFROBENCH also provides valuable insights into model behavior across different African language families and their unique linguistic features. All tasks within AFROBENCH are evaluated using both zero-shot and few-shot prompting to guide model responses. To ensure consistent and reliable evaluation, we implement task-specific response constraints to facilitate systematic extraction and analysis of model outputs. For completion, we compare against state-of-the-art (SoTA) encoder-only and encoder-decoder architectures that have previously demonstrated superior performance on individual tasks within the benchmark. This enables us to directly compare the performance of specialized models to general-purpose LLMs. Table 2 summarizes the tasks, the datasets used, number of languages covered, and total sample size.

3.1 Languages

We cover 64 African languages from seven language families (Afro-Asiatic, Atlantic-Congo, Austronesian, Indo-European, Mande, Nilotic, and English-Creole). 40 languages are from the Atlantic-Congo family, 12 from the Afro-Asiatic

Task	Dataset	Total Size	No. of Lang.	Per. Lang. size
POS	MasakhaPOS	12,190	20	500–700
NER	MasakhaNER-X	18,192	20	900–1000*
TC	SIB-200	11,220	55	204
	MasakhaNEWS	6,242	16	200–948
SA	AfriSenti	37,670	15	950–4500‡
	NollySenti	2,500	5	500
Intent	Injongo-Intent	10,880	17	640
Hate	AfriHate	14,250	15	323–1600
NLI	AfriXNLI	9,600	16	600
XQA	AfriQA	3,107	9	250–500
RC	Belebele	27,900	31	900
	NaijaRC	357	3	80–190
QA	Uhura-Arc-Easy	3,257	7	300–500
MMLU	AfriMMLU	8,500	17	500
	MMMLU	42,126	3	14042
Math	AfriMGSM	4,500	18	250
MT	Flores-200	58,696	58	1012
	MAFAND	29,155	21	1000–2000
	NTREX	48,000	24	2000
	Salt	3,500	7	500
Summ	XLSum	25,769	12	500–1300¶
ADR	AfriADR	7,567	5	1400–1600

Table 2: **AfroBench data statistics:** We detail the dataset evaluated per task, test set size and number of languages for each dataset as well as the range of sample per language, *excl. amh: 500 & luo: 185 (in MasakhaNER-X), ‡excl. tso: 254 (in AfriSenti), and ¶excl. arb: 4689 & eng: 11,535. (in XLSum). The tasks covered in the **Lite** version is highlighted in Grey.

family, seven from the Nilotic family, 2 Indo-European, 2 Creole languages, and 1 Austronesian language. Figure 2 shows the geographical distribution of the languages covered in AFROBENCH and a full list of languages is provided in Appendix F.

3.2 Evaluation tasks

Our evaluation spans multiple datasets across 15 NLP tasks. While some of these multilingual datasets cover languages from multiple continents, we focus specifically on the African language subsets, along with select high-resource languages (English, French, Portuguese, and Arabic), due to their widespread use in various African regions. Table 2 details the sample size and the number of languages evaluated per task per dataset. We present a breakdown of the tasks, sub-tasks, and specific datasets contained in AFROBENCH.

3.2.1 Text Classification

Sentiment Classification (SA): We evaluate NOLLYSENTI (Shode et al., 2023) and AFRISENTI (Muhammad et al., 2023). AFRISENTI is a tweet sentiment analysis dataset in 14 African languages, while NOLLYSENTI focuses on sentiment in movie reviews in 4 African languages.

Topic Classification (TC): We evaluate SIB-200 and MASAKHANEWS (Adelani et al., 2023) which cover 53 and 14 African languages, respectively. The topic categories could be general topic such as *business*, *entertainment*, and *health*.

Intent Classification: INJONGO-INTENT (Yu et al., 2025) is an intent classification dataset in 16 African languages. The goal is to classify an utterance into one of 40 intent types from different domains such as *Banking* (e.g. "freeze account"), *Home* (e.g. "play music"), *Kitchen and Dining* (e.g. "cook time"), and *Travel* (e.g. "plug type").

Hate Speech detection: AFRIHATE (Muhammad et al., 2025) is a multilingual hate speech and abusive language dataset covering 15 African languages for tweets. Each tweet can be categorized as *abusive*, *hate* or *neural* label.

Natural Language Inference (NLI): AFRIXNLI(Adelani et al., 2024b) is a dataset collection in 16 African languages where each sample is a pair of sentences (a premise and a hypothesis) and the task is to classify each pair as an *entailment*, *contradiction*, or *neural* pair.

3.2.2 Token Classification

Named Entity Recognition (NER): We evaluate entity recognition for 20 African languages on the MASAKHANER-X (Ruder et al., 2023)—an extension of MASAKHANER dataset (Adelani et al., 2021, 2022b) which converts NER tags from CoNLL format into a text generation task of predicting entities with a delimiter "$" between them.

Part-of-Speech Tagging (POS): MASAKHAPOS(Dione et al., 2023) is a part-of-speech tagging dataset in 20 African languages created from news articles. Each token is categorized into one of the 17 POS tags.

3.2.3 Reasoning:

Mathematical reasoning (Math): We evaluate on AFRIMGSM (Adelani et al., 2024b), an extension of the MGSM dataset to 17 African languages. Each question is at grade school level with a numerical answer.

3.2.4 Question Answering

Cross-Lingual Question Answering (XQA): AFRIQA (Ogundepo et al., 2023) is a cross-lingual QA task with questions in 10 African languages and context passages in English or French. The goal is to extract the correct answer span from the text, similar to a cross-lingual reading comprehension.

Reading Comprehension (RC): We evaluate on NAIJARC (Aremu et al., 2024), a multiple-choice reading comprehension dataset in 3 African languages and BELEBELE (Bandarkar et al., 2024), a multiple-choice reading comprehension dataset in 122 languages, including 29 African languages.

Knowledge QA: We focus on two human-translated **MMLU** datasets: OPENAI-MMLU and AFRIMMLU (Adelani et al., 2024b) covering 3 and 16 African languages respectively.[3] Both tasks span multiple subjects and follow a four-option multiple-choice format. However, AFRIMMLU covers only five subjects. We also extend our evaluation to the human-translated version of the *scientific* **Arc-Easy** benchmark in 6 African languages UHURA (Bayes et al., 2024).

3.2.5 Text Generation

Machine translation (MT): Our MT benchmark includes the following datasets: FLORES(Goyal et al., 2022), MAFAND(Adelani et al., 2022a), NTREX-128(Federmann et al., 2022) and SALT (Akera et al., 2022) covering 57, 21, 23 and 7 translation directions into African languages. All translations are from English, except for MAFAND, which includes a few languages with French as the source.

Summarization (Summ): Given a news article, our goal is to generate its summary based on the widely-used XL-SUM dataset (Hasan et al., 2021) covering 10 African languages.

Automatic Diacritics Restoration (ADR): This is a **new benchmark** we introduce called **AFRI-ADR**. Given a sentence in a language, say "*Sugbon sibesibe, Mama o gbagbo*" (in Yorùbá), the model's goal is to add the missing tonal marks and accents, say "*Ṣùgbọ́n síbẹ̀síbẹ̀, Màmá ò gbàgbọ́*". We cover 5 African languages for this task: *Ghomálá'*, *Fon*, *Igbo*, *Wolof*, and *Yorùbá*. To create AFRIADR, we selected the five languages with extensive use of diacritics from MAFAND MT dataset, then, we strip all accents and diacritics on each sentence, and use it as the "source" text, and the "target" as the fully diacritized texts. Table 3 provides the data size and an example sentence for each language in AFRIADR.

[3]https://huggingface.co/datasets/openai/MMMLU

Lang.	Size	Example sentence
Ghomálá'	1430	**Input:** A jwə guŋ tsə awɛ a lə nəŋ kwitə **Target:** Â jwə́ guŋ tsə́ awɛ́ a lə nə́ŋ kwítə́
Fon	1579	**Input:** Din ɔ, nu lɛɛ bi jɛwexo. **Target:** Din ɔ, nú lɛ́ɛ bǐ jɛ wexo.
Igbo	1500	**Input:** Akuko ndi ga-amasi gi: **Target:** Akụkọ ndị ga-amasị gị:
Wolof	1500	**Input:** Naari taggatkat lanu yu xaran lu kawe. **Target:** Ñaari tàggatkat lañu yu xarañ lu kawe.
Yorùbá	1558	**Input:** Isokan awon Oniroyin naa fe oro naa loju: **Target:** Íṣọ̀kan àwọn Oníròyìn náà fẹ ọ̀rọ̀ náà lójú:

Table 3: **AfriADR dataset**: Language, test size, and Example sentence

3.3 AfroBench-Lite: A cost-effective bench

Following the idea of Global-MMLU-Lite (Singh et al., 2024), which aims to create a cost-effective benchmark with fewer languages and samples. We introduce AFROBENCH-LITE, a subset of AFROBENCH covering 14 languages and seven datasets (and tasks): SIB-200 (TC), INJONGO (Intent), AFRIXNLI (NLI), BELEBELE (RC), AFRIMMLU (MMLU), AFRIMGSM (Math), and FLORES (MT). The selected languages are highly typologically diverse, and have varying resource levels (Kudugunta et al., 2023). These languages include: *English*, *Kiswahili*, *Kinyarwanda*, *Hausa*, *Amharic*, *isiXhosa*, *chiShona*, *isiZulu*, *Igbo*, *Yorùbá*, *Sesotho*, *Lingala*, *Oromo*, *Luganda*, and *Wolof*.

4 Experimental setup

4.1 Evaluation Framework

We model all tasks as text-generation problems, where we combine inputs with prompts to guide language models in generating outputs under specific constraints. To ensure robust evaluation, we employ multiple prompts for each task with with few-shot and zero-shot examples, which help maintain consistency and minimize potential bias across different models.

Our evaluation framework is fully integrated with Eleuther LM Evaluation Harness (Gao et al., 2024) with custom evaluation scripts to run open-source models.[4] However, for the proprietary models, we developed a custom framework for prompting various LLMs via API including open models available on TogetherAI API.[5] These tools are open source, easily accessible, and reproducible. Details

[4]https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/afrobench

[5]https://github.com/McGill-NLP/AfroBench/tree/main/prompt_with_API

of the custom framework and its integration with the Eleuther LM Evaluation Harness are provided in Appendix C.

4.2 Fine-tuned baselines

For the tasks with available training data, we use available task-specific trained models, such as NLLB-200 3.3B for MT, and fine-tuned multilingual encoders or encoder-decoder T5 models on applicable datasets. We fine-tune AfroXLMR (Alabi et al., 2022)—one of the SoTA BERT-style encoders for African languages for each NLU task. For summarization and ADR, we fine-tune AfriTeVa V2 Large (Oladipo et al., 2023) on the available training data of each task. While AfriTeVa V2 outperformed mT5 (Xue et al., 2021) overall, its tokenization failed for Fon language, so we fine-tune mT5-large, which has a more diverse tokenizer, for the language.

4.3 LLMs Evaluated

We evaluate two broad categories of Large Language Models (LLMs): **Open Models** and **Closed Models**. We evaluate 10 open models: `LLaMa 2 7B` (Touvron et al., 2023), `Gemma 1.1 7B` (Mesnard et al., 2024), LLAMA 3 series (3 8B, 3.1 8B & 3.1 70B) (Dubey et al., 2024), `LLaMaX 8B` (Lu et al., 2024) (an adapted LLaMa 3 8B to 100 languages), `AfroLlama 8B`(an adapted LLaMa 3 8B to Swahili, Xhosa, Zulu, Yoruba, Hausa and English languages),[6] GEMMA 2 (9B & 27B) (Riviere et al., 2024), and `Aya-101`(an instruction-tuned mT5 encoder-decoder model on massively multilingual prompted dataset). Finally, we evaluate on two popular proprietary models: `GPT-4o` and `Gemini-1.5 pro` (Reid et al., 2024). A full description of the LLMs evaluated is provided in Appendix B.

Prompts used for evaluation We make use of *five* different prompts in the evaluation of each task except the text generation tasks, and we report the best prompt in the paper. For text generation tasks, we reduce the number of prompts to *three* since the generation is often time-consuming and computationally expensive, especially for summarization tasks. Moreover, we observe that performance is less sensitive to prompt variations, unlike natural language understanding (NLU) tasks. The prompt templates are provided in Appendix H.

Few-shot evaluation We restrict the few-shot evaluation to top-performing open and proprietary models. We fixed the number of examples to *five*, except for AfriMGSM whose number of examples is *eight*.[7]

5 Results

5.1 AfroBench Evaluation

Table 4 presents the overall results across all 15 tasks and 22 datasets. We report only the best prompt results. Average results across all five prompts, along with confidence intervals, are provided in Appendix D.

Our **first** observation is that closed models such as `GPT-4o` and `Gemini-1.5 pro` outperform the best open model, `Gemma 2 27B` by $+12$ or more points on average. This shows that the gap in performance is wider for low-resource African languages compared to high-resource languages like English, especially with open models. **Secondly**, we find that the performance gap varies across tasks. Knowledge-intensive and reasoning tasks such as ARC-EASY, MMLU, MATH have the largest gaps of $+29.4$, $+19.9$, $+22.6$ points, respectively, when comparing `GPT-4o` to `Gemma 2 27B`. Overall, performance improves with newer LLM versions (e.g. `Gemma 1.1 7B` vs. `Gemma 2 9B` and `LLaMa 2 7B` vs. `LLaMa 3.1 8B`) and larger model sizes (`Gemma 2 9B` and `Gemma 2 27B`). This suggests that newer iterations of models are getting better on low-resource languages, though improvements on knowledge-intensive tasks remain limited. **Finally**, while LLMs have made significant progress, they still lag behind *fine-tuned baselines* (**FT. AVG**) when training data is available for a task. The performance gap is around $+11.5$ on average, highlighting the continued value of curating annotated datasets for low-resource languages, as LLMs still underperform. We provide task and per-language results in Appendix A and I.

5.2 AfroBench-Lite Evaluation

In the AFROBENCH-LITE evaluation, we restrict evaluation to seven LLMs, and seven tasks, comparing their performance to English.

Large gap in performance when compared to English One striking observation is that open models such as `LlaMa 3.1 70B` and `Gemma 2 27B` have competitive performance to closed models

[6]`https://huggingface.co/Jacaranda/AfroLlama_V1`

[7]8-shot samples is the standard setting for MGSM datasets

Tasks	POS	NER	SA	TC	Intent	Hate	NLI	XQA	RC	Arc-E	MMLU	Math	MT		Summ	ADR	ALL	FT.
	natural language understanding							*QA*		*knowledge*		*reasoning*	*text generation*					
Metrics	acc	F1	F1	acc	acc	F1	acc	F1	F1	acc	acc	EM	ChrF		BertScore	ChrF	AVG	AVG
Fine-tuned baselines													*en/fr-xx*	*xx-en/fr*				
AfroXLMR	**89.4**	**84.6**	**72.1**	74.4	**93.7**	**77.2**	61.4											
mT5/AfriTeVa V2 1B								52.5	N/A	N/A	N/A	N/A			**72.3**	**79.4**		**70.4**
NLLB 3.3B													**40.4**	**47.8**				
Prompt-based baselines																		
open models																		
Gemma 1.1 7B	38.6	27.9	43.3	45.3	9.4	24.3	34.0	17.4	38.1	32.2	28.6	4.6	11.7	9.7	49.1	50.8	29.1	29.7
LLaMa 2 7B	27.9	15.6	42.3	19.4	1.5	21.9	33.8	13.7	24.3	23.3	25.6	2.0	10.5	20.3	46.9	30.4	22.5	22.2
LLaMa 3 8B	48.5	22.7	43.6	37.0	2.1	27.8	35.4	12.6	27.6	32.0	27.4	5.1	15.9	27.7	66.2	26.1	28.6	28.6
LLaMaX 8B	41.6	0.0	51.9	49.8	5.6	28.6	40.8	2.2	29.7	39.9	28.3	4.0	22.7	35.0	50.7	49.4	30.0	29.0
LLaMa 3.1 8B	47.1	11.5	50.5	46.7	6.0	23.6	36.6	21.8	39.5	32.8	31.4	6.8	16.4	28.5	43.7	25.9	29.3	28.1
AfroLLaMa 8B	0.0	3.5	43.4	19.8	0.8	18.4	35.9	21.8	24.1	37.2	25.8	3.7	8.4	9.5	50.8	5.2	19.3	17.6
Gemma 2 9B	51.9	40.3	60.0	56.0	29.2	29.9	40.3	45.9	51.6	53.4	37.1	18.7	24.8	29.1	66.1	51.6	42.9	42.9
Aya-101 13B	0.0	0.0	63.4	70.3	42.4	31.0	51.5	**62.5**	60.7	59.6	30.9	4.4	23.4	37.9	52.4	50.4	40.1	37.7
Gemma 2 27B	55.1	50.8	63.4	62.4	33.0	45.5	42.8	50.5	53.9	56.3	40.5	27.0	27.9	32.9	66.4	55.1	47.7	48.3
LlaMa 3.1 70B	54.1	14.4	52.2	57.7	34.0	49.0	38.0	44	49.7	54.9	39.9	23.2	25.1	37.9	67.6	51.7	43.3	42.6
proprietary models																		
Gemini 1.5 pro	60.8	41.8	68.3	**76.7**	74.3	62.1	62.0	40.5	52.7	84.8	57.6	**52.3**	37.6	41.7	66.7	**55.6**	58.5	58.9
GPT-4o (Aug)	62.8	40.7	68.0	74.8	74.0	63.5	**64.3**	43.4	**69.2**	**85.7**	**60.4**	49.8	35.1	40.7	66.5	54.9	**59.6**	58.1

Table 4: **AfroBench Evaluation Results on Fine-Tuned Models and LLMs.** We cover 15 tasks, 22 datasets, and 64 African languages in the evaluation. The best closed and open LLMs are highlighted in Cyan . We **bolden** the best result per task in each column. We provide average on **ALL** tasks and on those with fine-tuned baselines (**FT**)

Model	Lang	Intent	TC	NLI	RC	MMLU	Math	MT *en/fr-xx*	AVG
Gemma 1.1 7B	*eng*	72.1	86.3	59.2	87.9	44.6	20.8	26.1	56.7
	africa	10.2	42.0	34.6	34.1	27.3	5.1	10.9	23.5
Gemma 2 9B	*eng*	36.3	82.5	70.7	**93.7**	69.8	68.8	67.9	70.0
	africa	27.8	64.0	40.9	49.3	36.1	21.7	37.2	39.6
Aya-101 13B	*eng*	78.0	82.8	67.0	86.1	42.8	11.6	64.2	61.8
	africa	40.2	76.0	52.4	59.7	30.3	4.9	31.8	42.2
Gemma 2 27B	*eng*	84.0	**89.3**	67.8	93.4	75.6	85.6	68.5	80.6
	africa	31.4	66.6	43.7	52.1	40.8	30.6	39.1	43.5
LLaMa 3.1 70B	*eng*	84.5	88.3	59.5	93.2	76.4	86.8	**71.6**	80.0
	africa	36.9	61.9	38.4	45.3	40.6	26.5	29.6	39.9
Gemini 1.5 pro	*eng*	**86.8**	88.7	88.5	69.6	**88.8**	86.8	69.1	82.6
	africa	75.6	81.3	63.6	54.4	62.6	57.7	44.2	62.8
GPT-4o (Aug)	*eng*	86.2	89.2	**89.2**	84.3	88.0	**88.8**	70.2	**85.1**
	africa	78.4	83.0	66.3	70.3	63.1	57.3	43.6	66.0

Table 5: **AfroBench-Lite Evaluation**: LLM baselines on 7 datasets spanning 14 African languages. Tasks were selected for broad NLP coverage, prioritizing language consistency. The best score per task is in **bold**.

on English with -5 to -2 point performance gap. However, when compared to African languages, GPT-4o and Gemini-1.5 pro achieve an average score better than Gemma 2 27B by over 20 points on AFROBENCH-LITE. These results suggest that current LLMs, especially open models, are biased toward *English* and a few high-resource languages. Adapting LLMs for a region of African languages could help bridge the gap. For instance, we see that continually pre-training LLaMa 3 8B, that resulted in LLaMaX 8B yields a modest performance gain of $+1.4$ points or more compared to vanilla LLaMa 3 8B in Table 4. However, to further improve performance, better adaptation techniques are needed.

Performance varies across languages Figure 3 shows the results for per-language performance scores of 14 languages in AFROBENCH-LITE. Our results show a correlation between performance and available monolingual text on the web (Kudugunta et al., 2023). We find that Swahili (swa) with over 2.4GB of monolingual text has the highest performance among the African languages, while Wolof with the smallest monolingual data (5MB) has the lowest performance. While these data size estimates are approximate, it shows that there is a need to invest more in developing language texts for many African languages for them to benefit in the LLM age. For most languages, GPT-4o gives the best overall results except for Amharic (amh) where Gemini-1.5 pro was better. For the open models, Gemma 2 27B outperforms other open models on 8 of the 14 languages, even better than LlaMa 3.1 70B, which is more than twice its number of parameters. Although Aya-101covers 100 languages in its pre-training and often achieves better performance on NLU tasks in AFROBENCH-LITE, it often struggles with math reasoning and MMLU, resulting in lower overall performance.

5.3 Few-shot Results

Table 6 shows the results of zero-shot and few-shot evaluation for three LLMs: Gemma 2 27B, Gemini-1.5 pro and GPT-4o. The benefit of few-shot varies for different LLMs and tasks. For GPT-4o, we find that across all tasks, there is an average improvement of $+1.8$ while the other LLMs

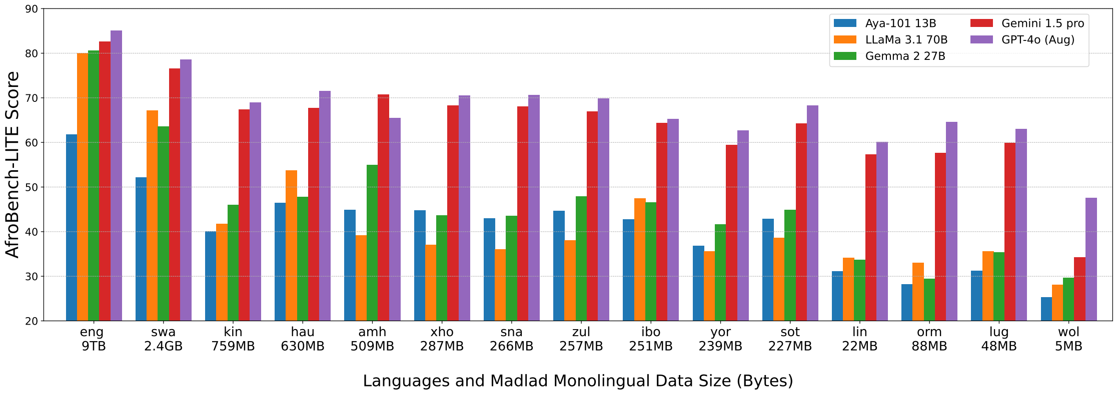

Figure 3: AfroBench-Lite performance of various models across African languages, plotted against the availability of monolingual data (MADLAD byte size).

Tasks	# shots	POS	NER	SA	TC	Intent	Hate	NLI	XQA	RC	MMLU	Math	MT *en/fr-xx*	MT *xx-en/fr*	SUMM	ADR	AVG
Gemma 2 27B	0-shot	55.1	**50.8**	58.6	57.3	35.2	45.5	42.8	50.5	53.6	39.9	27.0	32.4	32.4	66.4	55.1	46.8
	5-shot	43.9	14.5	59.7	62.5	56.7	57.3	56.0	**52.4**	58.3	44.8	27.5	22.7	34.9	55.5	31.2	45.2
Gemini 1.5 pro	0-shot	60.8	41.8	62.6	74.5	**74.3**	62.1	62.0	40.5	53.0	**60.2**	52.3	35.4	41.7	66.7	55.6	56.2
	5-shot	33.2	37.4	**64.5**	**77.3**	73.4	64.1	35.9	28.7	24.4	46.0	49.0	37.4	43.1	**70.4**	**63.4**	49.9
GPT-4o (Aug)	0-shot	**62.8**	40.7	62.6	72.5	74.0	63.5	**64.3**	43.4	69.1	60.0	49.8	31.5	41.0	66.5	54.9	57.1
	5-shot	62.4	45.0	62.3	72.9	71.6	**69.3**	64.2	40.0	**71.9**	59.7	**54.7**	**33.9**	**43.3**	67.9	62.7	**58.8**

Table 6: **Few-shot Evaluation.** The better score between each model's 0-shot and few-shot is in underlined.

dropped in performance on average. The tasks that benefit the most from the few-shot examples are math reasoning, hate speech detection and ADR with respective improvements of +4.9, +5.8, and +7.8 points. The result shows that few-shot examples are important for teaching LLM a new task it is unfamiliar with such as ADR since the rules of adding diacritics are not provided during the zero-shot, therefore, five examples provide a demonstration to the LLMs on how to perform the task especially for low-resource languages like Ghomálá' and Fon, which have very limited monolingual data online. These two languages improved by +16.4 and 7.2 points, respectively. While the other languages such as Igbo, Wolof and Yorùbá achieved over +5.0 boost in chrF scores. Similarly, for `Gemini-1.5 pro`, we observed a consistent performance boost for ADR with five demonstration examples.

For both `GPT-4o` and `Gemini-1.5 pro`, there is a significant boost in performance across all the text generation tasks we evaluated, which shows that the current models have weaker generative capabilities in these low-resource languages, unless provided with few-shot examples. For Hate speech, we provided a detailed explanation of the distinction between "abusive" content and "hate" in the prompt, but this is often confusing even for native speakers of the language, who often need examples of such sentences to improve annotation. We found that LLMs also require such additional examples to be able to better predict if a tweet is offensive. In general, `Gemma 2 27B` improved on several NLU tasks but showed no gains from few-shot examples in token classification, math reasoning, summarization and ADR tasks.

6 Discussion

6.1 Prompt Variability

In our evaluation, we present results for the `Best prompt` rather than the `Average results` over several prompts to ensure no LLM is penalized due to sensitivity to prompt templates. Here, we analyze the difference in the performance scores between the Best prompt and the average over five prompts (or three prompts for the NLG tasks).

Figure 4 presents results from our analysis across 18 tasks. Our **first observation** is that LLMs are not sensitive to different prompts when evaluating text generation tasks, all LLMs exhibit less than a 6 point difference, and the task that is the least sensitive is machine translation (FLORES). The **second observation** is that `Gemini-1.5 pro` exhibits the lowest sensitivity to different prompt templates on average. The gap in performance across different prompts is often small for several NLU tasks. Interestingly, we find that `GPT-4o` is highly sensitive to prompts for a tasks such as hate speech, cross-

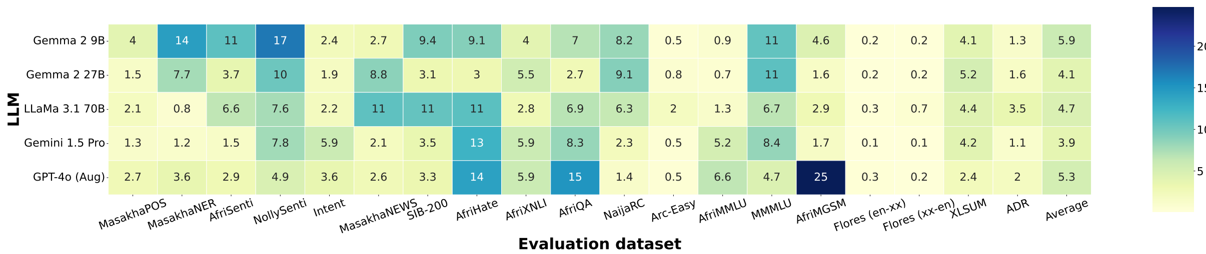

Figure 4: **Prompt Variability**: Heatmap of the difference between the Best and Average prompt results.

Lang.	Shot	Input / Output
ADR: `Gemini-1.5 pro`		
	Input:	A jwə guŋ tsə awɛ a lə nəŋ kwitə
bbj	**Target:**	Â jwə́ guŋ tsə́ awɛ́ a lə nə́ŋ kwítə́
	0-shot	à jwɛ̆ gûŋ tsə áwɛ́ à lə nə̂ŋ kẅitə (ChrF: 21.4)
	5-shot	a jwə́ guŋ tsə́ awɛ́ a lə nə́ŋ kwitə (ChrF: 81.6)
Hate speech: `Gemini-1.5 pro`		
	Input:	oloriburuku ni pupo ninu awon igbo yii
yor	**English:**	these igbo people are brainless
	Label:	hate
	0-shot	abuse
	5-shot	hate
Math reasoning: `GPT-4o`		
yor	**Input:**	Ryan gbin òdòdó 2 ní ojúmọ́ sí inú ọgbaà rẹ̀. Lẹ́yìn ọjọ́ 15, òdòdó mélòó ní ó ní tí 5 ò bá wù?
	English:	Ryan plants 2 flowers a day in his garden. After 15 days, how many flowers does he have if 5 did not grow?
	Answer:	25
	0-shot	ryan ní òdòdó 30 tí ó bá ń gbin 2 ní ojúmọ́
	8-shot	ìdáhùn: ryan gbin òdòdó 2 ní ojúmọ́. lẹ́yìn ọjọ́ 15, ó máa gbin òdòdó 2 * 15 = 30. tí 5 ò bá wù, ó ní òdòdó 30 - 5 = 25. ìdáhùn náà ni 25.

Table 7: **Qualitative Analysis** comparison of the 0-shot and 5-shot samples on ADR, Hate speech and Math.

lingual QA and math reasoning—which explains the large difference in performance scores. This analysis shows the benefit of using several prompts in evaluation, although the benefit for text generation tasks are limited. Finally, the largest variability is observed in the smaller `Gemma 2 9B` model, suggesting that smaller LLMs require more prompt template tuning than larger models, as evidenced by the lower sensitivity of GEMMA 2 27B.

6.2 Qualitative Analysis

Table 7 shows the benefit of few-shot examples on ADR, hate speech and math reasoning—the three tasks that improved the most with few-shot examples. For the ADR evaluation on Ghomálá', we observed an improvement of over 60.0 chrF point, and noticed that only a few characters had incorrect diacritics unlike the zero-shot setting. Similarly, for hate speech, without the few-shot example, the LLM focused on the abusive word "oloriburuku" (i.e. brainless), however, when we consider the *target* to tweet, it is obvious that it was referring to an entire tribe in Nigeria, which is "hate". With the definition of "hate" provided in the prompt, and examples provided, this becomes clearer to the model than without demonstration examples. Finally, for the math reasoning, in zero-shot setting, the LLM often produces *incorrect* or *incomplete* reasoning steps about the Yorùbá question which leads to an incorrect answer. However, when provided with few-shot in the language, `GPT-4o` came up with more appropriate reasoning steps, leading to the *correct* answer. This observation is particularly promising for many low-resource languages.

7 Conclusion

In this paper, we introduce a new benchmark, AFROBENCH, that aggregates existing evaluation datasets for African languages, and adds a *new* dataset focused on diacritics restoration. AFROBENCH comprises 15 NLP tasks, 22 datasets, and 64 African languages under-represented in NLP. We evaluate the performance of several closed and open LLMs on these tasks, showing that they consistently fall behind the fine-tuned baselines. We also observe a large performance gap compared to English, although the gap is smaller for closed models such as `GPT-4o` and `Gemini-1.5 pro`. Through this benchmark, we have created a leaderboard focusing on LLM evaluation for African languages, which will be maintained going forward with additional tasks, LLMs, and languages. We will be releasing our prompt and task configurations to Eleuther's *lm-eval*. We hope this encourages the development of more African-centric LLMs for African languages.[8] Our aim is to continuously add newer LLMs to the leaderboard, we demonstrate this by adding the following LLMs to the AFROBENCH-LITE: Lugha-LLaMa (an African-centric LLM) (Buzaaba et al., 2025), GPT-4.1, Gemini-2.0-Flash, and LLaMa 4 400B (Maverick), as shown in Appendix E.

[8]Our evaluation suite is available at: The-African-Research-Collective/afrobench-eval-suite.

8 Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Additional funding is provided by Microsoft via the Accelerating Foundation Models Research program. We are grateful for the support from IVADO and the Canada First Research Excellence Fund. We would also like to thank Google Cloud for the GCP credits awarded through the Gemma 2 Academic Program, and OpenAI for providing us with API credits through their Researcher Access API Program. Finally, we thank Israel Abebe for contributing the inference results of GPT-4.1 and Lugha-LLaMa to the AfroBench-Lite leaderboard.

9 Limitation

In today's NLP landscape, large language models are general-purpose models capable of performing multiple NLP tasks without the need for special training on these tasks. These models are often multilingual and can perform tasks in multiple languages. Our research examines how these models perform specifically with African languages, revealing performance disparities when compared to more resourced languages. In this section, we discuss some of the limitations of our research methodology and findings.

1. Training Data Transparency and Contamination: One of the challenges in evaluating large language models lies in the limited visibility into their training data composition. While organizations frequently publish training documentation, many reports lack comprehensive details about data mixtures and language distributions across different training stages. This lack of transparency impacts our findings in several ways. Without knowledge of the data mixture — or to what extent — our evaluation sets overlap with the training data. Thus, we cannot conclude that superior performance on certain tasks is a true demonstration of generalization or merely the model's exposure to similar content during training. In the context of African languages, knowledge of the training data enables us to examine other factors such as cross-lingual transfer, which may inform our interpretation of evaluation results. A clear understanding of training data composition serves as a crucial foundation for meaningful model evaluation. It helps establish the validity of performance metrics and provides essential context for interpreting results across different languages and tasks.

2. Limited Selection of LLMs and Evaluation Costs: We are only able to evaluate a limited set of LLMs due to the computational and financial costs associated with model access and inference. Language models are accessed using two primary methods; loading the pretrained checkpoints directly or using an API service. While providers like Together AI offer access to open-source models and companies like OpenAI provide proprietary model access, both approaches incur considerable costs that directly impact the scope of evaluation studies. In our evaluation, the costs were substantial, requiring approximately $2,500 each for `Gemini-1.5 pro` and `GPT-4o` model access, with an additional $1,200 for utilizing the Together AI platform. The total evaluation costs manifest in two key dimensions; first, when running the models locally, the GPU requirements for larger models are substantial; second, when using API services, the cost scales directly with the size of the evaluation dataset and number of models. These cost implications impose a limitation on the breadth and depth of our evaluation studies. We had to make strategic decisions about which models to include in our benchmark and how extensively to test them. This financial constraint introduces selection bias in terms of which models and tasks to prioritize, ultimately limiting the scope of our evaluation.

3. Long-tail Distribution of Languages Across Tasks & Datasets: Another limitation of AfroBench is the uneven distribution of languages across tasks and datasets. While our evaluation covers 64 languages in total, the coverage across tasks and datasets exhibits a long-tail distribution. As shown in Table F, 60% of the languages appear in fewer than 5 of the 21 datasets. This poses two challenges; first, it limits our ability to accurately assess the performance of LLMs across these underrepresented languages. Secondly, it highlights the gap in the availability of evaluation datasets even among low-resource languages. Without extensive dataset coverage for these languages, conclusions about LLM capabilities across these languages remain tentative.

4. Constraints in Machine Translation Metrics: Machine translation is often evaluated using BLEU and ROUGE, which rely on word-level recall and precision, and chrF, which operates at the character level. Research has shown these metrics sometimes demonstrate poor correlation with human judgments of translation quality. Other evaluation metrics that utilize embedding similarity, such as BERTScore (Zhang* et al., 2020) and COMET (Rei et al., 2020) / AfriCOMET (Wang et al., 2024), which leverages pretrained encoder models to generate scores by comparing translations against reference texts, are promising alternatives. However, these neural evaluation models have limited language coverage, making them unsuitable for many of the languages in our study. As a result, we rely on chrF++, which combines unigram and character n-gram overlap measurements. While this metric provides broader language coverage, it is a compromise between evaluation quality and practical applicability.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, et al. 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, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, et al. 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 Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, et al. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, et al. 2023. Masakhanews: News topic classification for african languages. *Preprint*, arXiv:2304.09972.

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

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Meta AI. 2024. Meta ai announces llama 3.1. Accessed: Feb 1, 2025.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

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. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Anuoluwapo Aremu, Jesujoba Oluwadara Alabi, Daud Abolade, Nkechinyere Faith Aguobi, Shamsuddeen Hassan Muhammad, and David Ifeoluwa Adelani. 2024. NaijaRC: A multi-choice reading comprehension dataset for nigerian languages. In *5th Workshop on African Natural Language Processing*.

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

Edward Bayes, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *ArXiv*, abs/2412.00948.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa 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. *arXiv preprint arXiv:2405.14782*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lughallama: Adapting large language models for african languages. *ArXiv*, abs/2504.06536.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

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. 2023. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

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.

Clémentine Fourrier, Nathan Habib, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

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

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Jacaranda Health, Mwongela Stanslaus, Patel Jay, Lusiji Sathy Rajasekharan, Lyvia, Piccino Francesco, Ukwak Mfoniso, and Sebastian Ellen. 2024. Afrollama v1: An instruction-tuned llama model for african languages. Accessed: Feb 12, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

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.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian

Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, and et al. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, Lilian D. A. Wanzare, Samuel Rutunda, Lukman Jibril Aliyu, Esubalew Alemneh, Oumaima Hourrane, Hagos Tesfahun Gebremichael, Elyas Abdi Ismail, Meriem Beloucif, Ebrahim Chekol Jibril, Andiswa Bukula, Rooweither Mabuya, Salomey Osei, Abigail Oppong, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Chiamaka Ijeoma Chukwuneke, Paul Röttger, Seid Muhie Yimam, and Nedjma Ousidhoum. 2025. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. *Preprint*, arXiv:2501.08284.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

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

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *Preprint*, arXiv:2305.06897.

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

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2024. Gpt-4o system card. Accessed: February 13, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, and et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

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.

Sebastian Ruder. 2021. Challenges and Opportunities in NLP Benchmarking. `http://ruder.io/nlp-benchmarking`.

Sebastian Ruder, J. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Ifeoluwa Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, R. Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Pratim Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *ArXiv*, abs/2305.11938.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint*.

Iyanuoluwa Shode, David Ifeoluwa Adelani, JIng Peng, and Anna Feldman. 2023. NollySenti: Leveraging transfer learning and machine translation for Nigerian movie sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998, Toronto, Canada. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Hilal Ermiş, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *ArXiv*, abs/2412.03304.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 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, Madian 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, Rashi 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. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Task Based Results

We group tasks using similar evaluation metrics to analyze model performance systematically.

B LLMs evaluated

Models are selected to cover a range of open- and closed-source LLMs with diverse parameter sizes, multilingual capabilities, and recent advancements. We prioritize models with strong multilingual support, accessibility for research, and relevance to African languages.

B.0.1 Open Models

These are LLMs whose architectures, weights, and often training datasets are publicly available, allowing researchers and practitioners to fine-tune or adapt them to specific use cases. These models promote transparency, replicability, and accessibility, particularly for low-resource language tasks.

Aya-101 (Üstün et al., 2024) is a T5-style encoder-decoder model fine-tuned for low-resource multilingual applications, including African languages. It was fine-tuned on a curated dataset, consisting of public multilingual corpora, and machine & human translated datasets from more than 100 languages. The model adopts a text-to-text paradigm and emphasizes cross-lingual transfer learning, allowing for robust generalization across various multilingual text-based tasks.

LLaMa 2 7B Chat LLaMa 2 (Touvron et al., 2023) is a collection of open-source pretrained and fine-tuned generative text models developed by Meta, ranging from 7 billion to 70 billion parameters. The 7B Chat variant allows for dialogue use cases. It employs an auto-regressive transformer architecture and has been fine-tuned using supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF). These models are pretrained on multiple languages, but have limited coverage of African languages.

LLaMa 3 8B Instruct Llama 3 (Dubey et al., 2024) is an updated variant of the LlaMa 2 series (Touvron et al., 2023). The models are instruction-fine-tuned to handle a wide range of text-based tasks. Similar to LLaMa 2, it supports multiple languages, though coverage of African languages remains limited. The number of parameters ranges from 8B to 70B; we use the 8B variant for this evaluation.

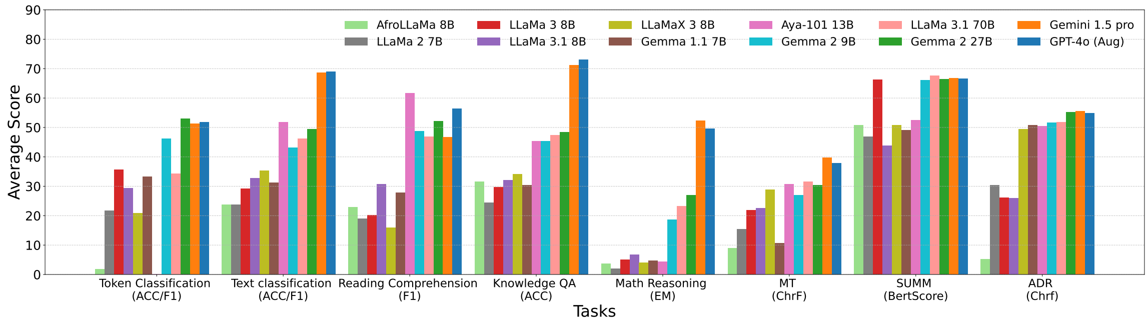

Figure 5: Performance of models across various NLP tasks, grouped by metric-based evaluation categories. Tasks include Token Classification, Text classification, Reading Comprehension QA, Knowledge QA, Math Reasoning, Machine Translation (MT), and Summarization (SUMM) and Diacritics Restoration (ADR).

LLaMa 3.1 Instruct (8B, 70B) LLaMa 3.1 (AI, 2024) is an updated variant of the LLaMa 3 series. Compared to LLaMa 3 (Dubey et al., 2024), LLaMa 3.1 (AI, 2024) introduces improvements in multilingual capabilities and general instruction-following. We use the instruction-tuned variants, fine-tuned for a broad range of NLP tasks. While these models support multiple languages, coverage of African languages remains limited. LLaMA 3.1 is available in parameter sizes ranging from 8B to 405B; due to computational cost, we evaluate only the 8B and 70B variants.

Gemma 1.1 7B IT (Mesnard et al., 2024) is a lightweight open model from Google, built from the same research and technology used to create the Gemini models. It is a text-to-text, decoder-only large language models, available in English, with open weights, pre-trained variants, and instruction-tuned variants. However, it does not have strong multilingual support. We evaluate the instruction-tuned 7B variant of this model.

Gemma 2 IT (9B, 27B) (Riviere et al., 2024) is an improved iteration of the Gemma model series optimized for efficiency. Compared to Gemma 1, Gemma 2 incorporates enhanced instruction-following capabilities and more robust parameter scaling. We evaluate the instruction-tuned variants of Gemma 2 at 9B and 27B parameter sizes.

AfroLlama-V1 (Health et al., 2024) is a decoder-only transformer model, optimized for African language applications. It leverages proprietary datasets, including text from social media, newspapers, and government publications in African languages. Its architecture is based on LLaMa 3 8B (Dubey et al., 2024), but includes additional pretraining on African-centric text.

B.0.2 Proprietary Models

These are proprietary systems developed and maintained by organizations. Their training data and architectures are typically undisclosed.

GPT-4o (Aug) GPT-4o (OpenAI, 2024) is an optimized version of OpenAI's GPT-4 model (OpenAI, 2023). It is an autoregressive omni model, trained end-to-end across text, vision, and audio on both public and proprietary data. While specific details about its architecture and datasets are not publicly disclosed, the GPT series is designed to adapt effectively to various language tasks, making it suitable for applications involving African languages. We evaluated the August 2024 version of this model.

Gemini 1.5 Pro 002. Gemini (Reid et al., 2024) is a cutting-edge proprietary model with strong multilingual capacity. It's a compute-efficient multimodal model whose training data is tailored for diverse linguistic contexts, including low-resource languages. Although its training data and architecture remain undisclosed, Gemini is designed to adapt effectively to various language tasks, making it suitable for applications involving African languages.

C Evaluation Tools and Framework

AfroBench and AfroBench-Lite are fully integrated with Eleuther LM Evaluation Harness (Gao et al., 2024) for open models, with sample run scripts and instructions on how to run the benchmark. We chose Eleuther LM Evaluation Harness due to its open-source and reproducible nature and widespread adoption within the industry. The evaluation methodology varies by task type: text classification and multiple-choice tasks are assessed using log-likelihood evaluation, which measures the probability of a prompt-generated continuation containing the expected response, while all other tasks utilize free-form generation approaches.

For proprietary models accessed through their API, we developed a custom framework to prompt and evaluate these models. This framework is also open-sourced with sample run scripts and instructions on how to reproduce the benchmarks. The same prompt and evaluation methodology for task is used in both the LM Evaluation Harness and our custom API framework.

D AfroBench Evaluation with Confidence Scores

We computed 95% confidence intervals for AfroBench results to quantify statistical significance. The calculation was based on the results of 5 prompts for each task (3 prompts for NLG tasks). Table 9 presents the average performance and confidence intervals across prompts to assess variability and significance.

E Newer LLM evaluation on AfroBench-Lite

We extend our evaluation for AFROBENCH-LITEto include newer LLMs such as Lugha-LLaMa (an African-centric LLM) (Buzaaba et al., 2025), GPT-4.1, Gemini-2.0-Flash, and LLaMa 4 400B (Maverick) in Table 8.

F Languages covered in the evaluation

Table 10 lists the languages and corresponding tasks evaluated.

Model	Lang	Intent	TC	NLI	RC	MMLU	Math	MT *en/fr-xx*	AVG
Lugha-Llama 8B	*eng*	16.7	43.6	46.8	22.4	31.8	6.4	51.3	31.3
	africa	4.1	34.1	36.7	23.0	25.2	1.8	22.1	21.0
Gemma 1.1 7B	*eng*	72.1	86.3	59.2	87.9	44.6	20.8	26.1	56.7
	africa	10.2	42.0	34.6	34.1	27.3	5.1	10.9	23.5
Gemma 2 9B	*eng*	36.3	82.5	70.7	93.7	69.8	68.8	67.9	70.0
	africa	27.8	64.0	40.9	49.3	36.1	21.7	37.2	39.6
LLaMa 3.1 70B	*eng*	84.5	88.3	59.5	93.2	76.4	86.8	71.6	80.0
	africa	36.9	61.9	38.4	45.3	40.6	26.5	29.6	39.9
Aya-101 13B	*eng*	78.0	82.8	67.0	86.1	42.8	11.6	64.2	61.8
	africa	40.2	76.0	52.4	59.7	30.3	4.9	31.8	42.2
Gemma 2 27B	*eng*	84.0	89.3	67.8	93.4	75.6	85.6	68.5	80.6
	africa	31.4	66.6	43.7	52.1	40.8	30.6	39.1	43.5
LLaMa 4 405B	*eng*	88.9	84.8	49.2	25	11.2	97.6	73	61.4
	africa	73.9	80.6	45.5	24.6	15.8	65.0	42.8	49.7
Gemma 3 27B	*eng*	79.6	87.3	65.5	93.4	74.2	87.6	68.9	79.5
	africa	55.2	74.2	51.2	62.4	44.4	47.5	33.1	52.6
Gemini 1.5 pro	*eng*	86.8	88.7	88.5	69.6	88.8	86.8	69.1	82.6
	africa	75.6	81.3	63.6	54.4	62.6	57.7	44.2	62.8
GPT-4o (Aug)	*eng*	86.2	89.2	89.2	84.3	88.0	88.8	70.2	85.1
	africa	78.4	83.0	66.3	**70.3**	**63.1**	57.3	43.6	66.0
Gemini 2.0 Flash	*eng*	87.6	86.8	87	63	80.8	92.8	73.1	79.7
	africa	82.5	**84.9**	66.5	56.8	57.8	**67.5**	**49.6**	66.5
GPT-4.1 (April)	*eng*	87.8	89.7	88.5	73.9	71.4	82.4	73.1	81.0
	africa	**84.4**	84.8	**67.5**	64.8	60.2	59.9	47.3	**67.0**

Table 8: **AfroBench-Lite Evaluation (NEW)**: LLM baselines on 7 datasets spanning 14 African languages (sorted by performance on African languages). Tasks were selected for broad NLP coverage, prioritizing language consistency. The best score per task is in **bold**.

Task	LLaMa2 7B	LLaMa3 8B	LLaMaX 8B	LLaMa3.1 8B	AfroLLaMa 8B	Gemma2 9B	Aya-101 13B	Gemma2 27B	LLaMa3.1 70B	Gemini1.5 Pro	GPT-4o (Aug)
POS	22.6±13.6	45.8±4.4	38.7±4.4	42.9±6.5	0.0±0.0	47.9±7.9	0.0±0.0	53.6±3.1	52.0±4.7	59.5±3.0	60.1±5.9
NER	11.1±10.7	17.3±8.3	0.0±0.0	7.7±5.6	2.9±2.2	25.9±30.8	0.0±0.0	43.1±11.4	12.9±5.8	40.6±3.6	37.1±6.8
SA	37.5±17.0	39.7±16.3	44.5±17.1	45.7±18.4	39.8±25.2	48.3±29.0	60.0±9.8	58.4±17.3	43.4±18.3	65.4±15.2	64.6±17.7
TC	15.3±14.5	24.6±26.9	23.5±32.0	37.5±26.4	16.9±22.1	51.6±15.9	68.9±4.4	59.4±8.7	47.0±17.5	73.5±10.2	73.3±4.9
Intent	0.8±1.5	0.9±2.3	3.1±3.8	4.0±5.0	0.3±1.0	29.2±5.6	42.4±4.6	33.0±4.9	31.8±7.4	68.4±12.2	70.4±6.6
Hate	16.8±10.8	21.8±11.0	23.0±12.5	19.3±5.9	15.2±8.1	21.3±13.0	28.7±—	36.6±15.0	36.5±29.3	49.7±33.5	49.5±37.6
NLI	33.4±1.5	33.7±2.7	35.0±6.8	34.3±3.8	34.2±4.4	36.3±6.6	48.3±5.3	37.3±7.3	35.2±5.4	56.1±15.9	58.4±11.4
XQA	10.4±5.7	9.6±5.6	2.0±0.5	14.1±14.3	19.2±5.2	39.3±13.8	61.9±1.6	47.7±7.6	37.1±8.7	34.8±11.8	31.6±25.0
RC	24.3±3.5	28.0±8.2	24.6±5.7	36.2±16.2	24.4±2.5	47.7±26.2	55.2±29.4	47.6±28.8	44.5±16.8	52.7±7.6	71.4±3.2
Arc-E	21.0±4.3	30.8±3.8	39.3±2.6	31.7±3.0	35.8±2.8	52.9±1.8	59.3±1.4	55.5±1.9	55.4±4.3	83.8±2.1	85.2±1.4
MMLU	24.5±2.4	26.7±2.2	28.0±1.4	30.3±4.5	25.1±1.9	34.8±8.8	30.4±4.0	38.9±9.6	37.9±8.6	50.7±12.2	55.3±15.5
Math	1.8±1.3	4.2±3.2	3.7±2.5	5.5±3.4	0.1±0.4	14.1±8.0	4.3±1.6	25.4±4.8	20.3±5.4	46.6±20.3	48.7±4.2
MT (en-xx)	7.9±7.1	15.0±4.7	21.9±4.6	16.1±2.5	7.4±3.2	24.5±1.2	23.0±2.9	27.5±2.8	24.7±5.2	37.6±1.9	34.4±2.9
MT (xx-en)	17.8±7.5	23.1±11.0	34.0±5.0	27.7±3.8	8.3±3.0	28.8±0.8	36.9±4.1	32.7±1.5	35.8±8.8	41.7±0.8	40.5±1.5
ADR	22.8±19.2	24.1±7.4	47.2±6.6	23.1±6.8	4.3±2.4	50.3±4.4	49.8±1.8	53.5±4.4	48.2±16.2	54.5±4.2	52.9±5.0

Table 9: Model performance based on average with standard deviation at 95% confidence intervals

	Language	Branch	Region (of Africa)	Script	# speakers
Afro-Asiatic	Algerian Arabic (arq)	Semitic	North	Arabic	36M
	Amharic (amh)	Ethio-Semitic	East	Ge’ez	57M
	Egyptian Arabic (arz)	Semitic	North	Arabic	41M
	Hausa (hau)	Chadic	West	Latin	77M
	Kabyle (kab)	Berber	North	Arabic	3M
	Oromo (orm)	Cushitic	East	Latin	37M
	Moroccan Arabic (ary)	Semitic	North	Arabic	29M
	Somali (som)	Cushitic	East	Latin	22M
	Tamasheq (taq)	Berber	East	Latin	1M
	Tamazight (tzm)	Berber	East	Latin	-
	Tigrinya (tig)	Ethio-Semitic	East	Ge’ez	9M
	Tunisian Arabic (aeb)	Semitic	North	Arabic	12M
Niger-Congo	Akan (aka)	Tano	West	Latin	10M
	Bambara (bam)	Mande	West	Latin	14M
	Bemba (bem)	Bantu	South/East/Central	Latin	4M
	Chichewa (nya)	Bantu	South-East	Latin	14M
	chiShona (sna)	Bantu	Southern	Latin	11M
	Chokwe (cjk)	Bantu	South/Central	Latin	1M
	Dyula (dyu)	Mande	West	Latin	3M
	Éwé (ewe)	Kwa	West	Latin	7M
	Fon (fon)	Volta-Niger	West	Latin	14M
	Ghomálá’ (bbj)	Grassfields	Central	Latin	1M
	Igbo (ibo)	Volta-Niger	West	Latin	31M
	isiXhosa (xho)	Bantu	Southern	Latin	19M
	isiZulu (zul)	Bantu	Southern	Latin	27M
	Kabiyè (kbp)	Gur	West	Latin	1M
	Kamba (kam)	Bantu	East	Latin	5M
Niger-Congo	Kikongo (kon)	Bantu	South/Central	Latin	5M
	Kikuyu (kik)	Bantu	East	Latin	8M
	Kimbundu (kmb)	Bantu	Southern	Latin	2M
	Kinyarwanda (kin)	Bantu	East	Latin	10M
	Kiswahili (swa)	Bantu	East/Central	Latin	71-106M
	Lingala (lin)	Bantu	Central	Latin	40M
	Luba-Kasai (lua)	Bantu	Central	Latin	6M
	Luganda (lug)	Bantu	Central	Latin	11M
	Lugbara (lgg)				
	Mossi (mos)	Gur	West	Latin	8M
	Nigerian Fulfulde (fuv)	Senegambia	West	Latin	15M
	N’Ko (nqo)	Mande	West	Latin	-
	Northern Sotho (nso)	Bantu	Southern	Latin	4M
	Rundi (run)	Bantu	East	Latin	11M
	Runyankole (nyn)				
	Sango (sag)	Ubangian	Central	Latin	5M
	Setswana (tsn)	Bantu	Southern	Latin	14M
	Southern Sotho (sot)	Bantu	Southern	Latin	7M
	Swati (ssw)	Bantu	Southern	Latin	1M
	Twi (twi)	Kwa	West	Latin	9M
	Tumbuka (tum)	Bantu	South/East	Latin	2M

Continued on next page

	Language	Branch	Region (of Africa)	Script	# speakers
	Umbundu (umb)	Bantu	Southern	Latin	7M
	Xitsonga (tso)	Bantu	Southern	Latin	7M
	Wolof (wol)	Senegambia	West	Latin	5M
	Yoruba (yor)	Volta-Niger	West	Latin	46M
Nilo-Saharan	Acholi (ach)	Nilotic	East	Latin	1.5M
	Ateso (teo)	Nilotic	East	Latin	2.8M
	Dinka (dik)	Nilotic	Central	Latin	4M
	Kanuri (knc)	Saharan	West/Central	Latin	10M
	Kanuri (knc)	Saharan	West/Central	Arabic	10M
	Luo (luo)	Nilotic	East	Latin	4M
	Neur (nus)	Nilotic	Central	Latin	2M
Austronesian	Malagasy (plt)	Malayo-Polynesian	Southern	Latin	25M
Indo-European	Afrikaans (afr)	Germanic	Southern	Latin	7M
	Mozambican Portuguese (pt-MZ)	Italic	South East	Latin	13M
Creoles	Nigerian Pidgin (pcm)	English-based	West	Latin	121M
	Kabuverdianu (kea)	Portuguese-based	West	Latin	1M

Table 10: **Languages covered in each of our evaluation tasks**: language family, region, script, number of L1/L2 speakers

	Classification									Reasoning	Question Answering						Generation						
Lang.	AFRIHATE	AFRISENTI	AFRIXNLI	INJONGO-INTENT	NOLLYSENTI	MASAKHANEWS	MASAKHANER	MASAKHAPOS	SIB-200	AFRIMGSM	AFRIMMLU	AFRIQA	BELEBELE	NAIJARC	OPENAI-MMLU	UHURA	AFRIADR	FLORES	MAFAND	NTREX-128	SALT	XL-SUM	**# Tasks**
aeb									✓									✓					2
ach																					✓		1
afr													✓					✓		✓			3
aka									✓									✓					2
amh	✓	✓	✓	✓		✓	✓		✓	✓	✓		✓			✓		✓	✓	✓			14
ara															✓							✓	2
arq	✓	✓																					2
ary	✓	✓							✓				✓					✓					5
arz									✓				✓					✓					3
bam							✓	✓	✓				✓					✓	✓				6
bbj							✓	✓									✓		✓				4
bem									✓			✓						✓		✓			4
cjk									✓									✓					2
dik									✓									✓					2
dyu									✓									✓					2
ewe			✓	✓			✓	✓	✓	✓	✓							✓	✓	✓			10
fon								✓	✓			✓					✓	✓	✓				6
fuv									✓														1
gaz									✓									✓					2
hau	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	19
ibo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	19
kab									✓									✓					2
kam									✓									✓					2
kbp									✓									✓					2
kea									✓									✓					2
kik									✓									✓					2
kin	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓						✓	✓	✓			13
kmb									✓									✓					2
knc									✓									✓					2
kon									✓									✓					2
lgg																					✓		1
lin			✓	✓		✓			✓	✓	✓		✓					✓					8
lua									✓									✓					2
lug			✓	✓		✓	✓	✓	✓	✓	✓							✓	✓		✓		11
luo							✓	✓	✓									✓	✓				5
mos							✓	✓	✓									✓	✓				5
nde																				✓			1
nso									✓									✓		✓			3
nus									✓									✓					2
nya							✓	✓	✓									✓	✓	✓			6

Continued on next page

Lang.	Classification									Reasoning	Question Answering						Generation						# Tasks
	AfriHate	AfriSenti	AfriXNLI	Injongo-Intent	NollySenti	MasakhaNEWS	MasakhaNER	MasakhaPOS	SIB-200	AfriMGSM	AfriMMLU	AfriQA	Belebele	NaijaRC	OpenAI-MMLU	Uhura	AfriADR	Flores	Mafand	NTREX-128	SALT	XL-SUM	
nyn																					✓		1
orm	✓	✓	✓	✓		✓				✓	✓									✓		✓	9
pcm	✓	✓				✓	✓	✓											✓			✓	7
plt									✓									✓		✓			3
run						✓			✓									✓					3
sag									✓									✓					2
sna			✓	✓		✓	✓	✓	✓	✓	✓		✓					✓	✓	✓			12
som	✓					✓			✓									✓		✓		✓	6
sot			✓	✓						✓	✓							✓					5
ssw									✓									✓		✓			3
swa	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓		✓	✓		✓	✓	✓	✓	✓	18
taq																		✓					1
teo																					✓		1
tir	✓	✓				✓			✓				✓					✓		✓		✓	8
tsn							✓	✓										✓	✓	✓			5
tso		✓							✓				✓										3
tum									✓									✓					2
twi		✓	✓	✓			✓	✓	✓	✓	✓	✓						✓	✓				11
tzm									✓									✓					2
umb									✓									✓					2
ven																				✓			1
wol			✓	✓			✓	✓	✓		✓		✓				✓	✓	✓	✓	✓		12
xho	✓		✓	✓		✓	✓		✓	✓	✓		✓					✓	✓	✓			13
yor	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	21
zul	✓		✓	✓			✓	✓	✓	✓	✓	✓	✓			✓		✓	✓	✓			14

Table 11: **Languages covered in each of our evaluation tasks**: check marks (✓) indicate that a language is covered by the task in that column. While 13 languages are covered by $\geq$ 10 tasks, 44 languages are covered by $\leq$ 5 tasks. SIB-200 and Flores have the broadest coverage of African languages. In general, classification and generation tasks have better coverage of African languages than reasoning and question answering tasks.

G Best Performing Prompt

The table below details which prompt performed best for each model on each dataset. The actual prompts can be found in Appendix H.

Task	Dataset	AfroLLaMa 8B	LLaMAX3 8B	LLaMa2 7b	LLaMa3 8B	LLaMa3.1 8B	LLaMa3.1 70B	Aya-101 13B	Gemma1.1 7b	Gemma2 9b	Gemma2 27b	Gemini 1.5 Pro	GPT-4o (Aug)
SA	AfriSenti	T4	T3	T5	T5	T5	T5	T3	T5	T5	T5	T3	T2
	NollySenti	T5	T3	T4	T3	T4	T4	T5	T4	T4	T4	T1	T3
TC	Masakhanews	T3	T3	T4	T3	T3	T3	T2	T2	T3	T3	T2	T2
	SIB	T3	T3	T5	T2	T2	T3	T4	T4	T5	T3	T5	T3
TokC	MasakhaNER	T4	T1	T5	T3	T3	T5	T1	T5	T2	T1	T3	T3
	MasakhaPOS	T1	T5	T1	T2	T2	T2	T1	T2	T2	T3	T3	T3
Intent	InjongoIntent	T1	T5	T4	T5	T3	T4	T5	T5	T5	T5	T5	T4
Hate	AfriHATE	T5	T4	T3	T1	T4	T4	T1	T4	T1	T4	T1	T4
NLI	AfriXNLI	T2	T1	T3	T2	T2	T2	T4	T2	T2	T2	T3	T3
XQA	AfriQA	T5	T4	T5	T5	T5	T2	T2	T2	T2	T2	T2	T2
RC	NaijaRC	T4	T5	T4	T1	T5	T5	T4	T3	T4	T5	T3	T2
	Belebele	T2	T5	T4	T1	T5	T5	T5	T3	T5	T5	T1	T1
Arc-E	Uhura-Arc Easy	T1	T4	T1	T5	T3	T2	T2	T1	T4	T4	T1	T3
MMLU	AfriMMLU	T5	T5	T1	T4	T3	T1	T2	T1	T1	T1	T1	T1
	Openai-MMLU	T5	T4	T3	T5	T5	T5	T5	T5	T3	T5	T1	T1
Math	AfriMGSM	T1	T4	T3	T4	T4	T4	T1	T1	T2	T4	T5	T2
MT	Flores en_xx	T3	T2	T1	T1	T2	T2	T1	T2	T2	T2	T2	T3
	Flores xx_en	T1	T3	T3	T1	T3	T2	T3	T1	T2	T2	T1	T2
	Mafand en_xx	T1	T2	T2	T2	T2	T2	T1	T1	T2	T2	T3	T1
	Mafand xx_en	T3	T2	T2	T2	T2	T2	T2	T1	T2	T1	T3	T1
	NTREX en_xx	T3	T2	T1	T1	T2	T2	T2	T1	T1	T3	T2	T2
	NTREX xx_en	T1	T2	T3	T1	T2	T2	T2	T1	T2	T2	T2	T2
	Salt en_xx	T2	T1	T1	T2	T1	T2	T3	T1	T3	T2	T2	T3
	Salt xx_en	T1	T2	T3	T1	T3	T3	T3	T1	T2	T2	T2	T2
Summ	XLSUM	T3	T3	T3	T3	T3	T3	T1	T3	T1	T1	T3	T1
ADR	ADR	T4	T4	T2	T4	T4	T3	T1	T3	T4	T3	T2	T1

Table 12: Best-performing prompts per model for each dataset. These prompts achieved the highest scores reported in the paper

H Prompt Bank

In this section, we list all prompts used in our experiments. We use zero-shot cross-lingual prompts, where the context and query are in English, while the input text is in the target African language. This approach leverages LLMs' stronger ability to follow instructions in English (Lin et al., 2021; Shi et al., 2022). The prompts are grouped by task category, as shown in Figure 2.

H.1 Natural Language Understanding

POS prompts:

Listing 1: MasakhaPOS Prompt 1

```
Please provide the POS tags for each word in the
     input sentence. The input will be a list of
     words in the sentence. The output format should
      be a list of tuples, where each tuple consists
      of a word from the input text and its
     corresponding POS tag label from the tag label
     set: ['ADJ', 'ADP', 'ADV', 'AUX', 'CCONJ, 'DET
     ', 'INTJ', 'NOUN', 'NUM', 'PART', 'PRON', '
     PROPN', 'PUNCT', 'SCONJ', 'SYM', 'VERB', 'X'].
Your response should include only a list of tuples,
     in the order that the words appear in the input
      sentence, including punctuations, with each
     tuple containing the corresponding POS tag
     label for a word.

Sentence: {{text}}
Output:
```

Listing 2: MasakhaPOS Prompt 2

```
You are an expert in tagging words and sentences in
     {{language}} with the right POS tag.

Please provide the POS tags for each word in the {{
     language}} sentence. The input is a list of
     words in the sentence. POS tag label set: [ '
     ADJ', 'ADP', 'ADV', 'AUX', 'CCONJ', 'DET', 'INTJ
     ', 'NOUN', 'NUM', 'PART', 'PRON', 'PROPN', '
     PUNCT', 'SCONJ', 'SYM', 'VERB', 'X' ]. The
     output format should be a list of tuples, where
      each tuple consists of a word from the input
     text and its corresponding POS tag label from
     the POS tag label set provided.
Your response should include only a list of tuples,
     in the order that the words appear in the input
      sentence, including punctuations, with each
     tuple containing the corresponding POS tag
     label for a word.

Sentence: {{text}}
Output:
```

Listing 3: MasakhaPOS Prompt 3

```
Acting as a {{language}} linguist and without making
      any corrections or changes to the text,
     perform a part of speech (POS) analysis of the
     sentences using the following POS tag label
     annotation ['ADJ', 'ADP', 'ADV', 'AUX', 'CCONJ',
      'DET', 'INTJ', 'NOUN', 'NUM', 'PART', 'PRON',
     'PROPN', 'PUNCT', 'SCONJ', 'SYM', 'VERB', 'X'].
      The input will be a list of words in the
     sentence. The output format should be a list of
      tuples, where each tuple consists of a word
     from the input text and its corresponding POS
     tag label from the POS tag label set provided.
Your response should include only a list of tuples,
     in the order that the words appear in the input
      sentence, including punctuations, with each
     tuple containing the corresponding POS tag
     label for a word.
```

```
Sentence: {{text}}
Output:
```

Listing 4: MasakhaPOS Prompt 4

```
Annotate each word in the provided sentence with the
      appropriate POS tag. The annotation list is
     given as: ['ADJ', 'ADP', 'ADV', 'AUX', 'CCONJ,
     'DET', 'INTJ', 'NOUN', 'NUM', 'PART', 'PRON', '
     PROPN', 'PUNCT', 'SCONJ', 'SYM', 'VERB', 'X'].
     The input sentence will be a list of words in
     the sentence. The output format should be a
     list of tuples, where each tuple consists of a
     word from the input text and its corresponding
     POS tag label from the POS tag label set
     provided\nYour response should include only a
     list of tuples, in the order that the words
     appear in the input sentence, including
     punctuations, with each tuple containing the
     corresponding POS tag label for a word.

Sentence: {{text}}
Output:
```

Listing 5: MasakhaPOS Prompt 5

```
Given the following sentence, identify the part of
     speech (POS) for each word. Use the following
     POS tag set:
NOUN: Noun (person, place, thing),
VERB: Verb (action, state),
ADJ: Adjective (describes a noun),
ADV: Adverb (modifies a verb, adjective, or adverb),
PRON: Pronoun (replaces a noun),
DET: Determiner (introduces a noun),
ADP: Adposition (preposition or postposition),
CCONJ: Conjunction (connects words, phrases, clauses
     )
PUNCT: Punctuation,
PROPN: Proper Noun,
AUX: Auxiliary verb (helper verb), \nSCONJ:
     Subordinating conjunction
PART: Particle,
SYM: Symbol,
INTJ: Interjection,
NUM: Numeral,
X: others. The output format should be a list of
     tuples, where each tuple consists of a word
     from the input text and its corresponding POS
     tag label key only from the POS tag set
     provided
Your response should include only a list of tuples,
     in the order that the words appear in the input
      sentence, including punctuations, with each
     tuple containing the corresponding POS tag
     label for a word.

Sentence: {{text}}
Output:
```

NER prompts:

Listing 1: MasakhaNER Prompt 1

```
Named entities refers to names of location,
     organisation and personal name.
For example, 'David is an employee of Amazon and he
     is visiting New York next week to see Esther'
     will be
PERSON: David $ ORGANIZATION: Amazon $ LOCATION: New
      York $ PERSON: Esther

Ensure the output strictly follows the format: label
     : entity $ label: entity, with each unique
     entity on a separate label line, avoiding
     grouped entities (e.g., avoid LOC: entity,
     entity) or irrelevant entries like none.

Text: {{text}}
Return only the output
```

Listing 2: MasakhaNER Prompt 2

```
You are working as a named entity recognition expert
     and your task is to label a given text with
     named entity labels. Your task is to identify
     and label any named entities present in the
     text. The named entity labels that you will be
     using are PER (person), LOC (location), ORG (
     organization) and DATE (date). Label multi-word
     entities as a single named entity. For words
     which are not part of any named entity, do not
     return any value for it.
Ensure the output strictly follows the format: label
     : entity $$ label: entity, with each unique
     entity on a separate label line, avoiding
     grouped entities (e.g., avoid LOC: entity,
     entity) or irrelevant entries like none. Return
     only the output

Text: {{text}}
```

Listing 3: MasakhaNER Prompt 3

```
You are a Named Entity Recognition expert in {{
     language}} language.
Extract all named entities from the following {{
     language}} text and categorize them into PERSON
     , LOCATION, ORGANIZATION, or DATE.
Ensure the output strictly follows the format; label
     : entity $$ label: entity, with each unique
     entity on a separate label line, avoiding
     grouped entities (e.g., avoid LOC: entity,
     entity) or irrelevant entries like none. Return
     only the output

Text: {{text}}
Return only the output
```

Listing 4: MasakhaNER Prompt 4

```
As a {{language}} linguist, label all named entities
     in the {{language}} text below with the
     categories: PERSON, LOCATION, ORGANIZATION, and
     DATE. Ensure the output strictly follows the
     format; label: entity $$ label: entity, with
     each unique entity on a separate label line,
     avoiding grouped entities (e.g., avoid LOC:
     entity, entity) or irrelevant entries like none
     . Return only the output.

Text: {{text}}
Return only the output
```

Listing 5: MasakhaNER Prompt 5

```
Provide a concise list of named entities in the text
     below. Use the following labels: PERSON,
     LOCATION, ORGANIZATION, and DATE. Ensure the
     output strictly follows the format; label:
     entity $$ label: entity, with each unique
     entity on a separate label line, avoiding
     grouped entities (e.g., avoid LOC: entity,
     entity) or irrelevant entries like none. Return
     only the output.

Text: {{text}}
Return only the output
```

Sentiment prompts:

Listing 1: AfriSenti Prompt 1

```
Does this statement; "{{tweet}}" have a Neutral,
     Positive or Negative sentiment? Labels only
```

Listing 2: AfriSenti Prompt 2

```
Does this {{language}} statement; "{{tweet}}" have a
     Neutral, Positive or Negative sentiment?
     Labels only
```

Listing 3: AfriSenti Prompt 3

```
You are an assistant able to detect sentiments in
     tweets.

Given the sentiment labels Neutral, Positive or
     Negative; what is the sentiment of the {{
     language}} statement below? Return only the
     labels.

text: {{tweet}}
label:
```

Listing 4: AfriSenti Prompt 4

```
Label the following text as Neutral, Positive, or
     Negative. Provide only the label as your
     response.

text: {{tweet}}
label:
```

Listing 5: AfriSenti Prompt 5

```
You are tasked with performing sentiment
     classification on the following {{language}}
     text. For each input, classify the sentiment as
     positive, negative, or neutral. Use the
     following guidelines:

Positive: The text expresses happiness, satisfaction
     , or optimism.
Negative: The text conveys disappointment,
     dissatisfaction, or pessimism.
Neutral: The text is factual, objective, or without
     strong emotional undertones.
If the text contains both positive and negative
     sentiments, choose the dominant sentiment. For
     ambiguous or unclear sentiments, select the
     label that best reflects the overall tone.
     Please provide a single classification for each
     input.

text: {{tweet}}
label:
```

Listing 6: NollySenti Prompt 1

```
Does this movie description "{{review}}" have a
     Positive or Negative sentiment? Labels only
```

Listing 7: NollySenti Prompt 2

```
Does this {{language} movie description; "{{review
     }}" have a Positive or Negative sentiment?
     Labels only
```

Listing 8: NollySenti Prompt 3

```
You are an assistant able to detect sentiment in
     movie reviews.

Given the sentiment labels Positive or Negative;
     what is the sentiment of the English statement
     below? Return only the labels

Review: {{review}}"
```

Listing 9: NollySenti Prompt 4

```
Label the following text as Positive, or Negative.
     Provide only the label as your response.

text: {{review}}
label:
```

Listing 10: NollySenti Prompt 5

```
You are tasked with performing sentiment
     classification on the following English text.
     For each input, classify the sentiment as
     positive, negative. Use the following
     guidelines:
```

```
Positive: The text expresses happiness, satisfaction
     , or optimism.
Negative: The text conveys disappointment,
     dissatisfaction, or pessimism.
If the text contains both positive and negative
     sentiments, choose the dominant sentiment. For
     ambiguous or unclear sentiments, select the
     label that best reflects the overall tone.
     Please provide a single classification for each
      input.

text: {{review}}
label:
```

Topic Classification prompts:

Listing 1: SIB Prompt 1

```
Given the categories science/technology, travel,
     politics, sports, health, entertainment, or
     geography; what category does the text: '{{text
     }}' belong to:
```

Listing 2: SIB Prompt 2

```
Does this {{language}} topic; '{{text}}' belong to
     one of the following categories: science/
     technology, travel, politics, sports, health,
     entertainment, or geography? category only
```

Listing 3: SIB Prompt 3

```
You are an assistant able to classify topics in
     texts.

Given the categories science/technology, travel,
     politics, sports, health, entertainment, or
     geography; what is the topic of the {{language
     }} statement below? Return only the category.

text: {{text}}
category: "
```

Listing 4: SIB Prompt 4

```
Label the following text as science/technology,
     travel, politics, sports, health, entertainment
     , or geography. Provide only the category as
     your response.

text: {{text}}
category:
```

Listing 5: SIB Prompt 5

```
You are tasked with performing topic classification
     on the following {{language}} text. For each
     input, classify the topic as science/technology
     , travel, politics, sports, health,
     entertainment, or geography. Use the following
     guidelines:

science/technology: The text discusses scientific
     discoveries, technological advancements, or
     related topics.
travel: The text describes travel experiences,
     destinations, or related topics.
politics: The text covers political events, policies
     , or related topics.
sports: The text talks about sports events, athletes
     , or related topics.
health: The text addresses health issues, medical
     advancements, or related topics.
entertainment: The text pertains to movies, music,
     celebrities, or related topics.
geography: The text involves geographical
     information, locations, or related topics.

If the text contains multiple topics, choose the
     dominant topic. For ambiguous or unclear topics
     , select the category that best reflects the
     overall content. Please provide a single
     classification for each input.

text: {{text}}
category:
```

Listing 6: MasakhaNEWS Prompt 1

```
Given the categories technology, business, politics,
      sports, health, entertainment, or religion;
     what category does the text: '{{headline}}'
     belong to:

Return only the one category
```

Listing 7: MasakhaNEWS Prompt 2

```
Does this {{language}} topic; '{{headline}}' belong
     to one of the following categories: technology,
      business, politics, sports, health,
     entertainment, or religion? category only
```

Listing 8: MasakhaNEWS Prompt 3

```
You are an assistant able to classify topics in
     texts.

Given the categories technology, religion, politics,
      sports, health, entertainment, or business;
     what is

text: {{headline}}
category:
```

Listing 9: MasakhaNEWS Prompt 4

```
Label the following text as technology, religion,
     politics, sports, health, entertainment, or
     geography. Provide only the category as your
     response.

text: {{headline}}
category:
```

Listing 10: MasakhaNEWS Prompt 5

```
You are tasked with performing topic classification
     on the following {{language}} text. For each
     input, classify the topic as technology,
     business, politics, sports, health,
     entertainment, or religion. Use the following
     guidelines:

technology: The text discusses scientific
     discoveries, technological advancements, or
     related topics.
politics: The text covers political events, policies
     , or related topics.
sports: The text talks about sports events, athletes
     , or related topics.
health: The text addresses health issues, medical
     advancements, or related topics.
entertainment: The text pertains to movies, music,
     celebrities, or related topics.
religion: The text talks about relgions, religious
     institutions and beliefs or related topics.
business: The text covers economy, business, or
     related topics.

If the text contains multiple topics, choose the
     dominant topic. For ambiguous or unclear topics
     , select the category that best reflects the
     overall content. Please provide a single
     classification for each input.

text: {{headline}}
category:
```

Intent Detection prompts:

Listing 1: IngongoIntent Prompt 1

```
Given the text: '{{text}}', classify it into one of
     these intents: [alarm, balance, bill_balance,
     book_flight, book_hotel, calendar_update,
     cancel_reservation, car_rental,
     confirm_reservation, cook_time, exchange_rate,
     food_last, freeze_account, ingredients_list,
     interest_rate, international_visa, make_call,
     meal_suggestion, min_payment, pay_bill,
     pin_change, play_music, plug_type, recipe,
     restaurant_reservation, restaurant_reviews,
     restaurant_suggestion, share_location,
     shopping_list_update, spending_history, text,
     time, timezone, transactions, transfer,
     translate, travel_notification,
     travel_suggestion, update_playlist, weather].
     Only output one intent from the list.
```

Listing 2: IngongoIntent Prompt 2

```
Analyze the text: '{{text}}'. Choose the most
     appropriate intent from these options: [alarm,
     balance, bill_balance, book_flight, book_hotel,
      calendar_update, cancel_reservation,
     car_rental, confirm_reservation, cook_time,
     exchange_rate, food_last, freeze_account,
     ingredients_list, interest_rate,
     international_visa, make_call, meal_suggestion,
      min_payment, pay_bill, pin_change, play_music,
      plug_type, recipe, restaurant_reservation,
     restaurant_reviews, restaurant_suggestion,
     share_location, shopping_list_update,
     spending_history, text, time, timezone,
     transactions, transfer, translate,
     travel_notification, travel_suggestion,
     update_playlist, weather]. Respond with only
     the selected intent.
```

Listing 3: IngongoIntent Prompt 3

```
You are a linguistic analyst trained to understand
     user intent. Based on the text: '{{text}}',
     choose the intent that best matches from this
     list: [alarm, balance, bill_balance,
     book_flight, book_hotel, calendar_update,
     cancel_reservation, car_rental,
     confirm_reservation, cook_time, exchange_rate,
     food_last, freeze_account, ingredients_list,
     interest_rate, international_visa, make_call,
     meal_suggestion, min_payment, pay_bill,
     pin_change, play_music, plug_type, recipe,
     restaurant_reservation, restaurant_reviews,
     restaurant_suggestion, share_location,
     shopping_list_update, spending_history, text,
     time, timezone, transactions, transfer,
     translate, travel_notification,
     travel_suggestion, update_playlist, weather].
     Return only the intent.
```

Listing 4: IngongoIntent Prompt 4

```
You are a English linguistic analyst trained to
     understand {{language}} user intent. Based on
     the {{language}} text: "{{text}}", choose the
     intent that best matches from this list: [alarm
     , balance, bill_balance, book_flight,
     book_hotel, calendar_update, cancel_reservation
     , car_rental, confirm_reservation, cook_time,
     exchange_rate, food_last, freeze_account,
     ingredients_list, interest_rate,
     international_visa, make_call, meal_suggestion,
      min_payment, pay_bill, pin_change, play_music,
      plug_type, recipe, restaurant_reservation,
     restaurant_reviews, restaurant_suggestion,
     share_location, shopping_list_update,
     spending_history, text, time, timezone,
     transactions, transfer, translate,
     travel_notification, travel_suggestion,
     update_playlist, weather]. Return only the
     intent.
```

Listing 5: IngongoIntent Prompt 5

```
The following text is in {{language}}: '{{text}}'.
     Given the list of intents: [alarm, balance,
     bill_balance, book_flight, book_hotel,
     calendar_update, cancel_reservation, car_rental
     , confirm_reservation, cook_time, exchange_rate
     , food_last, freeze_account, ingredients_list,
     interest_rate, international_visa, make_call,
     meal_suggestion, min_payment, pay_bill,
     pin_change, play_music, plug_type, recipe,
     restaurant_reservation, restaurant_reviews,
     restaurant_suggestion, share_location,
     shopping_list_update, spending_history, text,
     time, timezone, transactions, transfer,
     translate, travel_notification,
     travel_suggestion, update_playlist, weather],
     identify the intent expressed in the text.
     Return only the identified intent.
```

Hate Speech prompts:

Listing 1: AfriHate Prompt 1

```
I am providing you with the definition Hate speech,
     Abusive language and Normal tweets.
Hate speech is a language content that expresses
     hatred towards a particular group or individual
      based on their political affiliation, race,
     ethnicity, religion, gender, sexual orientation
     , or other characteristics. It also includes
     threats of violence
Abusive language is any form of bad language
     expressions including rude, impolite, insulting
      or belittling utterance intended to offend or
     harm an individual.
Normal does not contain any bad language.

Tweet: {{tweet}}

Which category does the tweet above belong to: 'Hate
     ', 'Abuse' or 'Normal'. Pick exactly one
     category. Return only the label
```

Listing 2: AfriHate Prompt 2

```
Read the following label definitions and provide a
     label without any explanations.

Hate: Hate speech is public speech that expresses
     hate or encourages violence towards a person or
      group based on something such as race,
     religion, gender, ethnicity, sexual orientation
      or other characteristics.

Abusive: Abusive and offensive language means verbal
      messages that use words in an inappropriate
     way and may include but is not limited to
     swearing, name-calling, or profanity. Offensive
      language may upset or embarrass people because
      it is rude or insulting.

Normal: Normal language is neither hateful nor
     abusive or offensive. It does not contain any
     bad language.

Text: {{tweet}}
Label:
```

Listing 3: AfriHate Prompt 3

```
Read the following text and definitions:

Text: {{tweet}}.

Definitions:
Hate: Hate speech is public speech that expresses
     hate or encourages violence towards a person or
      group based on something such as race,
     religion, gender, ethnicity, sexual orientation
      or other characteristics.

Abuse: Abusive and offensive language means verbal
     messages that use words in an inappropriate way
      and may include but is not limited to swearing
     , name-calling, or profanity. Offensive
     language may upset or embarrass people because
     it is rude or insulting.
```

```
Normal: Normal language is neither hateful nor
     abusive or offensive. It does not contain any
     bad language.

Which of these definitions (hate, abuse, normal)
     apply to this tweet?, return only the label
```

Listing 4: AfriHate Prompt 4

```
Read the following definitions and text to
     categorize:

Definitions:
Hate: Hate speech is public speech that expresses
     hate or encourages violence towards a person or
      group based on something such as race,
     religion, gender, ethnicity, sexual orientation
      or other characteristics.

Abuse: Abusive and offensive language means verbal
     messages that use words in an inappropriate way
      and may include but is not limited to swearing
     , name-calling, or profanity. Offensive
     language may upset or embarrass people because
     it is rude or insulting.

Normal: Normal language is neither hateful nor
     abusive or offensive. It does not contain any
     bad language.

Text: {{tweet}}.

Which of these definitions (hate, abuse, normal)
     apply to this tweet? Return only the label
```

Listing 5: AfriHate Prompt 5

```
You will be given a text snippet and 3 category
     definitions.
Your task is to choose which category applies to
     this text.

Your text snippet is: {{tweet}}.

Your category definitions are:
HATE category definition: Hate speech is public
     speech that expresses hate or encourages
     violence towards a person or group based on
     something such as race, religion, gender,
     ethnicity, sexual orientation or other
     characteristics.

ABUSE category definition: Abusive and offensive
     language means verbal messages that use words
     in an inappropriate way and may include but is
     not limited to swearing, name-calling, or
     profanity. Offensive language may upset or
     embarrass people because it is rude or
     insulting.

NORMAL category definition: Normal language is
     neither hateful nor abusive or offensive. It
     does not contain any bad language.

Does the text snippet belong to the HATE, ABUSIVE,
     or the NORMAL category? Thinking step by step
     answer HATE, ABUSIVE, or NORMAL capitalizing
     all the letters.
Explain your reasoning FIRST, then output HATE,
     ABUSIVE, or NORMAL. Clearly return the label in
      capital letters.
```

Natural Language Inference prompts:

Listing 1: AfriXNLI Prompt 1

```
Please identify whether the premise entails or
     contradicts the hypothesis in the following
     premise and hypothesis. The answer should be
     exact entailment, contradiction, or neutral.

Premise: {{premise}}
Hypothesis: {{hypothesis}}.
Is it entailment, contradiction, or neutral?
```

Listing 2: AfriXNLI Prompt 2

```
{{premise}}
Question: {{hypothesis}} True, False, or Neither?
Answer:
```

Listing 3: AfriXNLI Prompt 3

```
Given the following premise and hypothesis in {{
     language}}, identify if the premise entails,
     contradicts, or is neutral towards the
     hypothesis. Please respond with exact '
     entailment', 'contradiction', or 'neutral'.

Premise: {{premise}}
Hypothesis: {{hypothesis}}
```

Listing 4: AfriXNLI Prompt 4

```
You are an expert in Natural Language Inference (NLI
     ) specializing in {{language}} language.
Analyze the premise and hypothesis given in {{
     language}}, and determine the relationship
     between them.
Respond with one of the following options: '
     entailment', 'contradiction', or 'neutral'.

Premise: {{premise}}
Hypothesis: {{hypothesis}}
```

Listing 5: AfriXNLI Prompt 5

```
Based on the given statement, is the following claim
      'true', 'false', or 'inconclusive'.

Statement: {{premise}}
Claim: {{hypothesis}}
```

H.2 Question Answering

CrosslingualQA prompts:

Listing 1: AfriQA Prompt 1

```
Your task is to answer a question given a context.
Make sure you respond with the shortest span
     containing the answer in the context.
Question: {{question_lang}}
Context: {{context}}
Answer:
```

Listing 2: AfriQA Prompt 2

```
Your task is to answer a question given a context.
     The question is in {{language}}, while the
     context is in English or French.
Make sure you respond with the shortest span in the
     context that contains the answer.
Question: {{question_lang}}
Context: {{context}}
Answer:
```

Listing 3: AfriQA Prompt 3

```
Given the context, provide the answer to the
     following question.
Ensure your response is concise and directly from
     the context.
Question: {{question_lang}}
Context: {{context}}
Answer:
```

Listing 4: AfriQA Prompt 4

```
You are an AI assistant and your task is to answer
     the question based on the provided context.
Your answer should be the shortest span that
     contains the answer within the context.
Question: {{question_lang}}
Context: {{context}}
Answer:
```

Listing 5: AfriQA Prompt 5

```
Using the context, find the answer to the question.
Respond with the briefest span that includes the
     answer from the context.
Question: {{question_lang}}
Context: {{context}}
Answer:
```

Reading Comprehension prompts:

Listing 1: Belebele Prompt 1

```
P: {{passage}}
Q: {{question}}
A: {{option_1}}
B: {{option_2}}
C: {{option_3}}
D: {{option_4}}
Please choose the correct answer from the options
     above:
```

Listing 2: Belebele Prompt 2

```
Passage: {{passage}}
Question: {{question}}
1: {{option_1}}
2: {{option_2}}
3: {{option_3}}
4: {{option_4}}
Please select the correct answer from the given
     choices
```

Listing 3: Belebele Prompt 3

```
Context: {{passage}}
Query: {{question}}
Option A: {{option_1}}
Option B: {{option_2}}
Option C: {{option_3}}
Option D: {{option_4}}
Please indicate the correct option from the list
     above:
```

Listing 4: Belebele Prompt 4

```
{{passage}}
Based on the above passage, answer the following
     question:
{{question}}
Choices:
A) {{option_1}}
B) {{option_2}}
C) {{option_3}}
D) {{option_4}}
Please provide the correct answer from the choices
     given
```

Listing 5: Belebele Prompt 5

```
Read the passage: {{passage}}
Then answer the question: {{question}}
Options:
A. {{option_1}}
B. {{option_2}}
C. {{option_3}}
D. {{option_4}}
Please choose the correct option from the above list
```

Listing 6: NaijaRC Prompt 1

```
P: {{story}}
Q: {{question}}
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}
Please choose the correct answer from the options
     above
```

Listing 7: NaijaRC Prompt 2

```
Passage: {{story}}
Question: {{question}}
1: {{options_A}}
2: {{options_B}}
3: {{options_C}}
4: {{options_D}}
Please select the correct answer from the given
     choices
```

Listing 8: NaijaRC Prompt 3

```
Context: {{story}}
Query: {{question}}
Option A: {{options_A}}
Option B: {{options_B}}
Option C: {{options_C}}
Option D: {{options_D}}
Please indicate the correct option from the list
     above
```

Listing 9: NaijaRC Prompt 4

```
{{story}}
Based on the above passage, answer the following
     question
{{question}}
Choices:
A) {{options_A}}
B) {{options_B}}
C) {{options_C}}
D) {{options_D}}
Please provide the correct answer from the choices
     given
```

Listing 10: NaijaRC Prompt 5

```
Read the passage: {{story}}
Then answer the question: {{question}}
Options:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Please choose the correct option from the above list
```

H.3 Knowledge

Arc-E prompts:

Listing 1: UHURA Prompt 1

```
You are a virtual assistant that answers multiple-
     choice questions with the correct option only.

Question: {{question}}

Choices:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Answer:
```

Listing 2: UHURA Prompt 2

```
Choose the correct option that answers the question
     below:
```

```
Question: {{question}}

Choices:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Answer: .
```

Listing 3: UHURA Prompt 3

```
Answer the following multiple-choice question by
     picking 'A', 'B', 'C', or 'D'

Question: {{question}}

Options:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Answer:
```

Listing 4: UHURA Prompt 4

```
Question: {{question}}

Options:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Answer:
```

Listing 5: UHURA Prompt 5

```
Which of the following options answers this question
     : {{question}}

Options:
A. {{options_A}}
B. {{options_B}}
C. {{options_C}}
D. {{options_D}}
Answer:
```

MMLU prompts:

Listing 1: OpenAIMMLU Prompt 1

```
Q: {{Question}}
A: {{A}}
B: {{B}}
C: {{C}}
D: {{D}}
Please choose the correct answer from the options
     above
```

Listing 2: OpenAIMMLU Prompt 2

```
Question: {{Question}}
1: {{A}}
2: {{B}}
3: {{C}}
4: {{D}}
Please select the correct answer from the given
     choices
```

Listing 3: OpenAIMMLU Prompt 3

```
Input Question: {{Question}}
Option A: {{A}}
Option B: {{B}}
Option C: {{C}}
Option D: {{D}}
Please indicate the correct option from the list
     above
```

Listing 4: OpenAIMMLU Prompt 4

```
Critically analyze the question and select the most
     probable answer from the list:
{{Question}}
Choices:
A) {{A}}
B) {{B}}
C) {{C}}
D) {{D}}
```

Listing 5: OpenAIMMLU Prompt 5

```
Answer the question and pick the correct answer from
      the options:
{{Question}}
Options:
A. {{A}}
B. {{B}}
C. {{C}}
D. {{D}}
Please choose the correct option from the above list
```

Listing 6: AfriMMLU Prompt 1

```
You are a highly knowledgeable and intelligent
     artificial intelligence model answers multiple-
     choice questions about {{subject}}.

Question: {{question}}
Choices:
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}

Answer:
```

Listing 7: AfriMMLU Prompt 2

```
As an expert in {{subject}}, choose the most
     accurate answer to the question below. Your
     goal is to select the correct option 'A', 'B',
     'C', or 'D' by understanding the nuances of the
      topic.

Question: {{question}}
Choices:
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}

Answer:
```

Listing 8: AfriMMLU Prompt 3

```
You are a subject matter expert in {{subject}}.
     Utilizing your expertise in {{subject}}, answer
      the following multiple-choice question by
     picking 'A', 'B', 'C', or 'D'.

Question: {{question}}
Choices:
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}

Answer:
```

Listing 9: AfriMMLU Prompt 4

```
Analyze each question critically and determine the
     most correct option based on your understanding
      of the subject matter
Question: {{question}}
Choices:
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}

Answer:
```

Listing 10: AfriMMLU Prompt 5

```
Given your proficiency in {{subject}}, please answer
      the subsequent multiple-choice question
Question: {{question}}
Choices:
A: {{options_A}}
B: {{options_B}}
C: {{options_C}}
D: {{options_D}}

Answer:
```

H.4 Reasoning

Math prompts: from IROKOBENCH (Adelani et al., 2024b)

Listing 1: AfriMGSM Prompt 1

```
{{question}}
Step-by-step Answer:
```

Listing 2: AfriMGSM Prompt 2

```
Give direct numerical answers for the question
     provided.

Question: {{question}}
Step-by-step Answer:
```

Listing 3: AfriMGSM Prompt 3

```
Solve the following math question

Question: {{question}}
Step-by-step Answer:
```

Listing 4: AfriMGSM Prompt 4

```
Answer the given question with the appropriate
     numerical value, ensuring that the response is
     clear and without any supplementary information
     .

Question: {{question}}
Step-by-step Answer:
```

Listing 5: AfriMGSM Prompt 5

```
For mathematical questions provided in {{language}}
     language. Supply the accurate numeric step by
     step answer to the provided question.

Question: {{question}}
Step-by-step Answer:
```

H.5 Text Generation

Machine Translation prompts

Listing 1: Machine Translation Prompt 1

```
{{source_lang}} sentence: {{source_text}}
{{arget_lang}} sentence:
```

Listing 2: Machine Translation Prompt 2

```
You are a translation expert. Translate the
     following {{source_lang}} sentences to {{
     target_lang}}

{{source_lang}} sentence: {{source_text}}
{{target_lang}} sentence:
```

Listing 3: Machine Translation Prompt 3

```
As a {{source_lang}} and {{target_lang}} linguist,
     translate the following {{source_lang}}
     sentences to {{target_lang}}.

{{source_lang}} sentence: {{source_text}}
{{target_lang}} sentence:
```

Summarization prompts

Listing 1: XL-SUM Prompt 1

```
Provide a summary of the document written in {{
     language}}. Ensure that you provide the summary
      in {{language}} and nothing else.

Document in {{language}}: {{text}}

Summary:
```

Listing 2: XL-SUM Prompt 2

```
Summarize the document below in triple backticks and
      return only the summary and nothing else.

{{text}}
```

Listing 3: XL-SUM Prompt 3

```
You are an advanced Summarizer, a specialized
     assistant designed to summarize documents in {{
     language}}. Your main goal is to ensure
     summaries are concise and informative.
Ensure you return the summary only and nothing else.

Document: {{text}}

Summary:
```

Diacritics Restoration prompts

Listing 1: AFRIADR Prompt 1

```
Please restore the missing diacritics in the
     following sentence: {{text}}.
Return output sentence only
```

Listing 2: AFRIADR Prompt 2

```
Given a sentence without diacritics, add the
     appropriate diacritics to make it grammatically
      and semantically correct.
Sentence: {{text}}.
Return output sentence only
```

Listing 3: AFRIADR Prompt 3

```
This text is in {{language}}. Restore all
     diacritical marks to their proper places in the
      following sentence: {{text}}. Return output
     sentence only
```

Listing 4: AFRIADR Prompt 4

```
You are a linguist specializing in diacritical marks
      for {{language}}. Add the appropriate
     diacritics to this {{language}} sentence: {{
     text}}. Return output sentence only
```

Listing 5: AFRIADR Prompt 5

```
You are a linguist specializing in diacritical marks
      for {{language}}. Diacritics are essential for
      proper pronunciation and meaning in {{language
     }}. You are tasked with converting {{language}}
      sentences  without diacritics into their
     correctly accented forms. Here's the input: {{
     text}}. Return output sentence only
```

I Detailed Results Per Language

This appendix presents detailed per-language performance results for each dataset. We group them by the task category shown in Figure 2. Each figure shows model performance using the best prompt for each model–dataset pairs.

I.1 Natural Language Understanding (NLU)

I.1.1 POS

MasakhaPOS

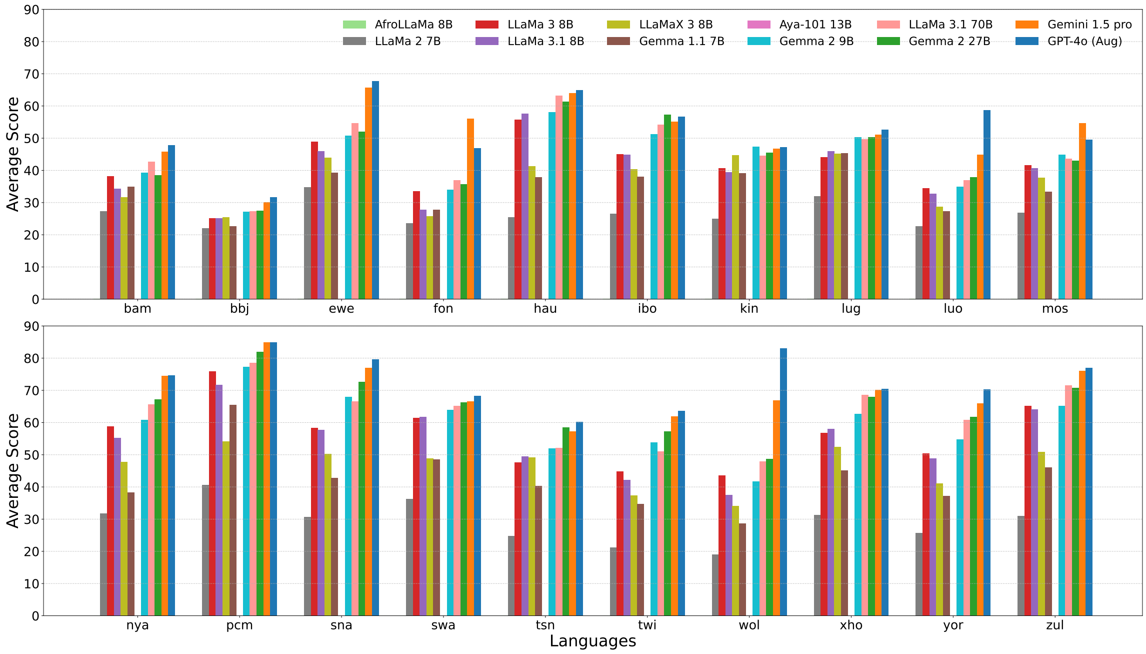

Figure 6: Per-language performance results for the MasakhaPOS dataset.

I.1.2 NER

MasakhaNER

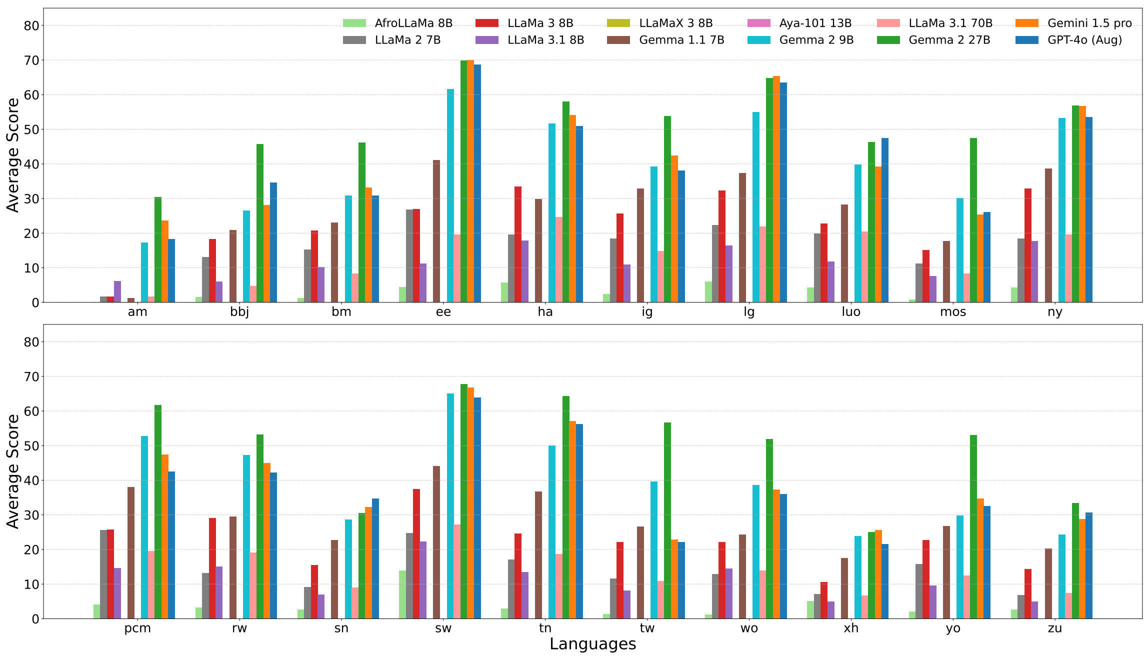

Figure 7: Per-language performance results for the MasakhaNER dataset.

I.1.3 Sentiment Analysis

AfriSenti

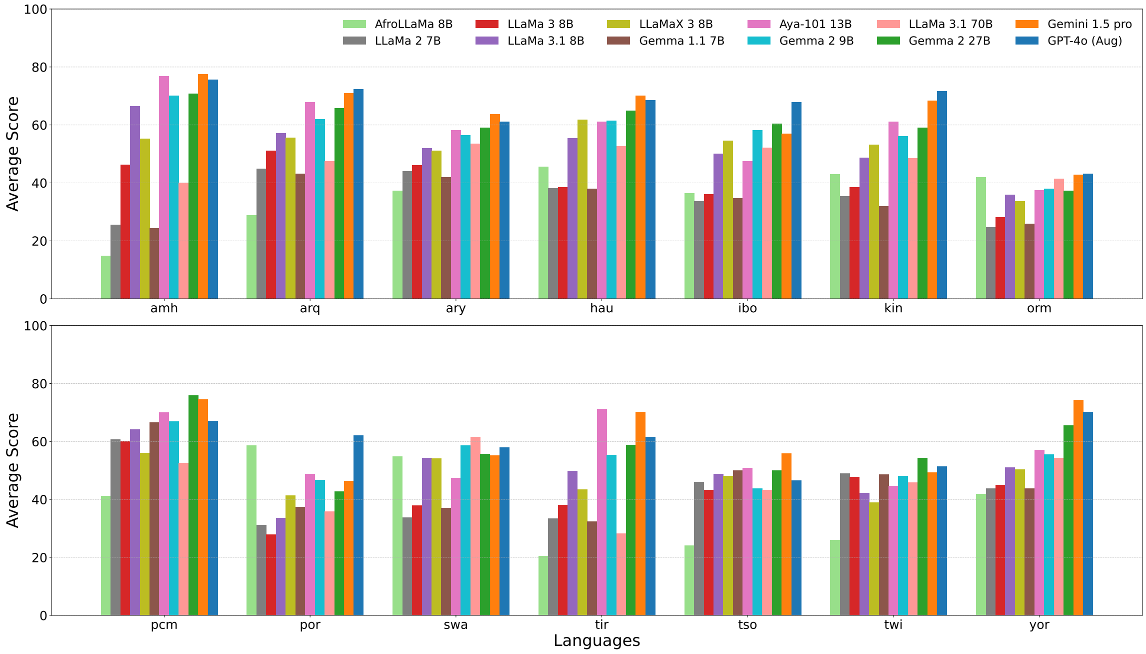

Figure 8: Per-language performance results for the AfriSenti dataset.

NollySenti

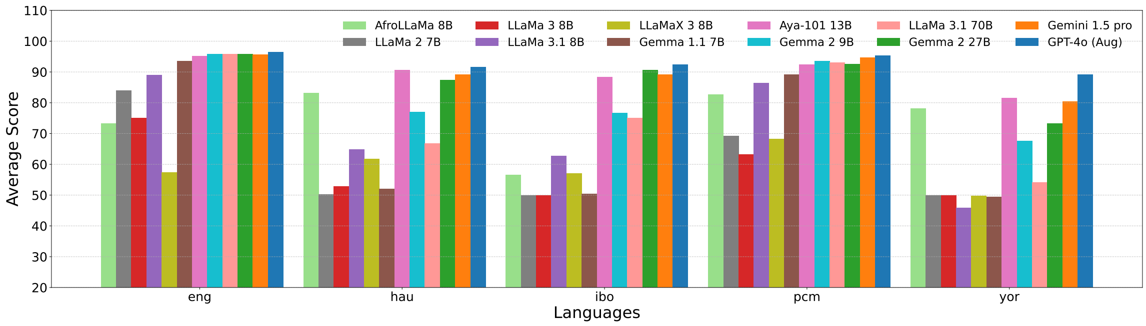

Figure 9: Per-language performance results for the NollySenti dataset.

I.1.4 Intent Detection

Injongo Intent

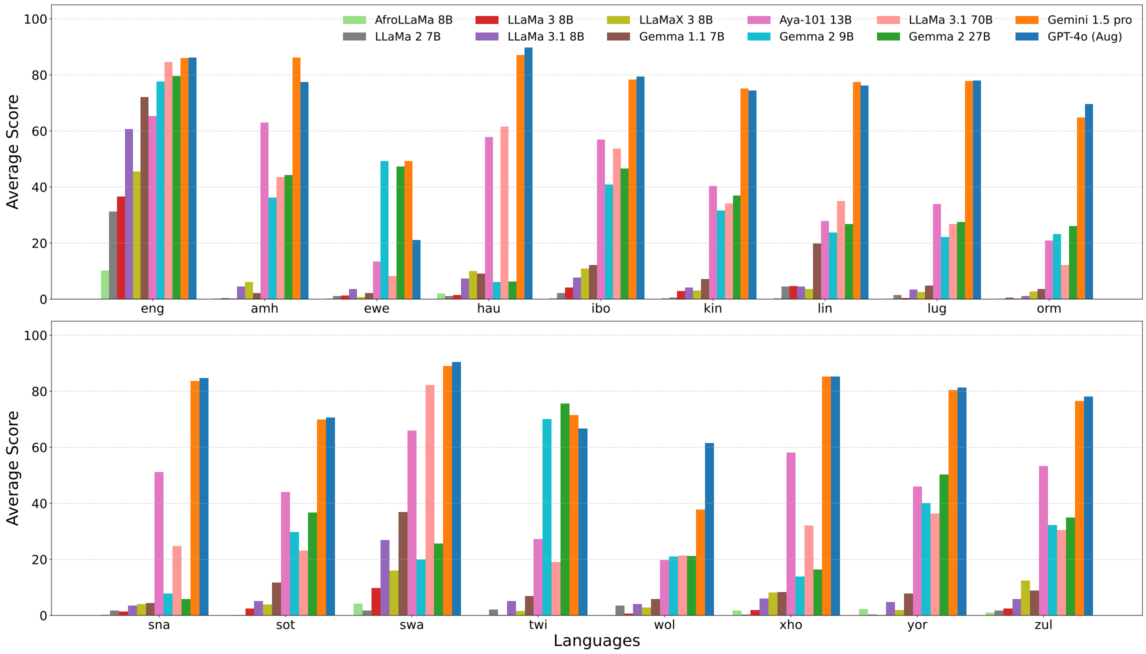

Figure 10: Per-language performance results for the InjongoIntent dataset.

I.1.5 Topic Classification

MasakhaNEWS

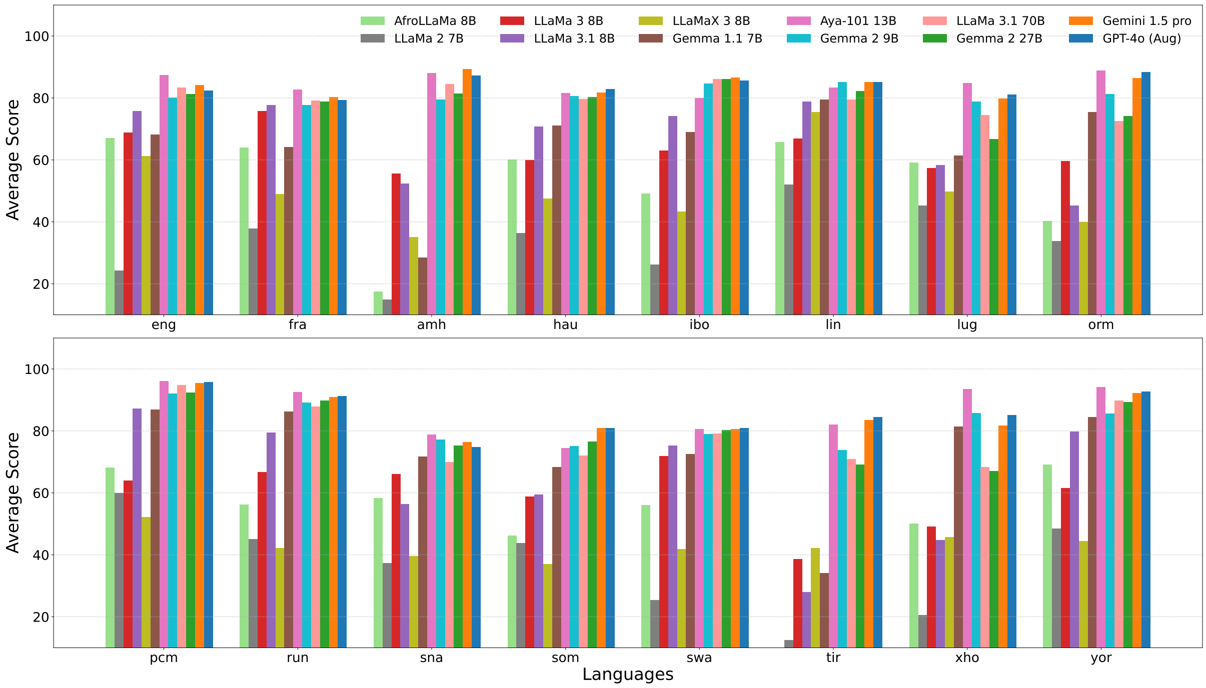

Figure 11: Per-language performance results for the MasakhaNEWS dataset.

SIB

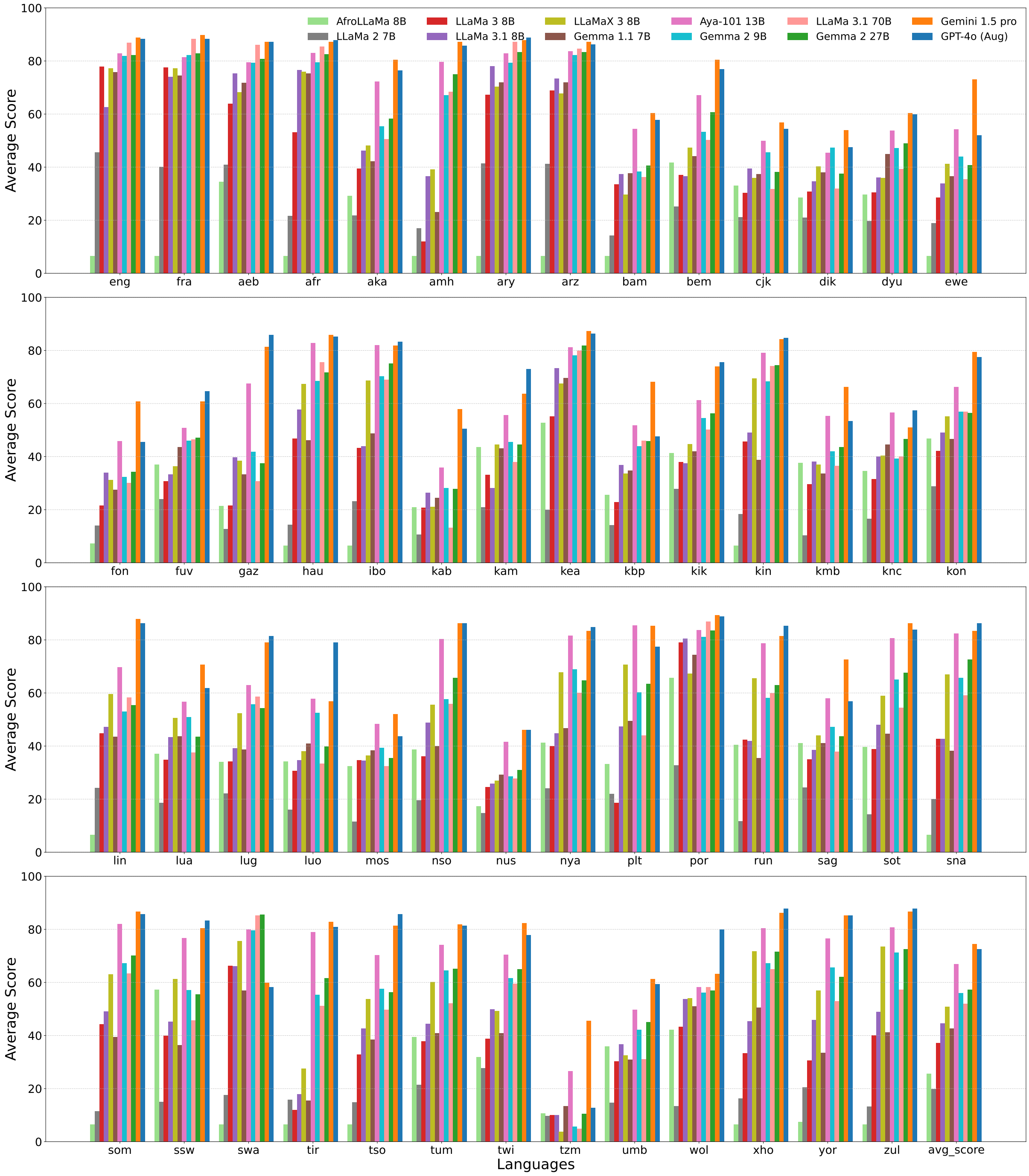

Figure 12: Per-language performance results for the SIB dataset.

I.1.6 Hate Speech: AfriHate

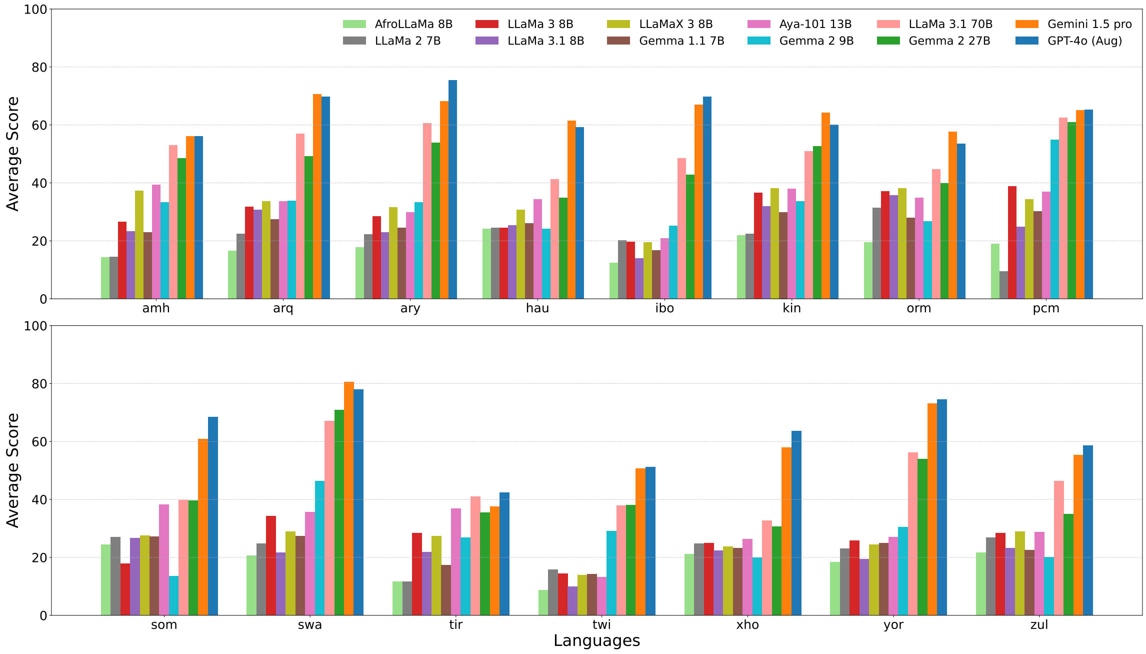

Figure 13: Per-language performance results for the AfriHate dataset.

I.2 Natural Language Inference

AfriXNLI

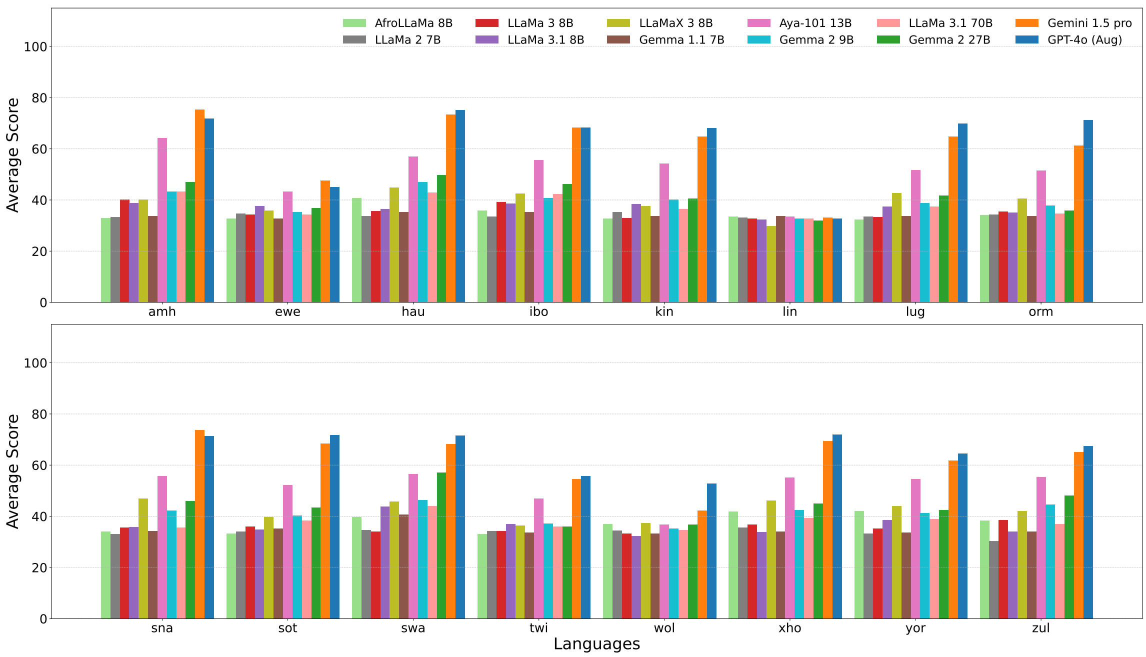

Figure 14: Per-language performance results for the AFRIXNLI dataset.

I.3 Question Answering

I.3.1 Cross-lingual Question Answering

AfriQA

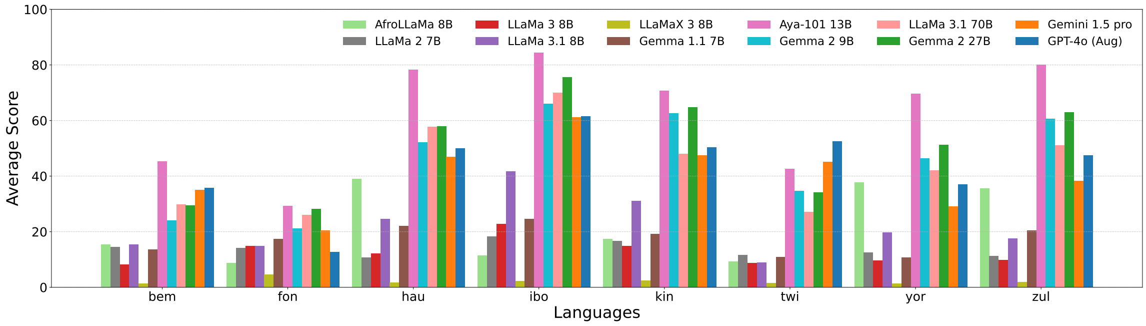

Figure 15: Per-language performance results for the AFRIQA dataset.

I.3.2 Reading Comprehension

Belebele

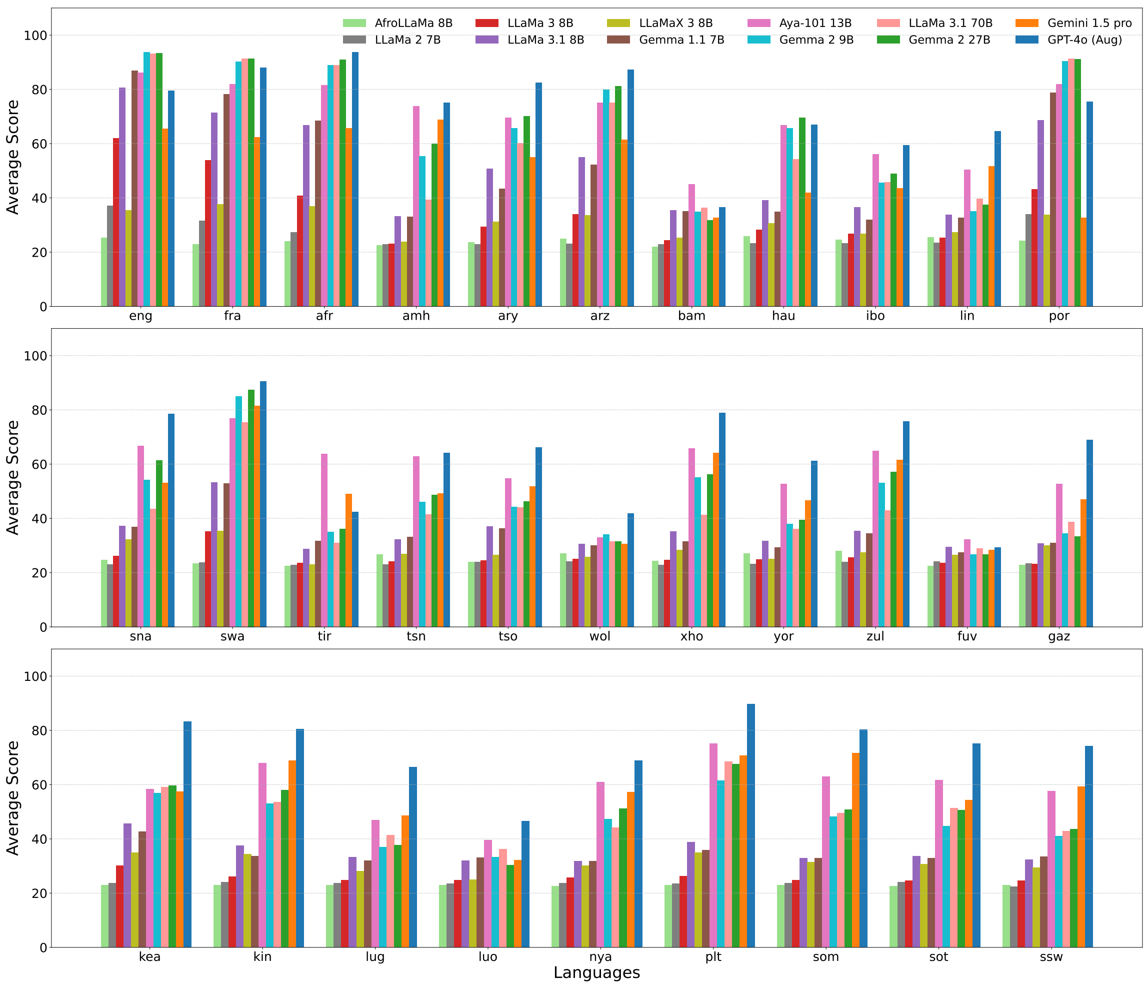

Figure 16: Per-language performance results for the BELEBELE dataset.

NaijaRC

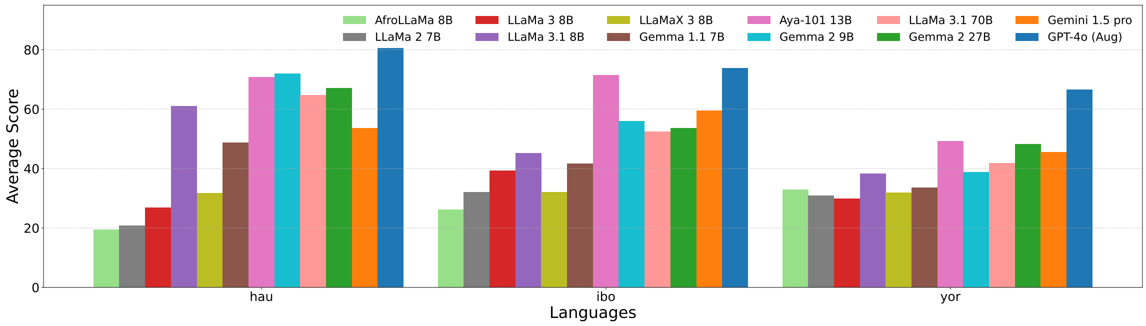

Figure 17: Per-language performance results for the NAIJARC dataset.

I.4 Knowledge

I.4.1 Arc-E

UHURA

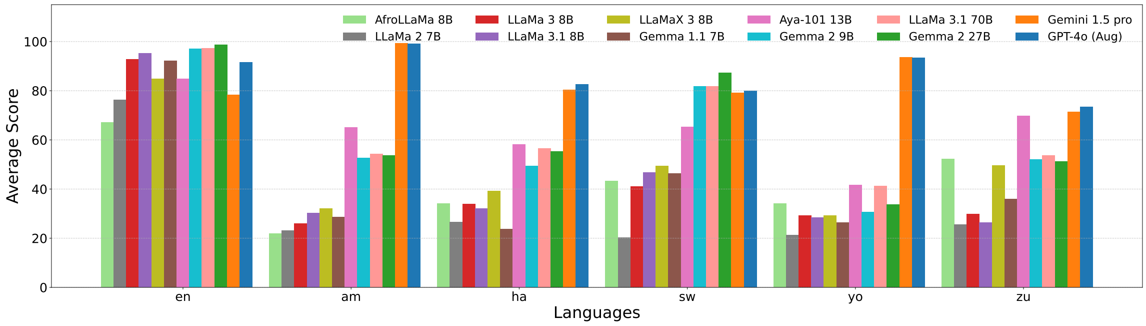

Figure 18: Per-language performance results for the UHURA dataset.

I.4.2 MMLU

OpenAIMMLU

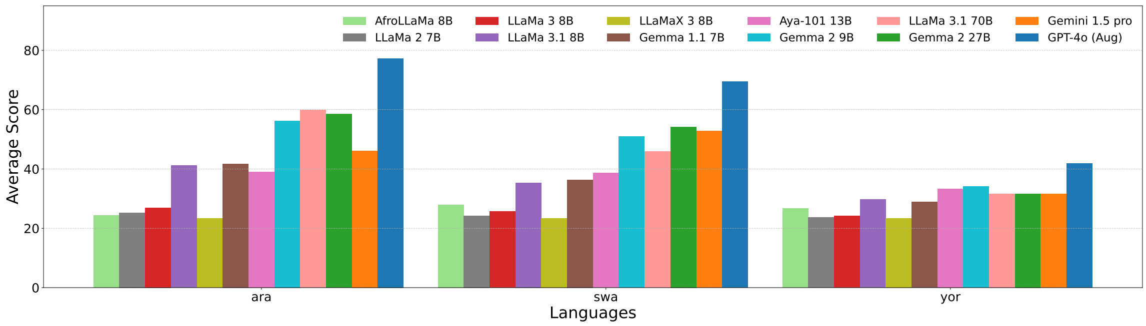

Figure 19: Per-language performance results for the OPENAI-MMLU dataset.

AfriMMLU

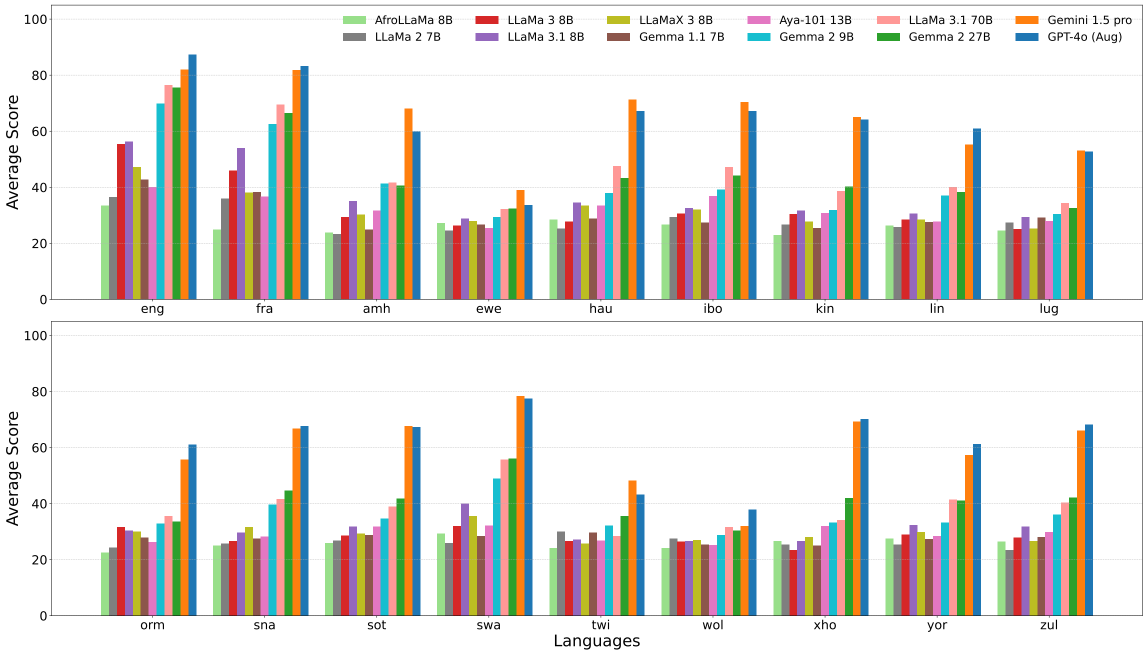

Figure 20: Per-language performance results for the AFRIMMLU dataset.

I.5 Reasoning

AfriMGSM

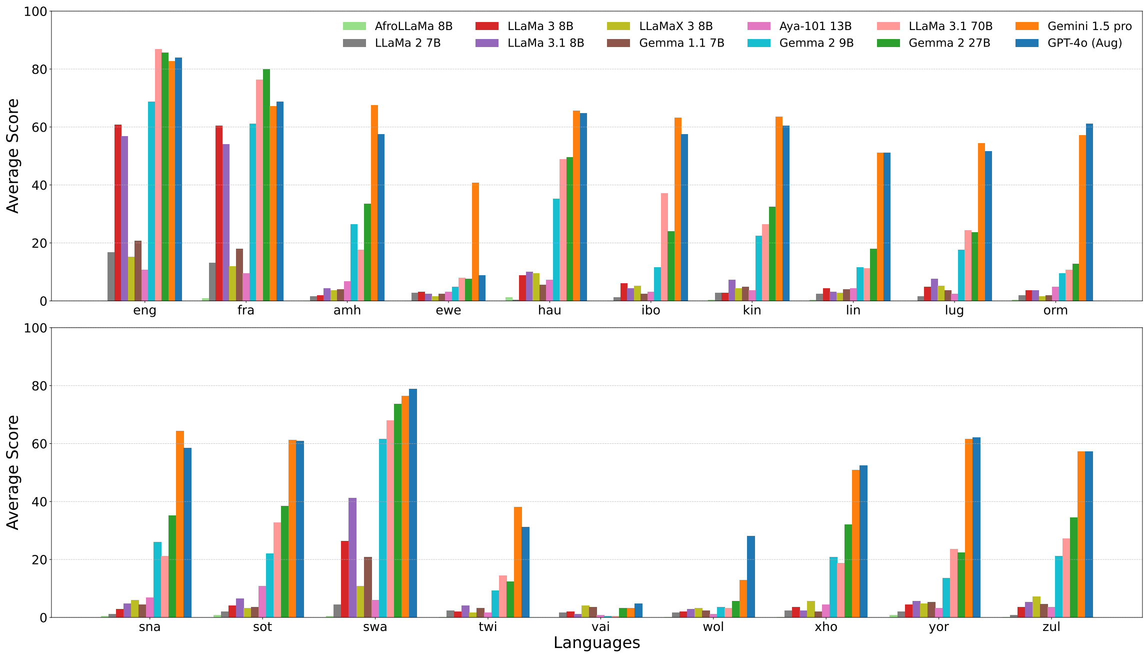

Figure 21: Per-language performance results for the AFRIMGSM dataset.

I.6 Text Generation

I.6.1 Machine Translation

SALT (*en/fr-xx*)

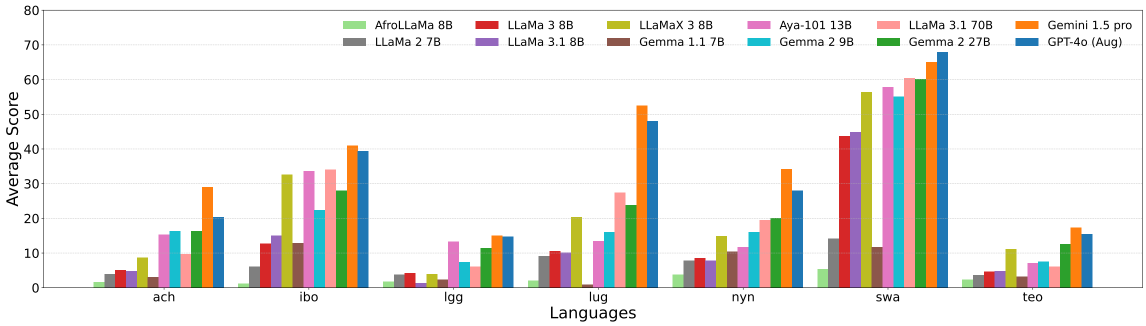

Figure 22: Per-language performance results for the SALT dataset (*en/fr-xx*).

SALT (*xx-en/fr*)

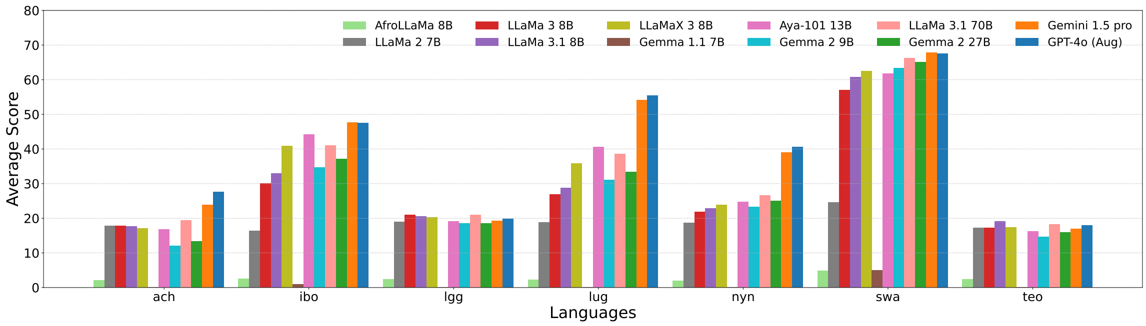

Figure 23: Per-language performance results for the SALT dataset (*xx-en/fr*).

MAFAND (*en-xx/fr*)

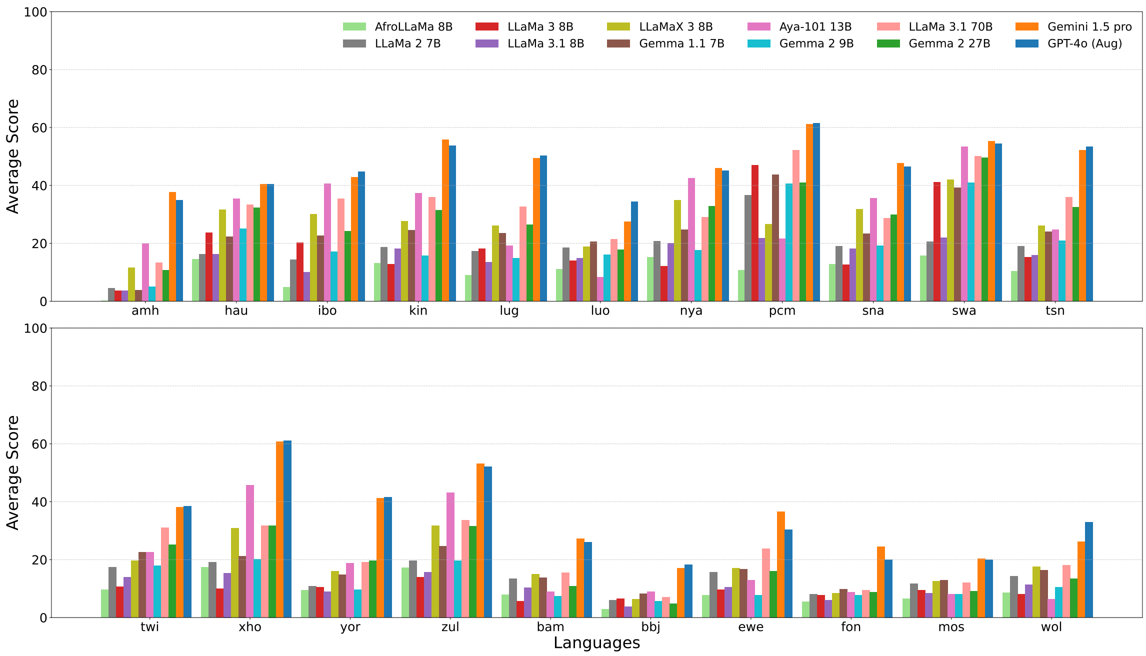

Figure 24: Per-language performance results for the MAFAND dataset.

MAFAND (*xx-en/fr*)

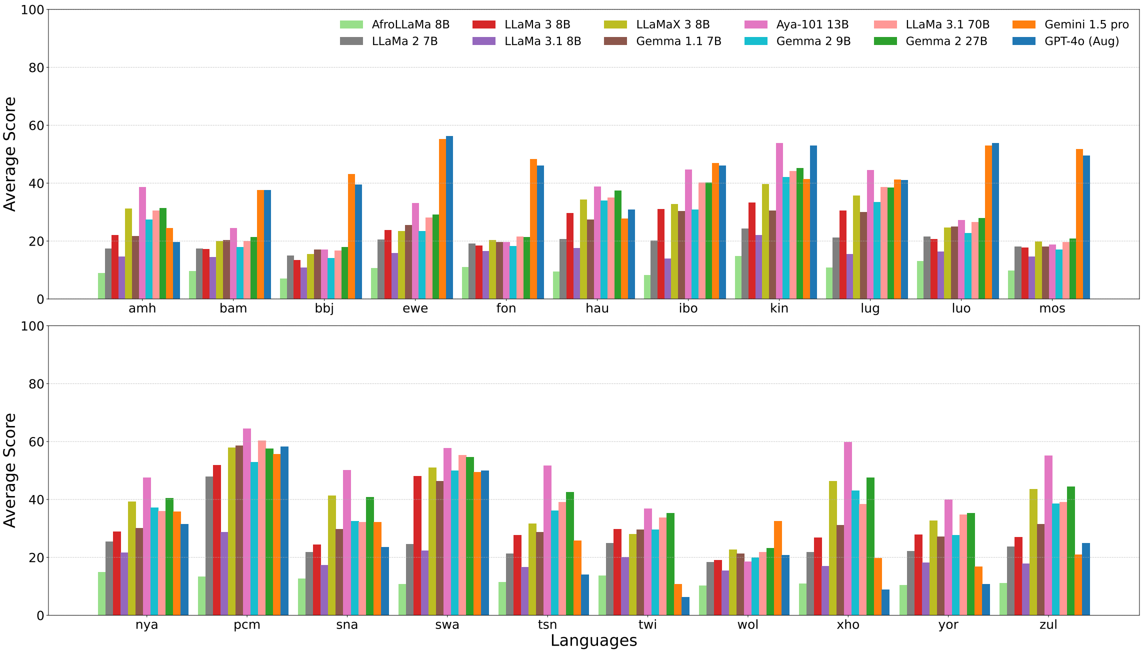

Figure 25: Per-language performance results for the MAFAND dataset.

NTREX (*en/fr-xx*)

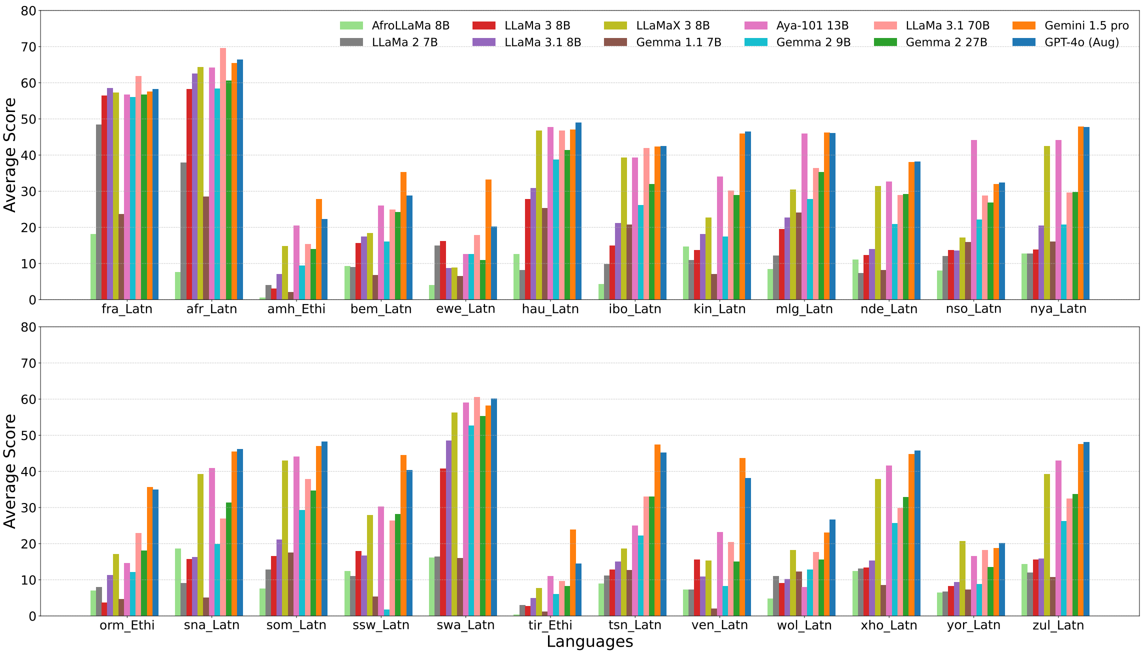

Figure 26: Per-language performance results for the NTREX-128 dataset (*en/fr-xx*).

NTREX (*xx-en/fr*)

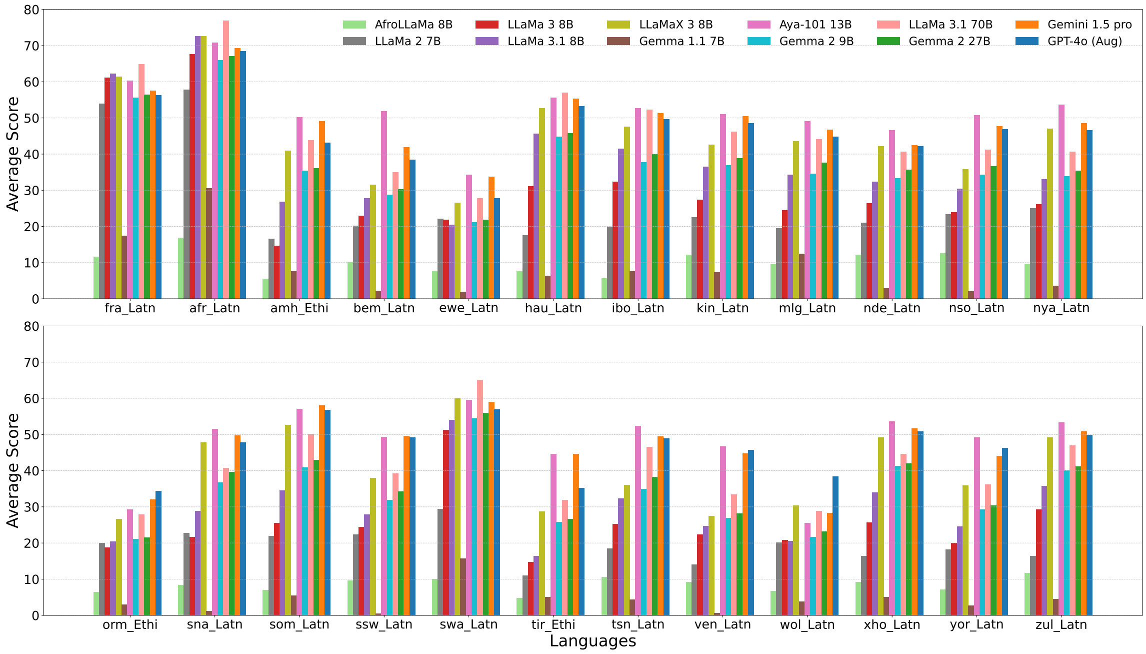

Figure 27: Per-language performance results for the NTREX-128 dataset (*xx-en/fr*).

Flores (African Languages only and French) (*en/fr-xx*)

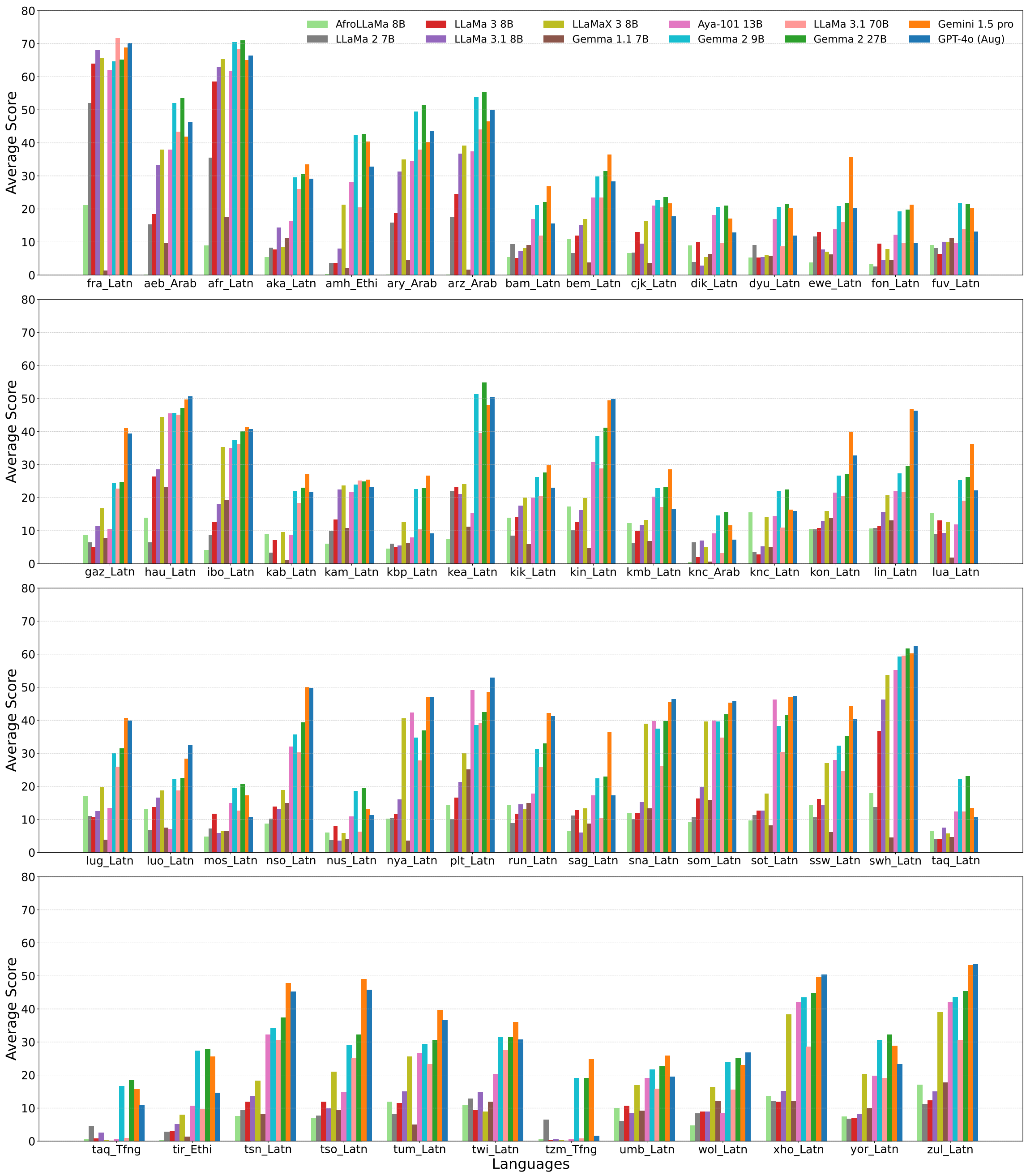

Figure 28: Per-language performance results for the FLORES dataset (*en/fr-xx*).

Flores (African Languages only and French) (*xx-en/fr)*

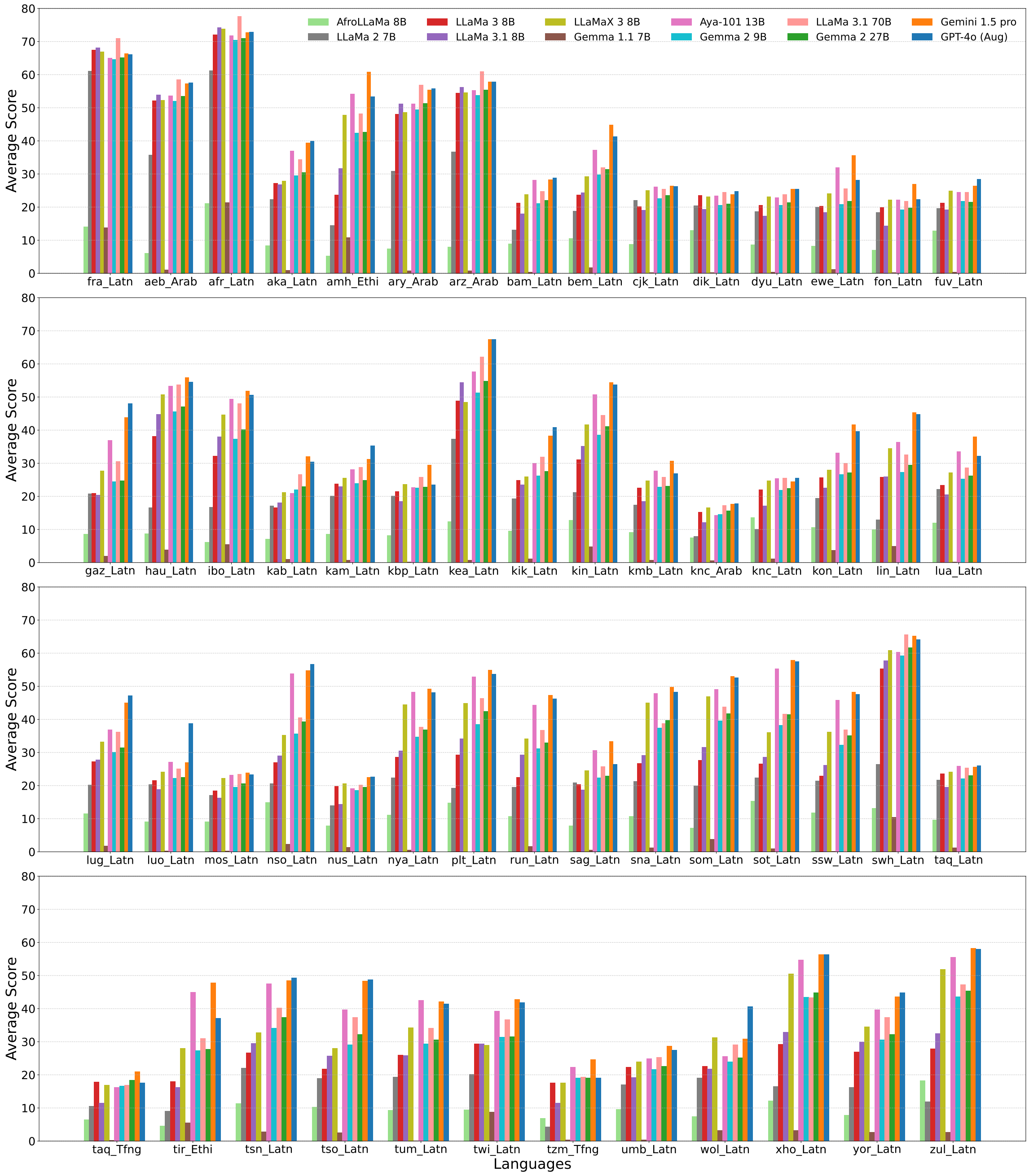

Figure 29: Per-language performance results for the FLORES dataset (*xx-en/fr*).

I.6.2 Summarization

XL-SUM

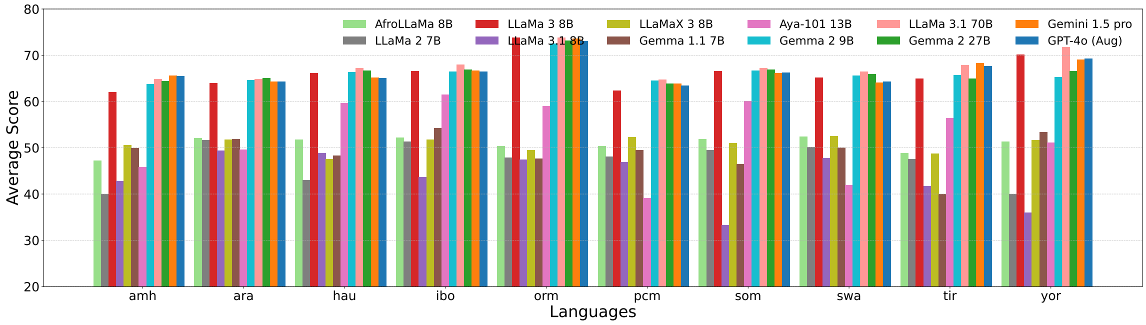

Figure 30: Per-language performance results for the XL-SUM dataset.

I.6.3 Diacritics Restoration

AFRIADR

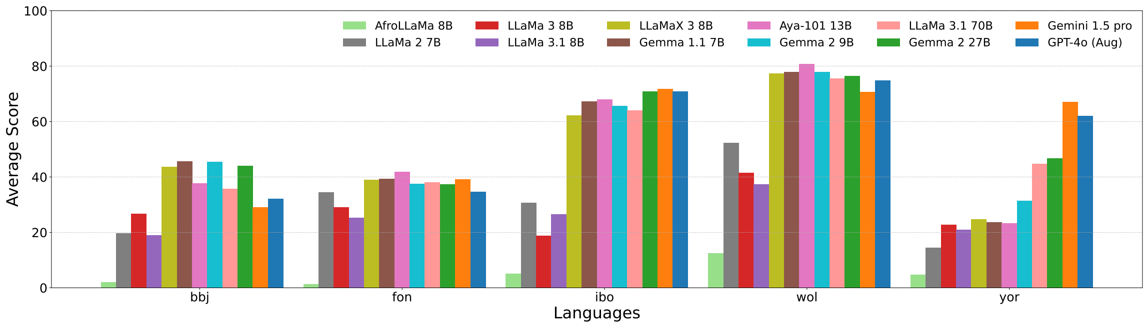

Figure 31: Per-language performance results for the AFRIADR dataset.