IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models

David Ifeoluwa Adelani[1,2*], Jessica Ojo[1,3*], Israel Abebe Azime[4*], Jian Yun Zhuang[5], Jesujoba O. Alabi[4*], Xuanli He[6], Millicent Ochieng[7], Sara Hooker[8], Andiswa Bukula[9], En-Shiun Annie Lee[10], Chiamaka Chukwuneke[11], Happy Buzaaba[12], Blessing Sibanda[*], Godson Kalipe[*], Jonathan Mukiibi[13*], Salomon Kabongo[14*], Foutse Yuehgoh[15*], Mmasibidi Setaka[9], Lolwethu Ndolela[*], Nkiruka Odu[*], Rooweither Mabuya[9], Shamsuddeen Hassan Muhammad[16], Salomey Osei[17*], Sokhar Samb[18*], Tadesse Kebede Guge[19*], Tombekai Vangoni Sherman[20], Pontus Stenetorp[6]

[*]Masakhane NLP, [1]Mila, McGill University, [2]Canada CIFAR AI Chair, [3]Lelapa AI, [4]Saarland University, [5]University of Toronto, [6]University College London, [7]Microsoft Research Africa, [8]Cohere For AI, [9]SADiLaR, [10]Ontario Tech University, [11]Lancaster University, [12]Princeton university, [13]Makerere University, [14]Leibniz Universität Hannover, [15]Le CNAM, [16]Imperial College London, [17]Universidad de Deusto, [18]DAUST, [19]Haramaya University.

Abstract

Despite the widespread adoption of Large language models (LLMs), their remarkable capabilities remain limited to a few high-resource languages. Additionally, many low-resource languages (*e.g.,* African languages) are often evaluated only on basic text classification tasks due to the lack of appropriate or comprehensive benchmarks outside of high-resource languages. In this paper, we introduce IrokoBench—a human-translated benchmark dataset for 17 typologically-diverse low-resource African languages covering three tasks: natural language inference (AfriXNLI), mathematical reasoning (AfriMGSM), and multi-choice knowledge-based question answering (AfriMMLU). We use IrokoBench to evaluate zero-shot, few-shot, and translate-test settings (where test sets are translated into English) across 10 open and six proprietary LLMs. Our evaluation reveals a significant performance gap between high-resource languages (such as English and French) and low-resource African languages. We observe a significant performance gap between open and proprietary models, with the highest performing open model, Gemma 2 27B only at 63% of the best-performing proprietary model GPT-4o performance. In addition, machine translating the test set to English before evaluation helped to close the gap for larger models that are English-centric, such as Gemma 2 27B and LLaMa 3.1 70B. These findings suggest that more efforts are needed to develop and adapt LLMs for African languages.

1 Introduction

In recent years, the capabilities of large language models (LLMs) have greatly improved, from coherent chat experiences to solving complex and knowledge-intensive tasks like mathematical reasoning, coding, and question answering (QA) (OpenAI et al., 2024; Jiang et al., 2024; Gemini-Team et al., 2024). These models have also demonstrated the ability to quickly learn new and challenging tasks with few in-context learning examples and through chain-of-thought reasoning (Brown et al., 2020; Shi et al., 2022; Wei et al., 2022). However, most state-of-the-art (SoTA) LLMs are primarily trained on high-resource languages (HRLs), resulting in sub-optimal performance for languages unseen during pre-training (Touvron et al., 2023; Ojo et al., 2023). Furthermore, this language coverage bias is reflected in the evaluation stage, predominantly conducted in English and a few other HRLs.

There has been considerable effort to create benchmarks for African languages, but they typically cover simpler tasks, or are specific to narrow tasks such as machine translation, and, more recently, reading comprehension (Bandarkar et al., 2023; Aremu et al., 2023). Some diverse reasoning benchmarks have included Swahili—the most spoken native African language—for tasks like commonsense reasoning (Ponti et al., 2020) and natural language inference (Conneau et al., 2018). Consequently, current multilingual evaluations of LLMs do not accurately reflect their capabilities in reasoning and knowledge-intensive tasks across the majority of African languages.

Furthermore, the few comprehensive evaluations that exist across languages often rely on machine translation of English benchmarks (Singh et al., 2024). While automatic translation from English benchmarks is a popular approach given the cost

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2732–2757
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

and time investment required for human translation, it often suffers from noise and biases (Vanmassenhove et al., 2021; Lee et al., 2022; Khiu et al., 2024; Hartung et al., 2023; Savoldi et al., 2021) or fail to reflect cultural context (Wang et al., 2022; Ji et al., 2023; Pudjiati et al., 2022). Automatic curation may also amplify any of the ubiquitous issues with the quality of broad pretraining sets (Luccioni and Viviano, 2021; Kreutzer et al., 2022; Ferrara, 2023).

In this paper, we seek to address both the *diversity and breadth* of evaluation coverage. We introduce IROKOBENCH, a human curated benchmark dataset for 17 typologically diverse African languages which encompasses three complex tasks: natural language inference (NLI), mathematical reasoning, and multi-choice knowledge-based QA. The datasets were created by human translating a subset of English cross-lingual NLI (XNLI) (Conneau et al., 2018), English Multilingual Grade School Math (MGSM) (Shi et al., 2023), and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021a), evaluation datasets into each of the 16 languages using professional translators.

We conduct a large-scale evaluation of IROKOBENCH to assess zero-shot, few-shot, and translate-test settings (where test sets are translated into English) performance across 10 open and six proprietary LLMs. Our main contributions can be enumerated as follows:

1. **We introduce and release IROKOBENCH**, a human-translated benchmark that includes 16 languages from various geographical regions in Africa, all with varying degrees of "low-resourcedness" (Joshi et al., 2020).
2. **Sharp cliff in performance across all models on low-resource languages** Our evaluation shows a large gap (~45% on average) between the performance of high-resource languages (*e.g.,* English) and African languages on all LLMs evaluated. Notably, Swahili performs better than other African languages, likely due to its large corpus on the web.
3. **Models generally perform poorly in in-language evaluation** This can be attributed to the inability of current SoTA LLMs to respond in the native languages of the users. Machine translating the test set to English before evaluation helped to close the gap for English-centric models; however, requiring users to always translate their prompts to English may not be a desirable behavior.
4. **IROKOBENCH highlights the performance divide between open and proprietary models on low-resource languages.** We find that proprietary closed models generally outperform open models for African languages. However, even these proprietary models exhibit substantial performance drops, due to the limited monolingual web data for African languages. The lowest performance is observed in languages such as *Ewe*, *Lingala*, *Luganda*, *Twi* and *Wolof*, which each have less than 50 million characters of available data (Kudugunta et al., 2023). Among the tasks evaluated, AfriMGSM proves most challenging for LLMs, followed by AfriMMLU and AfriXNLI.

We release IrokoBench on GitHub [1] and HuggingFace under the CC BY-SA 4.0 license [2] to further multilingual evaluation and research.

2 Related Work

Multilingual Evaluation of LLMs: The evaluation of multilingual capabilities of LLMs has garnered significant attention. This has led to an increase in research that explores their performance across diverse linguistic landscapes (Ahuja et al., 2023a,b; Lai et al., 2023b; Hendy et al., 2023; Bang et al., 2023; Üstün et al., 2024; Singh et al., 2024). Despite this growing interest, there remains a notable lack of representation of African low-resource languages in these studies. Ojo et al. (2023); Azime et al. (2024) address a broader spectrum of African languages, aligning more closely with our research. However, their study focuses on conventional NLP tasks, such as text classification, question answering, and text generation tasks. To address the lack of difficult benchmarks, a few works automatically translated MMLU benchmarks (Lai et al., 2023a), but they do propagate errors of machine translation (MT) engines. Moreover, this is not applicable to low-resource languages with low-quality MT systems (Adelani et al., 2022a; Costa-jussà et al., 2024). Our research advances this by evaluating LLMs on more complex tasks using newly developed, human-annotated benchmarks specifically for African languages.

[1] https://github.com/masakhane-io/masakhane-nlu

[2] https://huggingface.co/collections/masakhane/irokobench-665a21b6d4714ed3f81af3b1

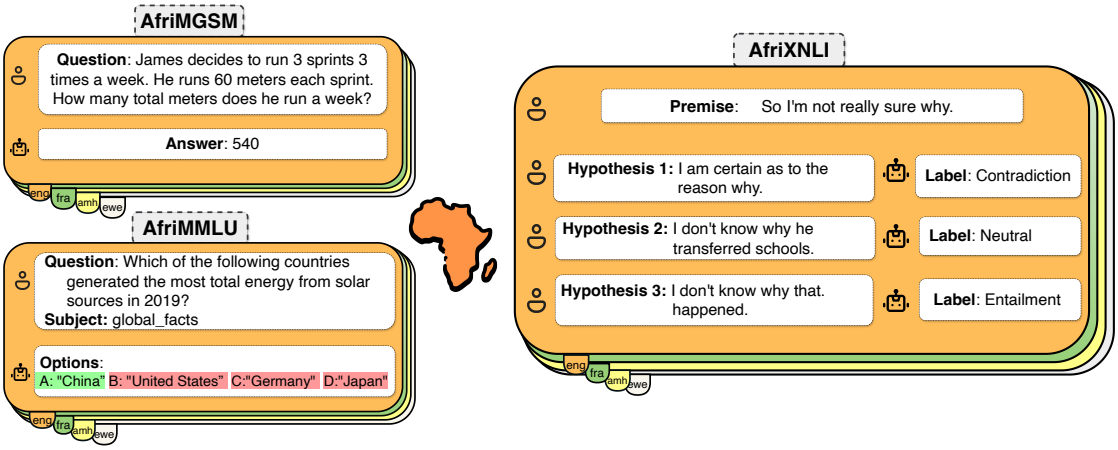


Figure 1: **Task Description for IROKOBENCH datasets.** Both AfriMGSM and AfriMMLU focus on QA, while AfriXNLI focuses on natural language inference between two pairs of sentences. For clarity, this figure provides examples in English.

African Benchmark Datasets: Due to the limited representation of African languages in the field of NLP, there has been a growing effort to create benchmark datasets for African languages to enable research on these languages. Initiatives such as Masakhane have been instrumental in the creation of standard benchmark for tasks such as machine translation (Adelani et al., 2022a), named entity recognition (Adelani et al., 2021, 2022b), part of speech tagging (Dione et al., 2023), news topic classification (Adelani et al., 2023), and sentiment analysis (Muhammad et al., 2023). There are also several multilingual benchmark datasets that cover a few African languages, such as SIB-200 (Adelani et al., 2024), Flores (Goyal et al., 2022; Costa-jussà et al., 2024), Aya dataset and Collection (Singh et al., 2024) and Taxi1500 (Ma et al., 2023). However, despite all these efforts, African languages still lack quality and more difficult datasets, our work fills this gap.

3 IROKOBENCH

3.1 Languages covered by IROKOBENCH

IROKOBENCH[3] cover 19 languages, 17 native African languages, and two European languages (English and French) which are widely spoken and official in many African countries. English and French are also the source languages we translated from. The 17 diverse and widely spoken African languages are from four regions of Africa: seven from West Africa (Ewe, Hausa, Igbo, Twi, Vai, Wolof, Yoruba), five from East Africa (Amharic, Kinyarwanda, Luganda, Swahili, and Oromo), four from Southern Africa (chiShona, isiXhosa, isiZulu, and Sesotho), and Central Africa (Lingala). These languages are from three language families in Africa: one from Mande, three from Afro-Asiatic and 13 from the Niger-Congo family—where we cover eight Bantu languages. Table 9 provides an overview of the languages covered, including their family, regions, and the number of native speakers.

3.2 Tasks covered by IROKOBENCH

The selection of these tasks is primarily driven by their coverage across various domains and downstream tasks for diverse use cases. Additionally, they enable the evaluation of logical, abstract, and reasoning capabilities in LLMs, which is the hallmark of human intelligence (Bowman et al., 2015; Hendrycks et al., 2021b). Figure 1 provides examples of the different tasks covered in our datasets. We provide their descriptions below:

AfriXNLI The task of NLI involves the classification of a pair of sentences—a premise and a hypothesis as *entailment*, *neutral*, or *contradiction* semantic relation. For example, the sentence "so I'm not really sure why" *contradicts* "I am certain as to the reason why" but has a *neutral* relation to "I don't know why he transferred schools". Here, we human translate the English portion of XNLI (a multilingual dataset comprising 15 languages, including Swahili) into the 15 African languages (excluding Swahili). While the original XNLI dataset has over 2,500/5,000 as DEV/TEST split, Each language in AfriXNLI has only 450 DEV instances and 600 test instances. We selected an equal num-

[3]Our benchmark name comes from Ìrókò— a large hardwood in West Africa that is very durable for making bench.

ber of instances from the 10 domains of XNLI. The task is evaluated using the *accuracy metric* since the dataset has balanced classes.

AfriMMLU This is a multi-choice knowledge QA curated from freely available online sources by undergraduate and graduate students in the USA. The subjects cover simple general knowledge questions like "global fact" to highly-technical questions like "professional law" and "professional medicine". MMLU are often grouped into *STEM*, *Humanities*, *Social Sciences*, and *Others* category. We focus on five subjects that we believe are culturally unbiased (or international) and that are simpler to translate since many of the subjects covered are only taught in African countries using English or French, making it extremely difficult to translate highly technical subjects, especially the STEM subjects. Table 1 shows the five subjects covered: two social science subjects (high-school geography and high-school microeconomics), one STEM subject (elementary mathematics), one humanities subject (international law), and one OTHER category (global facts). In total, we translated 608 question-answer pairs, with 500 instances in the test split, 100 questions per subject. The task is evaluated using the *option prediction accuracy*.

AfriMGSM This is a QA task with questions obtained from grade school mathematical word problems created by human problem writers. AfriMGSM expands the original MGSM dataset (Shi et al., 2023), which contains 250 QA pairs and 11 languages (including Swahili), to 15 more languages. The dataset consists of 8 training examples for few-shot and chain-of-thought prompting and 250 as a test set. We evaluated this task using the *Exact Match metric*, which is popularly used for QA tasks.

3.3 Data collection process

Translation We recruited language coordinators for each of the 16 African languages and French, and asked them to recruit professional translators to translate the sentences. The translation process took about two months, they started with XNLI, then MGSM and MMLU. Each translator received an appropriate remuneration for their work. [4] Most of the translators translated from English except for *Ewe*, *Lingala* and *Wolof* translators that translated from French since they are from the Francophone region of Africa. Additionally, we translated the MMLU dataset to French by professional translators and from French to these three languages. Many of the Francophone translators understand French and English but are more fluent in French, so they could cross-check from English if the French sentences were not clear enough.

Quality control Regarding quality control, language coordinators reviewed and corrected any poorly translated sentences. Translators received payment only after this phase to ensure the quality of translations. For additional checks, we computed COMET (Rei et al., 2020) quality estimation (QE) scores between the human translation and the original sentences based on AfriCOMET QE metric (Wang et al., 2024). In general, the distribution of the scores (range between 0 and 1) reflect that most translated sentences are between 0.7 and 1.0 for about 13 language pairs except for *Lingala*, *Twi*, and *Wolof* where the average is around 0.5. Further analysis shows that we cannot rely on these scores for those three languages since they are not covered in the pre-training of the original AfroXLMR encoder (Alabi et al., 2022) used to build the AfriCOMET QE metric. Similar findings were reported in the original AfriCOMET QE paper that Twi had worse correlation with human judgement (*i.e.,* 0.279 for Pearson, and 0.060 for Spearman) (Wang et al., 2024). We provide further analysis of the COMET scores in Appendix A.3.

3.4 LLMs used for evaluation

Open LLMs We evaluate on two *encoder-decoder open LLM*: mT0-XXL-MT (Muennighoff et al., 2023), and Aya-101 (Ustun et al., 2024) that have been instruction fine-tuned and multilingual T5 pre-trained on 101 languages (mT0 and Aya-101) models. Furthermore, these models are also all designed to be *massively multilingual* and explicitly optimized to work outside of English. The languages covered during instruction tuning differ for different models, mT0 and Aya covered 46 and 101 languages respectively.

Additionally we evaluate on eight *decoder-only open LLM* models: BLOOMZ 7B (BigScience-Workshop et al., 2023), Gemma 2 (9B & 27B) (Riviere et al., 2024), LLaMa 3 8B (Meta, 2024), LLaMa 3.1 (8B & 70B) (Dubey et al., 2024), Command-R (August version) (Cohere, 2024), and

[4] We recruited a logistic company in Kenya that managed all recruitment and payments—each country has different rates. For example, we paid $549.78 for the translation of 1020 XNLI samples in South Africa, $355.86 in Nigeria.

Dataset	No. of languages	No. instances Train/Dev/Test	Subjects / Domains	Average Length Train/Dev/Test
AfriMGSM	17 (excl. Swahili, inc. Vai)	8 / - / 250	grade school mathematics	25 / - / 46
AfriMMLU	17 (incl. French)	25 / 83 / 500	elementary mathematics, high-school geography, International law, global facts, high school microeconomics	18 / 17 / 17
AfriXNLI	15 (excl. Swahili)	- / 450 / 600	face-to-face, telephone, oxford university press (oup), fiction, travel, government, nineeleven, letters, slate, verbatim	- / 10 / 10 (hyp.) - / 18 / 17 (pre.)

Table 1: **The IROKOBENCH datasets:** dataset name, number of African languages covered, data split, and the subjects or domains covered. We included English, French, and Swahili in all benchmarks.

LLaMaX 3 8B (Lu et al., 2024). These models' weights are openly available under various licenses, ranging from fully permissive to non-commercial, research-only licenses. We evaluate the instruction-tuned variant of these models. All models were pre-trained from scratch except LLaMaX that undergo continue pre-training on 100 languages including 13 languages covered in IROKOBENCH, except Ewe, Twi, Kinyarwanda and Vai. We used LLaMAX3-8B-Alpaca, instruction-tuned on English Alpaca (Taori et al., 2023).

Closed LLM We limit our evaluation to only OpenAI GPT (3.5-0125, 4-Turbo-0125, 4o-mini-07-18, 4o-08-06) (OpenAI, 2024), Gemini-1.5-Pro (Reid et al., 2024), and Claude OPUS (Anthropic, 2024) models. Recent work has shown that proprietary models tend to exhibit better multilingual capabilities (Ahuja et al., 2023b), although specifics regarding their pre-training and instruction fine-tuning processes are not disclosed.

3.5 Evaluation Settings

Evaluation Set-up We conduct two types of evaluations: *in-language* and *translate-test* evaluation, where test instances are automatically translated into English using a machine translation (MT) engine. For MT, we use NLLB-200 (3.3B) (Costa-jussà et al., 2024). In both in-language and translate-test setups, we perform cross-lingual transfer experiments from English and zero-shot evaluations by prompting LLMs. Few-shot evaluations are performed only for the three best models (two open and one closed) in the in-language setting. For AfriMGSM in both settings, we use **Chain-of-Thought (COT)** reasoning.

We use the EleutherAI LM Evaluation Harness (`lm-eval`) tool (Biderman et al., 2024)—a popular evaluation tool that is helping to standardize LLM evaluation, especially for open models on HuggingFace Model Hub. For closed models, we employ a **verbalizer** (Gao et al., 2021; Schick and Schütze, 2021) for prediction and evaluation. All models are prompted with **five different templates**. We provide more details in Appendix A.2. [5]

Cross-lingual transfer experiments We first conduct a study on cross-lingual transfer in a supervised learning setting by fine-tuning the English training data ($400K$ instances) from Conneau et al. (2018) and evaluating on the remaining languages. This experiment focuses solely on the NLI task due to the availability of training data for supervised learning. The evaluation employs several masked language models, including XLM-R (Conneau et al., 2020), Serengeti (Adebara et al., 2023), AfroXLMR-{base, large} (Alabi et al., 2022), AfroXLMR-76L (Adelani et al., 2024). We report the result of the best model in the paper and others in Appendix A.6.

Zero- and few-shot evaluation In a zero-shot setting, we use the prompts detailed in subsection A.2. For few-shot evaluations, we conduct a 5-shot assessment for both AfriMMLU and AfriXNLI, and an 8-shot assessment for AfriMGSM.

4 Results

4.1 Overall Results

Large performance gaps between high-resource languages and African languages Table 2 shows the result of zero-shot evaluation for various LLMs. On average, there is a significant performance gap between African languages and English (up to 28%) and French (up to 19%) on the best LLM. The best LLM for African languages is GPT-4o, with an average performance of 59.0 across the evaluated tasks. Finally, as shown in Table 9, the languages with the lowest performance have the least monolingual data on the web.

Large performance gaps between closed and open weights models Our results, as presented in Table 2, indicate that the closed models Claude

[5]We make use of Cohere API for Command-R inference.

Model	size	AfriXNLI in-lang.	AfriXNLI translate test	AfriMMLU in-lang.	AfriMMLU translate test	AfriMGSM in-lang.	AfriMGSM translate test	Ave. in-lang.	Ave. translate test	Ave. English lang.	Ave. French lang.
AfroXLMR-76L	559M	**65.7**	**63.6**								
mT0-XXL-MT	13B	51.0	49.9	27.9	28.4	2.9	2.5	27.3	26.9	34.5	33.1
Aya-101	13B	51.5	50.2	29.7	31.1	4.6	7.9	28.6	29.7	39.3	35.3
BLOOMZ 7B	7B	39.4	47.6	24.1	27.9	1.7	1.9	21.7	25.8	31.4	28.9
LLaMa 3 8B	8B	35.4	38.2	28.1	31.8	3.9	33.4	22.5	34.4	52.8	45.8
LLaMa 3.1 8B	8B	36.6	43.6	31.1	41.1	7.2	30.1	24.9	38.2	57.5	53.4
LLaMaX 3 8B	8B	40.8	33.3	29.3	35.2	4.8	9.2	24.9	25.9	38.5	33.1
Gemma 2 9B	9B	40.3	43.3	35.4	44.7	19.8	39.4	31.9	42.5	64.6	58.2
Gemma 2 27B	27B	42.8	49.0	39.9	48.8	28.5	**46.1**	37.1	48.0	76.3	69.9
LLaMa 3.1 70B	70B	38.0	42.8	39.4	51.3	24.6	45.6	34.0	46.5	73.5	65.5
Command-R	35B	43.4	57.0	27.8	40.8	5.7	38.3	25.6	45.4	71.7	63.0
Claude Opus	UNK	58.1	56.4	43.0	47.6	25.3	32.7	42.3	45.6	73.3	64.0
Gemini-1.5-Pro	UNK	59.4	49.9	**60.2**	53.1	**55.4**	**44.3**	58.3	49.1	82.6	75.3
GPT-3.5-Turbo	UNK	42.1	45.5	38.1	46.8	10.6	37.1	30.2	43.1	71.9	62.9
GPT-4o-mini	UNK	54.2	56.7	45.5	50.2	35.4	42.3	45.0	49.7	84.7	76.7
GPT-4-Turbo	UNK	59.5	57.0	54.2	52.1	45.2	43.5	52.9	**50.9**	84.6	76.5
GPT-4o	UNK	64.3	52.1	60.0	**54.1**	52.6	42.6	**59.0**	49.6	**86.9**	**78.1**
								33.9	40.2	62.8	56.3

Table 2: **Main results:** Average performance of various LLMs on all tasks (ave. excl. eng, fra, and vai). Except for AfriMGSM, which uses the Exact Match metric, others use the Accuracy. The best result is in **Bold** and second best underlined. The top-2 open and closed models are in **gray**. We report only the **best prompt** (others in appendix).

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
Prompt LLMs in African Language																			
Aya-101 (t2)	40.0	36.6	31.6	25.4	33.4	36.8	30.8	27.8	28.0	26.2	28.2	31.8	32.2	26.8	25.2	32.0	28.4	29.8	29.7
Gemma 2 27B (t1)	75.6	66.4	40.6	32.4	43.2	44.2	40.2	38.2	32.6	33.6	44.6	41.8	56.0	35.6	30.4	42.0	41.0	42.2	39.9
LLaMa 3.1 70B (t1)	76.4	69.4	41.6	32.2	47.6	47.2	38.6	40.0	34.4	35.6	41.6	39.0	55.8	28.4	31.6	34.2	41.4	40.4	39.4
GPT-4o (t1)	**87.4**	**83.2**	**59.8**	**33.6**	**67.2**	**67.2**	**64.2**	**61.0**	**52.8**	**61.0**	**67.6**	**67.4**	**77.4**	**43.2**	**37.8**	**70.2**	**61.2**	**68.2**	**60.0**
Translate-Test (Eval. in English)																			
Aya-101 (t2)		37.8	32.4	28.6	31.0	31.6	31.8	33.6	27.2	28.6	32.4	34.2	31.2	31.8	26.6	30.8	34.2	32.2	31.1
Gemma 2 27B (t1)		62.4	57.6	40.4	50.0	50.0	50.6	47.0	42.8	46.6	49.8	55.6	59.8	39.8	31.2	55.2	52.6	51.6	48.8
LLaMa 3.1 70B (t1)		67.4	55.6	**44.8**	50.6	55.8	55.6	53.8	**46.8**	49.4	53.6	59.6	63.0	41.2	**32.8**	55.4	49.6	52.8	51.3
GPT-4o (t1)		**76.4**	**62.8**	43.8	**54.0**	**57.4**	**58.2**	**54.4**	46.4	**54.8**	**55.6**	**63.2**	**67.4**	**44.0**	32.0	**62.2**	**52.4**	**57.4**	**54.1**

Table 3: **AfriMMLU results in in-language and translate-test scenarios**: Option prediction accuracy per language. Average computed on only African languages. The **best prompt template** for each model is in **bracket**.

Opus, Gemini-1.5-Pro, GPT-4-Turbo, and GPT-4o consistently outperform the open models on IROKOBENCH. The top-2 closed models achieve average performance scores ranging from 52.9 to 59.0 across all tasks. The performance gap between the best closed model (GPT-4o) and the best open model (Gemma 2 27B) is 21.9. Notably, the largest performance differences for *in-language* setting are observed in the AfriMMLU and AfriMGSM tasks, where GPT-4o outperforms Gemma 2 27B by 20.1 and 24.1, respectively. For the AfriXNLI task, mT0-XXL-MT and Aya-101 perform better than bigger models like Gemma 2 27B and LLaMa 3.1 70B with 27B and 70B parameters respectively.

Majority of models perform worse for in-language prompting Most users would prefer to prompt in their native language; however, we find that almost all models we benchmark perform better with prompts translated into English. Only a few exceptions, such as GPT-4-Turbo, and GPT-4o, perform better in the in-language evaluation. Specifically, Command R and LLaMa 3.1 8B benefit the most from the translate-test approach, showing average improvements of +19.8 and +13.3, respectively. Notably, Gemma 2 27B achieves the best overall results for the AfriMGSM task with the translate-test, outperforming GPT-4o by +3.5. We attribute this boost to the fact that these models are heavily English-centric.

4.2 Task-specific results

Here, we examine the individual language performance per task, comparing which task prefers *in-language v.s. translate-test* evaluation. We compare the performance on a subset of LLMs: Aya-

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
Elementary Mathematics	88.0	83.0	76.0	55.0	73.0	84.0	77.0	71.0	80.0	81.0	75.0	78.0	87.0	68.0	61.0	76.0	80.0	80.0	75.1
Global Facts	67.0	60.0	44.0	32.0	56.0	49.0	60.0	54.0	41.0	52.0	48.0	49.0	59.0	33.0	36.0	56.0	42.0	57.0	48.0
High School Geography	93.0	86.0	56.0	25.0	67.0	64.0	57.0	55.0	48.0	54.0	69.0	68.0	76.0	34.0	27.0	63.0	53.0	73.0	55.6
High School Microeconomics	98.0	90.0	54.0	26.0	70.0	58.0	55.0	63.0	44.0	55.0	59.0	62.0	82.0	32.0	19.0	78.0	51.0	61.0	54.3
International Law	90.0	91.0	64.0	36.0	71.0	75.0	75.0	76.0	55.0	65.0	78.0	72.0	86.0	45.0	38.0	84.0	75.0	76.0	66.9
Average	87.2	82.0	58.8	34.8	67.4	66.0	64.8	63.8	53.6	61.4	65.8	65.8	78.0	42.4	36.2	71.4	60.2	69.4	60.0

Table 4: **GPT-4o AfriMMLU results by subjects**: Option prediction accuracy per language. Languages with at least 60% accuracy in four subjects are in Cyan .

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
Prompt LLMs in African Language																			
Aya-101 (t1)	10.8	9.6	6.8	3.2	7.2	3.2	3.6	4.4	2.4	4.8	6.8	10.8	6.0	1.6	1.2	4.4	3.2	3.6	4.6
Gemma 2 27B (t4)	85.6	**80.0**	33.6	7.6	49.6	24.0	32.4	18.0	23.6	12.8	35.2	38.4	73.6	12.4	5.6	32.0	22.4	34.4	28.5
LLaMa 3.1 70B (t4)	**86.8**	76.4	17.6	8.0	48.8	37.2	26.4	11.2	24.4	10.8	21.2	32.8	68.0	14.4	3.2	18.8	23.6	27.2	24.6
GPT-4o (t2)	84.0	68.8	**57.6**	**8.8**	**64.8**	**57.6**	**60.4**	**51.2**	**51.6**	**61.2**	**58.4**	**60.8**	**78.8**	**31.2**	**28.0**	**52.4**	**62.0**	**57.2**	**52.6**
Translate-Test (Eval. in English)																			
Aya-101 (t1)		8.4	8.4	6.4	7.6	6.0	10.0	6.4	6.8	7.6	6.8	10.4	9.2	6.0	8.4	8.8	8.8	8.0	7.9
Gemma 2 27B (t4)		70.8	53.2	30.0	**54.0**	**44.0**	**55.2**	47.2	34.4	**46.0**	**48.0**	54.4	69.6	**29.2**	**21.6**	48.4	**48.4**	**54.0**	**46.1**
LLaMa 3.1 70B (t4)		**73.6**	**54.8**	**30.4**	48.0	43.2	52.8	**48.0**	**35.6**	44.0	46.8	**55.2**	**72.0**	26.4	20.4	**49.6**	**48.4**	**54.0**	45.6
GPT-4o (t2)		70.0	48.4	23.6	46.8	39.2	51.6	44.8	35.2	42.0	42.4	54.8	68.0	26.4	16.0	45.6	47.2	49.2	42.6

Table 5: **AfriMGSM results in in-language and translate-test scenarios**: Exact Match score per language. Average computed on only African languages. The **best prompt template** for each model is in **bracket**.

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
Prompt LLMs in African Language																			
AfroXLMR-76L	88.2	**83.3**	**78.5**	**58.3**	**73.3**	**70.0**	**65.8**	**33.3**	**68.0**	**69.3**	**70.8**	**70.8**	**73.3**	**59.5**	**51.8**	**73.0**	**63.2**	**72.5**	**65.7**
Aya-101 (t4)	67.0	59.7	64.2	43.2	57.0	55.5	54.3	33.5	51.7	51.5	55.7	52.2	56.5	47.0	36.7	55.2	54.5	55.3	51.5
Gemma 2 27B (t2)	67.8	63.3	47.0	36.8	49.7	46.2	40.5	32.0	41.7	35.8	46.0	43.5	57.0	36.0	36.7	45.0	42.5	48.0	42.8
LLaMa 3.1 70B (t2)	57.3	50.7	43.2	34.3	42.8	42.3	36.5	32.8	37.5	34.7	35.5	38.3	44.0	36.0	34.7	39.3	39.0	37.0	38.0
GPT-4o (t3)	**89.2**	82.3	71.8	45.0	75.2	68.2	68.0	32.7	69.8	71.2	71.3	71.8	71.5	55.8	52.7	72.0	64.5	67.5	64.3
Translate-Test (Eval. in English)																			
AfroXLMR-76L		**83.0**	**73.7**	**54.3**	**67.2**	**66.0**	**63.0**	32.8	**65.7**	**65.8**	**71.2**	**70.2**	**73.0**	**56.8**	**47.5**	**74.2**	**63.7**	**72.0**	**63.6**
Aya-101 (t4)		61.2	60.7	40.8	53.8	52.3	50.3	**33.0**	48.7	51.3	54.0	56.0	54.0	44.5	39.3	56.8	51.0	57.0	50.2
Gemma 2 27B (t2)		60.8	55.8	45.5	48.5	49.7	47.7	31.5	51.3	52.3	51.7	50.2	55.7	47.2	40.2	54.8	50.2	51.5	49.0
LLaMa 3.1 70B (t2)		58.8	42.8	39.3	45.2	39.8	36.2	32.7	44.3	46.0	51.8	50.8	52.5	38.8	34.5	44.8	41.8	42.8	42.8
GPT-4o (t3)		73.8	63.8	42.8	54.8	53.3	52.8	32.3	54.3	55.5	58.2	58.0	58.5	45.7	38.8	58.2	53.5	52.8	52.1

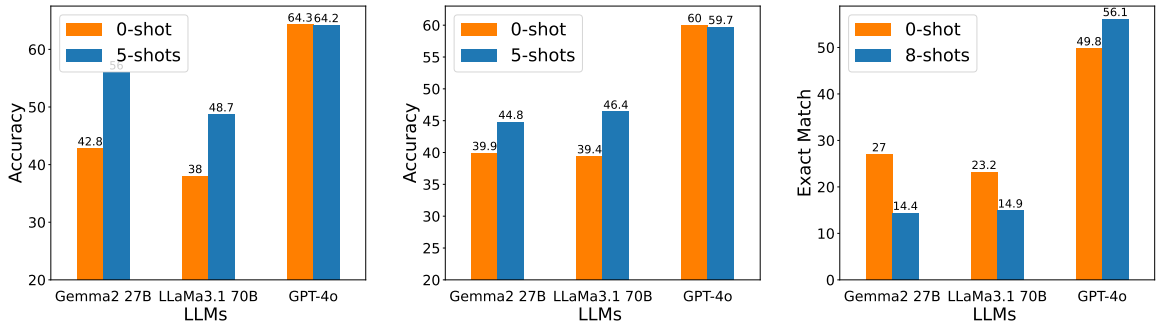
Table 6: **AfriXNLI results in in-language and translate-test scenarios**: Option prediction accuracy per language. Average computed on only African languages. The **best prompt template** for each model is in **bracket**.

101, Gemma 2 27B LLaMa 3.1 70B, and GPT-4o. Other LLMs are in Appendix A.4.

AfriMMLU evaluation is better when prompting in-language for closed LLMs Table 3 shows the result of different LLMs on AfriMMLU using *in-language* vs. *translate-test*. For GPT-4o, we found *in-language* prompting to be generally better. However, for open models like LLaMa 3.1 70B and Gemma 2 27B, we find a large improvement in the performance of the *translate-test* for 15 out of 16 African languages. The only language that did not improve is wol, probably due to poor MT performance on NLLB-200, as reported in Costa-jussà et al. (2024). This shows the benefit of translate-test for prompting English-centric LLMs when evaluating low-resource languages.

AfriMMLU performance by subjects Table 4 shows the result of GPT-4o by subjects. Interestingly, *elementary math* achieved the best overall accuracy, where 13 out of 16 African languages achieved at least 70% despite struggling with AfriMGSM. The performance difference to AfriMGSM may be due to the multi-choice options of MMLU, which may be slightly easier than free-form answer. *International law* also achieved impressive performance with 10 out of 16 languages achieving 70%. All languages struggle the most with *Global facts* including English and French. Similarly, many African languages find it difficult to answer questions in *geography* and *microeconomics* subjects. This presents an opportunity for improving LLMs for the education domain.

AfriMGSM performance receives significant boost with translate-test for open models Table 5 shows that we can achieve a significant boost in performance with the translate-test on all lan-



(a) AfriXNLI (b) AfriMMLU (c) AfriMGSM

Figure 2: **Few-shot evaluation** on IROKOBENCH, we performed 8-shots for AfriMGSM and 5-shots for the others.

	AfriXNLI						AfriMMLU						AfriMGSM					
Model	**t1**	**t2**	**t3**	**t4**	**t5**	**Ave.**	**t1**	**t2**	**t3**	**t4**	**t5**	**Ave.**	**t1**	**t2**	**t3**	**t4**	**t5**	**Ave.**
Aya-101	47.7	49.9	45.7	**51.5**	48.3	$48.6_{\pm 2.2}$	29.6	**29.7**	**29.7**	**29.7**	29.6	$29.7_{\pm 0.1}$	**4.4**	4.1	4.2	**4.4**	4.3	$4.3_{\pm 0.1}$
Gemma 2 27B	38.3	**40.3**	33.4	34.1	35.4	$36.3_{\pm 2.9}$	**39.9**	39.2	38.7	39.2	39.1	$39.2_{\pm 0.4}$	22.3	26.3	25.0	**27.0**	26.2	$25.4_{\pm 1.9}$
GPT-4o	55.3	49.6	**64.3**	62.9	59.7	$58.4_{\pm 6.0}$	**60.0**	55.8	59.0	39.6	59.6	$54.8_{\pm 8.7}$	46.7	**49.8**	48.0	49.5	49.6	$48.7_{\pm 1.3}$

Table 7: **Ablation results:** Effect of using different prompts (t1, t2, t3, t4, and t5) for IROKOBENCH datasets. Best prompt results are in **bold**. Average computed on only African languages. Second best prompt are underlined.

guages we evaluated. We hypothesize that the current LLMs are better at reasoning in English than other languages. Gemma 2 27B and LLaMa 3.1 70B improved by +17.6 and +21.0 respectively when questions are asked in English, while closed models like GPT-4o dropped in performance.

4.3 Few-shot results, Cross-Lingual Transfer, Sensitivity to Prompt Templates

Cross-lingual transfer in-language achieves better results When there is large enough labeled data in English, we could leverage this cross-lingual signal for zero-shot evaluation. We trained on 400k English NLI examples, and we performed zero-shot transfer in in-language and translate-test setting. Table 6 shows that cross-lingual transfer using an Africa-centric smaller language model (AfroXLMR-76L) gave better results than prompting LLMs on average. AfroXLMR-76L has been pre-trained on all languages in AfriXNLI, which explains the impressive performance. However, for multilingual encoders that have not seen some of the languages, prompting GPT-4o seems to be better, as shown in Appendix A.6.

Impact of few-shot vs zero-shot Figure 2 shows the few-shot results for the IROKOBENCH datasets leveraging Gemma 2 27B, LLaMa 3.1 70B and GPT-4o when we provide few examples in in-language setting. We found out that Gemma 2 27B and LLaMa 3.1 70B LLM consistently benefited the least from additional few shots examples for classification tasks (AfriXNLI and AfriMMLU) where Gemma 2 27B improved by +13.2 and +4.9 respectively. Similarly, LLaMa 3.1 70B improved on both AfriXNLI and AfriMMLU tasks with +10.7 and +7.0. However, for reasoning tasks, only GPT-4o improved in performance by 6.3, other LLMs dropped in performance, probably due to their inability to reason in non-English languages. Surprisingly, GPT-4o did not benefit from additional examples for the classification tasks.

Prompt t2

```
{{premise}}

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

Prompt t3

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

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

Prompt t4

```
You are an expert in Natural Language Inference (NLI)
specializing in the {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}
```

Table 8: Three different prompts preferred by different models for AfriXNLI

Sensitivity to prompt templates To understand whether sensitivity to prompts impacts results, we perform an ablation for all the IROKOBENCH tasks where we evaluate the performance of five different prompts (see subsection A.2). Table 7 shows the results of five prompts we tested for (three of the prompts most preferred by different models are

shown in Table 8). On AfriXNLI, we find that simpler prompt (Nie et al., 2020) i.e. `{{premise}} Question: {{hypothesis}} True, False, or Neither?`, have better results for the open models like Aya-101 and Gemma 2 27B, while GPT-4o prefers `t3` where a detailed task description is provided. The best prompt for Aya-101 is `t4` where the `{{language}}` name is mentioned, which shows additional language information may be useful in improving performance.

On AfriMMLU, we find GPT-4o perform worse for `t4`.[6]. However, other models are not very sensitive to the use of different prompts. In general, we do not find AfriMGSM to be sensitive to different prompts. In subsection A.4, we provide the results of five prompt templates for all LLMs evaluated.

5 Conclusion

In this paper, we introduced IROKOBENCH, a new benchmark for evaluating large language models (LLMs) on African languages. IROKOBENCH comprises three datasets focused on different tasks: natural language inference (AfriXNLI), multi-choice knowledge QA (AfriMMLU), and mathematical reasoning (AfriMGSM). Unlike previous benchmarks, which primarily involve simple text classification tasks, these datasets assessed the LLMs' abilities in complex and knowledge-intensive areas. Our evaluation revealed a significant performance gap between high-resource languages (*e.g.,* English and French) and African languages. Additionally, we observed a substantial disparity in performance between open models and proprietary models, with the latter generally outperforming the former, particularly in mathematical reasoning tasks. We hope that IROKOBENCH will serve as a valuable benchmark for evaluating future LLMs developed or adapted for African languages.

Limitations Our benchmark has a few limitations: (1) **The benchmark is human-translated** which may include some translationese effects, it would have been better if they are all generated in the native African languages. However, this parallel translation allows us to evaluate and compare the same sentences in all these languages. (2) **We only cover three language families in Africa**, Nilo-Saharan, Austronesian, and Khoisan language groups are missing, one of the reason we excluded them is either lack of contact with professional translators or limited translation budget, we hope to extend to more languages in the future.

[6]Analyze each question critically and determine the most correct option based on your understanding of the subject matter " Question: {question}. Choices A: {choice1}, B:{choice2}, C: {choice3}, D: {choice4}

Acknowledgment

This work was supported by Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. Additional support was provided through compute credits from Oracle and the Cohere For AI Research Grant. We extend our gratitude to Charles Riley for facilitating our connection with the Vai translator for the MGSM data translation.

We are also deeply thankful to OpenAI for granting API credits through their Researcher Access API program to Masakhane, enabling the evaluation of GPT-3.5 and GPT-4 LLMs. Similarly, we appreciate Google for providing GCP credits via the Gemma 2 Academic Program, which supported the Gemini-1.5-Pro inference. Lastly, we would like to thank Hailey Schoelkopf and Lintang Sutawika for their invaluable assistance with the EleutherAI *lm-eval* tool.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. 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, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 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, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 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, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

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, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *ArXiv*, abs/2311.07463.

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.

Anthropic. 2024. Claude — anthropic.com. `https://www.anthropic.com/claude`. [Accessed 01-06-2024].

Anuoluwapo Aremu, Jesujoba O. Alabi, and David Ifeoluwa Adelani. 2023. Yorc: Yoruba reading comprehension dataset. *Preprint*, arXiv:2308.09768.

Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Enhancing amharic-llama: Integrating task specific and generative datasets. *arXiv preprint arXiv:2402.08015*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

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

BigScience-Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Cohere. 2024. Command R — cohere.com. `https://cohere.com/command`. [Accessed 01-06-2024].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, 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, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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 Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, and et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Gemini-Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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.

Kai Hartung, Aaricia Herygers, Shubham Kurlekar, Khabbab Zakaria, Taylan Volkan, Sören Gröttrup, and Munir Georges. 2023. Measuring sentiment bias in machine translation. *Preprint*, arXiv:2306.07152.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *ArXiv*, abs/2302.09210.

Meng Ji, Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. *Cultural and Linguistic Bias of Neural Machine Translation Technology*, page 100–128. Studies in Natural Language Processing. Cambridge University Press.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400:

A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023a. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023b. 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.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *Preprint*, arXiv:2305.08487.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. https://ai.meta.com/blog/meta-llama-3/. [Accessed 01-06-2024].

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

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

OpenAI. 2024. Introducing ChatGPT. https://openai.com/index/chatgpt/. [Accessed 01-06-2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Danti Pudjiati, Ninuk Lustyantie, Ifan Iskandar, and Tira Nur Fitria. 2022. Post-editing of machine translation: Creating a better translation of cultural specific terms. *Language Circle: Journal of Language and Literature*, 17(1):61–73.

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, 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, Sammy Jerome, and et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Preprint*, arXiv:2104.06001.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

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. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

A. Ustun, 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. *ArXiv*, abs/2402.07827.

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.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Jiayi Wang, David 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 Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, 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.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in

neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Appendix

A.1 Language covered

Table 9 provide the languages covered in the IROKOBENCH, their language family, regions located in Africa, number of speakers, and size of monolingual data available on the web based on the MADLAD cleaned corpus (Kudugunta et al., 2023)—we only report number of characters in mega bytes. Additionally, we added an indication whether this language is covered in the pre-training of Aya-101 and BLOOMZ 7B LLMs.

A.2 Prompts Template and Evaluation Tool

We use the EleutherAI LM Evaluation Harness (`lm-eval`) tool (Biderman et al., 2024)—a popular evaluation tool that is helping to standardize LLM evaluation. The tool allows for three types of evaluation: *log-likelihood*, *perplexity*, and *generation*. The log-likelihood is more suitable for multiple-choice tasks since it helps to restrict the model's option to fewer choices—more appropriate for weaker models. However, the log-likelihood approach cannot evaluate the generative capabilities of LLMs to generate coherent and relevant answers. Moreover, closed models are only accessible via API and do not provide access to the log probabilities, making it impossible to use the log-likelihood approach. To extract the correct answers for the task, we employed a **verbalizer** (Gao et al., 2021; Schick and Schütze, 2021). For AfriMGSM, we used the default verbalizer provided by the tool. However, for AfriXNLI and AfriMMLU, we manually created a verbalizer for the closed models and used the *log-likelihood* request type for the open models.

The prompt templates used for evaluation of different tasks are in Table 11, Table 12 and Table 10.

A.3 AfriCOMET metric scores for XNLI translation

We employ AfriCOMET evaluation metrics, as developed by Wang et al. (2024), to automatically assess the quality of translations for our newly created benchmarks. Figure 3 depicts the histogram of scores obtained from AfriCOMET for AfriXNLI, illustrating promising results and offering compelling evidence for the effectiveness of our translations (Amharic, Yorùbá, isiZulu). However, the performance of this metric depends on if the language we are evaluating is covered in the pre-training of the base model of the metric i.e. AfroXLMR-large. In the case of Lingala, Twi and Wolof, the performance of the metric does not correlate with the human translation since they are not covered in AfroXLMR. Similar findings were reported in the original AfriCOMET QE paper that Twi had worse correlation with human judgement (*i.e.,* 0.279 for Pearson, and 0.060 for Spearman) (Wang et al., 2024).

A.4 Task-specific results for all models

We provide the entire results of all LLMs and all prompts on AfriMMLU, AfriXNLI and AfriMGSM tasks are shown in Table 14, Table 13 and Table 15. We performed evaluation on 5 prompts for all models except Claude Opus which we limit to one prompt due to API inference cost.

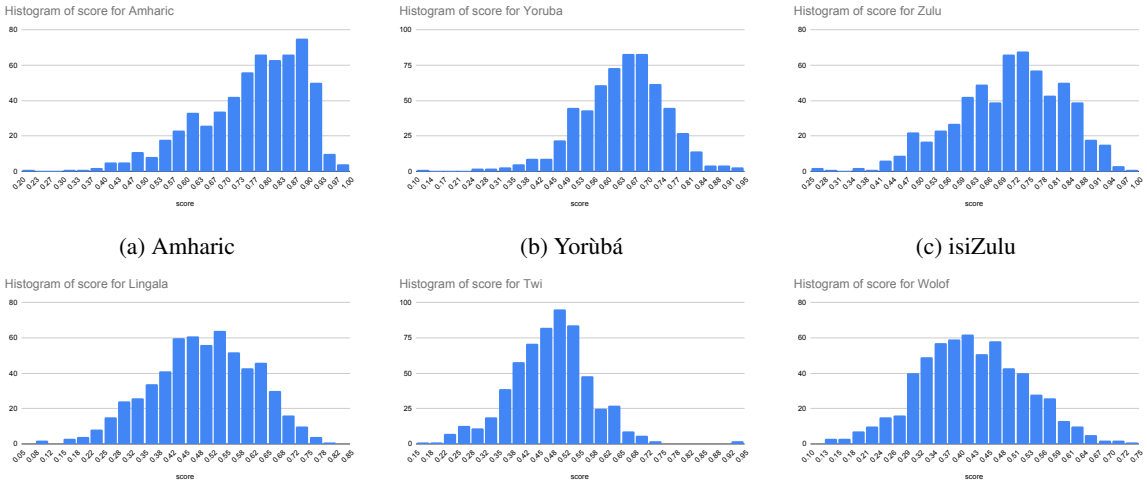
A.5 Comparison between in-language and translate-test results

We provide the results comparing the in-language and translate-test results on all LLMs in Table 16, Table 17, and Table 18.

A.6 Cross-lingual transfer results for XNLI

In Table 19, we compare different multilingual masked language model (MLM) performance on African languages. XLM-R-large has 559M parameters and is trained on 100 languages, but only a few African languages are covered (`amh`, `hau`, `orm`, `swa`, and `xho`). Serengeti, on the other hand, has been pre-trained on all languages in IROKOBENCH, but it only has 240M parameters. AfroXLMR was adapted from XLM-R through continual pre-training on 17 African languages including 11 in IROKOBENCH (`amh`, `hau`, `ibo`, `kin`, `orm`, `sna`, `sot`, `swa`, `xho`, `yor`, and `zul`). AfroXLMR-76L follows the same technique by performing continual pre-training on XLM-R-large on 76 languages (72 African), all languages covered in IROKOBENCH are part of its pre-training.

We found Africa-centric MLM to perform better on average than massively multilingual models like XLM-R-large. Serengeti and AfroXLMR-



(a) Amharic (b) Yorùbá (c) isiZulu

(d) Lingala (e) Twi (f) Wolof

Figure 3: Evaluation of **AfriXNLI** translations using AfriCOMET metric scores.

base improved over larger-sized XLM-R-large by $+3.5$ and $+5.3$ points, respectively. Similarly, fine-tuning AfroXLMR-large, a larger version of AfroXLMR-base, results in an improved boost in performance with 11.8 points. The best overall results were achieved by AfroXLMR-76L with a 16.1 boost in performance over XLM-R-large. This is probably because all the languages are used in pretraining. We make use of AfroXLMR-76L has the baseline for all LLMs. Interestingly, we find GPT-4o to be competitive or better than other MLMs except the AfroXLMR-76L on average.

Language	Family/branch	Region	# speakers	# chars in MADLAD (MB)	In Aya	In BLOOMZ
English (eng)	Indo-European / Germanic	Across Africa	1457M	9,000,000MB	✓	✓
French (fra)	Indo-European /Romance	Across Africa	310M	1,000,000MB	✓	✓
Kiswahili (swa)	Niger-Congo / Bantu	East & Central Africa	71M-106M	2,400MB	✓	✓
Kinyarwanda (kin)	Niger-Congo / Bantu	East Africa	10M	749MB	✓	✓
Hausa (hau)	Afro-Asiatic / Chadic	West Africa	77M	630MB	✓	✘
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	East Africa	57M	509MB	✓	✘
isiXhosa (xho)	Niger-Congo / Bantu	Southern Africa	19M	287MB	✓	✓
chiShona (sna)	Niger-Congo / Bantu	Southern Africa	11M	266MB	✓	✓
isiZulu (xho)	Niger-Congo / Bantu	Southern Africa	27M	257MB	✓	✓
Igbo (ibo)	Niger-Congo / Volta-Niger	West Africa	31M	251MB	✓	✓
Yorùbá (yor)	Niger-Congo / Volta-Niger	West Africa	46M	239MB	✓	✓
Sesotho (sot)	Niger-Congo / Bantu	Southern Africa	13M	227MB	✓	✓
Oromo (orm)	Afro-Asiatic / Cushitic	East Africa	37M	88MB	✘	✘
Luganda (lug)	Niger-Congo / Bantu	Central Africa	11M	48MB	✘	✓
Ewe (ewe)	Niger-Congo / Kwa	West Africa	7M	33MB	✘	✓
Twi (twi)	Niger-Congo / Kwa	West Africa	9M	25MB	✓	✓
Lingala (lin)	Niger-Congo / Bantu	Central Africa	40M	22MB	✘	✓
Wolof (wol)	Niger-Congo / Senegambia	West Africa	5M	5MB	✘	✓
Vai (vai)	Mande	West Africa	140,000	-	✘	✘

Table 9: **Languages covered in IROKOBENCH**: including language family, region, number of L1 & L2 speakers, size of monolingual data on the web (in MADLAD corpus)

Prompt1
`Question: {question}`
`Answer:`

Prompt2
`Give direct numerical answers for the question provided.`
`Question: {question}`
`Answer:`

Prompt3
`Solve the following math question.`
`Question: {question}`
`Answer:`

Prompt4
`Answer the given question with the appropriate numerical value, ensuring that the response is clear and without any supplementary information.`
`Question: {question}`
`Answer:`

Prompt5
`For mathematical questions provided in language language. Supply the accurate numeric answer to the provided question`
`Question: {question}`
`Answer:`

Table 10: Five different prompt used for prompt sensitivity experiments in AfriMGSM

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?
Prompt 2 {{premise}} Question: {hypothesis} True, False, or Neither? Answer:
Prompt 3 Given the following premise and hypothesis in English, identify if the premise entails, contradicts, or is neutral towards the hypothesis. Please respond with exact 'entailment', 'contradiction', or 'neutral'. Premise: {premise} Hypothesis: {hypothesis}
Prompt 4 You are an expert in Natural Language Inference (NLI) specializing in the {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}
Prompt 5 Based on the given statement, is the following claim 'true', 'false', or 'inconclusive'. Premise: {premise} Hypothesis: {hypothesis}

Table 11: Five different prompt used for prompt sensitivity experiments in AfriXNLI

Prompt 1

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

Question: {question}
 A: {choice1}
 B: {choice2}
 C: {choice3}
 D: {choice4}
Answer:
```

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}
 A: {choice1}
 B: {choice2}
 C: {choice3}
 D: {choice4}
Answer:
```

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}
 A: {choice1}
 B: {choice2}
 C: {choice3}
 D: {choice4}
Answer:
```

Prompt 4

```
Analyze each question critically and determine the most correct option based on your understanding of the subject matter

 A: {choice1}
 B: {choice2}
 C: {choice3}
 D: {choice4}
Answer:
```

Prompt 5

```
Given your proficiency in {subject}, please answer the subsequent multiple-choice question with 'A', 'B', 'C', or 'D'.

Question: {question}
 A: {choice1}
 B: {choice2}
 C: {choice3}
 D: {choice4}
Answer:
```

Table 12: Five different prompt used for prompt sensitivity experiments in AfriMMLU

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
AfroXLMR-R-76L	88.2	83.3	78.5	58.3	73.3	70.0	65.8	33.3	68.0	69.3	70.8	70.8	73.3	59.5	51.8	73.0	63.2	72.5	65.7
mT0-XXL-MT																			
t1	61.2	58.5	52.0	35.0	49.2	50.7	43.2	34.8	44.5	36.3	46.7	51.3	51.8	42.5	34.2	51	42.2	52.8	46.5
t2	63.5	61.5	58.3	38.5	57.5	56.5	51.3	33.5	54.2	47.2	54.8	56.0	55.5	48.8	39.7	56.3	52.3	55.0	51.0
t3	42.3	39.5	36.5	33.3	33.3	33.5	33.5	33.3	33.5	33.3	33.5	33.5	35.5	33.5	33.2	33.3	33.7	34.0	33.8
t4	34.0	34.5	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.5	33.5	33.5	33.3	33.5	33.3	33.3	33.4
t5	51.2	50.3	45.7	39.7	48.7	46.7	45.3	33.8	45.0	40.7	47.0	46.2	46.5	43.3	39.2	46.5	44.5	47.7	44.1
ave	50.4	48.9	45.2	36.0	44.4	44.1	41.3	33.8	42.1	38.2	43.1	44.1	44.6	40.3	35.9	44.1	41.2	44.6	41.4
Aya-101																			
t1	59.8	55.2	54.2	42.5	51.3	52.2	47.2	34.0	47.8	47.8	49.5	50.8	52.0	47.0	37.0	51.0	48.3	49.8	47.7
t2	61.5	60.5	58.2	42.2	56.8	54.7	48.7	34.5	44.5	51.7	55.8	54.7	55.2	48.2	37.3	53.7	48.2	54.3	49.9
t3	56.7	54.0	51.2	38.7	49.7	47.3	46.5	33.3	47.0	45.2	48.2	49.0	48.7	43.8	35.3	49.0	48.3	50	45.7
t4	67.0	59.7	64.2	43.2	57.0	55.5	54.3	33.5	51.7	51.5	55.7	52.2	56.5	47.0	36.7	55.2	54.5	55.3	51.5
t5	51.7	53.7	51.2	43.3	48.0	48.2	47.2	33.8	48.5	49.8	49.8	48.0	50.0	46.5	35.8	48.7	49.5	49.5	46.7
ave	59.3	56.6	55.8	42.0	52.6	51.6	48.8	33.8	47.9	49.2	51.8	50.9	52.5	46.5	36.4	51.5	49.8	51.8	48.3
BLOOMZ 7B																			
t1	54.8	51.0	36.2	35.3	36.7	39.2	39.3	32.2	38.8	35.8	42.5	40.0	43.2	37.7	35.0	40.0	42.0	39.7	38.3
t2	60.3	56.0	36.8	35.7	36.5	44.7	38.5	33.8	41.5	35.2	43.3	40.5	45.8	37.5	36.0	39.8	45.2	40.3	39.4
t3	38.3	39.0	33.3	32.8	34.2	33.8	32.2	32.5	35.7	33.5	34.2	31.8	36.3	34	33.3	35.3	39.7	34.0	34.2
t4	33.5	36.2	32.3	34.0	34.2	32.8	35.2	33.3	33.2	32.8	32.8	33.8	34.0	36.7	33.7	35.3	34.5	34.0	33.9
t5	36.2	38.7	35.8	33.3	35.0	36.0	35.3	33.0	36.0	34.7	36.7	35.7	37.3	36.0	34.0	36.7	36.2	36.0	35.5
ave	44.6	44.2	34.9	34.2	35.3	37.3	36.1	33.0	37.0	34.4	37.9	36.4	39.3	36.4	34.4	37.4	39.5	36.8	36.3
LLaMa 3 8B																			
t1	34.3	34.9	32.8	34.9	33.3	32.1	34.0	33.2	33.8	32.4	33.3	34.8	33.1	33.5	32.7	33.9	34.1	35.0	33.6
t2	43.9	47.4	40.2	34.3	35.6	39.1	32.9	32.7	33.3	35.4	35.5	35.9	34.1	34.3	33.3	36.8	35.17	38.5	35.4
t3	33.3	33.2	33.75	32.2	33.2	33.0	32.8	33.9	33.2	33.7	33.9	33.4	32.8	32.3	32.7	33.5	33.3	33.4	33.2
t4	36.2	35.0	32.4	34.3	33.3	33.0	33.3	34.1	33.2	33.5	32.3	33.3	32.9	31.3	32.9	33.1	33.3	33.2	33.1
t5	33.3	33.8	33.1	33.2	33.8	33.3	35.0	30.6	33.3	33.8	34.0	33.2	32.8	34.8	33.0	34.2	33.3	32.9	33.4
ave	36.2	36.9	34.5	33.8	33.8	34.1	33.6	32.9	33.4	33.8	33.8	34.1	33.1	33.2	32.9	34.3	33.8	34.6	33.7
LLaMa 3.1 8B																			
t1	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
t2	53.0	50.2	38.8	37.7	36.5	38.5	38.3	32.3	37.5	35.0	35.8	34.8	43.8	37.0	32.3	33.8	38.5	34.0	36.6
t3	33.7	34.5	35.8	33.7	35.7	34.0	35.2	35.0	33.3	33.5	34.8	34.7	34.0	32.8	35.8	33.8	34.0	34.7	34.4
t4	38.2	35.7	34.5	34.3	41.7	35.3	33.2	35.2	33.2	33.5	36.3	33.8	34.8	31.2	34.8	34.0	35.7	33.3	34.7
t5	36.2	36.3	32.3	32.8	34.0	33.5	35.3	31.2	33.5	30.2	34.0	33.2	34.3	32.5	31.5	32.0	33.3	30.3	32.7
ave	38.9	38.0	35.0	34.4	36.2	34.9	35.1	33.4	34.2	33.1	34.9	34.0	36.1	33.4	33.6	33.4	35.0	33.1	34.3
LLaMaX 3 8B																			
t1	53.0	49.3	40.2	35.8	44.8	42.5	37.7	29.8	42.7	40.5	47.0	39.7	45.7	36.3	37.3	46.2	44.0	42.0	40.8
t2	35.5	38.7	35.2	34.2	34.0	32.3	32.0	35.3	33.0	33.0	33.8	33.2	33.8	33.8	32.2	33.0	33.0	32.8	33.4
t3	34.5	33.7	34.2	34.5	33.0	32.7	35.2	34.7	33.2	33.2	33.2	33.2	34.0	35.2	33.2	35.8	33.7	33.0	33.9
t4	33.7	33.2	33.7	33.7	33.3	33.3	36.7	34.2	33.0	33.5	34.3	33.2	34.5	33.0	33.7	34.7	33.0	32.8	33.8
t5	33.3	33.7	33.3	33.0	32.8	33.5	34.8	30.7	33.2	33.2	33.7	33.3	33.8	34.8	33.0	33.5	33.3	34.5	33.4
ave	38.0	37.7	35.3	34.2	35.6	34.9	35.3	32.9	35.0	34.7	36.4	34.5	36.4	34.6	33.9	36.6	35.4	35.0	35.0
Gemma 2 9B																			
t1	70.67	62.5	45.2	34.2	39.7	38.2	38.3	33.2	36.3	35.0	39.7	39.2	46.0	36.5	33.7	39.3	39.5	39.2	38.3
t2	55.3	50.7	43.2	35.3	47.0	40.7	40.2	32.8	38.8	37.8	42.3	40.2	46.3	37.2	35.2	42.5	41.3	44.5	40.3
t3	34.5	33.7	29.5	32.5	40.2	34.3	36.3	34.8	31.3	31.3	34.3	30.8	34.7	32.7	33.7	35.7	31.3	31.3	33.4
t4	33.2	34.5	31.5	32.3	39.0	35.5	35.7	34.8	30.8	32.8	35.8	31.3	36.8	33.3	35.0	33.8	35.2	31.5	34.1
t5	37.0	39.7	34.5	33.2	36.8	35.7	33.7	35.0	35.2	35.0	36.8	38.7	35.5	33.0	34.3	38.0	35.2	36.5	35.4
ave	46.1	44.2	36.8	33.5	40.5	36.9	36.8	34.1	34.5	34.4	37.8	36.0	39.9	34.5	34.4	37.9	36.5	36.6	36.3
Gemma 2 27B																			
t1	61.5	55	43.3	33.5	41.33	37.2	35.7	33.2	34.7	33.5	38.2	35.8	45.0	34.5	33.5	35.2	35.7	36.8	36.7
t2	67.8	63.3	47.0	36.8	49.7	46.2	40.5	32.0	41.7	35.8	46.0	43.5	57.0	36.0	36.7	45.0	42.5	48.0	42.8
t3	43.2	46.0	32.0	31.7	36.2	35.0	35.0	33.3	34.7	33.2	37.0	37.7	41.5	33.2	32.3	37.3	34.0	33.7	34.9
t4	62	57.0	35.0	31.8	43.0	39.7	37.0	33.0	37.8	32.2	40.8	37.2	49.5	34.5	33.8	38.0	37.8	37.3	37.4
t5	41.3	41.7	33.8	33.5	35.5	37.7	33.5	34.5	35.7	33.8	34.7	35.5	34.2	33.7	33.7	38.8	34.3	35.2	34.9
ave	55.2	52.6	38.2	33.5	41.1	39.1	36.3	33.2	36.9	33.7	39.3	37.9	45.4	34.4	34.0	38.9	36.9	38.2	37.3
LLaMa 3.1 70B																			
t1	57.3	33.7	33.3	33.3	33.7	33.7	33.7	34.3	33.8	33.3	34.0	33.5	36.8	33.3	33.3	33.5	33.5	33.3	33.8
t2	57.3	50.7	43.2	34.3	42.8	42.3	36.5	32.8	37.5	34.7	35.5	38.3	44.0	36.0	34.7	39.3	39.0	37.0	38.0
t3	59.5	65.2	33.5	36.3	36.7	36.2	34.8	33.5	34.2	31.3	31.5	31.5	35.3	37.3	34.3	34.0	35.0	34.0	34.3
t4	37.3	57.5	33.2	34.7	45.8	36.3	35.2	32.5	38.0	32.8	39.7	39.0	49.2	33.8	34.7	33.5	35.8	37.0	36.9
t5	43.3	42.0	33.8	32.8	34.3	33.5	32.7	33.0	34.0	31.2	31.7	32.5	37.2	32.0	33.8	33.2	31.5	32.8	33.1
ave	51.0	49.8	35.4	34.3	38.7	36.4	34.6	33.2	35.5	32.7	34.5	35.0	40.5	34.5	34.2	34.7	35.0	34.8	35.2
CommandR (Aug)																			
t1	76.1	67.2	39.6	32.9	36.7	35.6	41.4	34.2	35.7	42.3	41.4	41.2	52.9	34.3	34.6	38.0	40.1	37.7	38.7
t2	57.7	57.9	36.5	34.0	37.8	39.2	35.9	30.7	37.6	36.3	37.4	34.6	42.0	35.3	32.0	37.3	43.1	35.3	36.6
t3	79.9	75.7	45.7	38.9	45.1	42.4	45.6	34.5	39.0	44.8	48.9	45.1	55.3	41.8	34.5	45.9	45.9	40.8	43.4
t4	75.4	63.4	46.9	38.9	43.0	42.1	43.0	36.2	42.1	43.1	45.3	39.1	49.7	36.4	33.8	44.0	39.9	42.8	41.6
t5	63.3	72.2	44.7	35.9	40.3	40.4	36.7	33.8	39.3	37.0	37.6	38.7	57.1	33.1	35.6	46.3	43.8	36.6	39.8
ave	70.5	67.3	42.7	36.1	40.6	39.9	40.5	33.9	38.7	40.7	42.1	39.7	51.4	36.2	34.1	42.3	42.6	38.6	40.0
gpt-3.5-turbo-0125																			
t1	65.8	62.7	36.5	33.3	36.2	39.7	39.8	29.5	41.0	39.5	39.8	39.3	49.8	39.5	34.5	42.0	41.2	40.5	38.9
t2	59.5	56.5	25.7	32.0	34.7	36.5	34.3	30.3	33.5	33.5	35.8	33.7	45.7	34.5	31.7	35.0	36.2	33.8	34.2
t3	64.2	54.5	35.8	38.3	34.8	37.0	38.2	32.3	35.3	36.7	37.3	35.8	43.0	36.2	36.8	37.7	35.7	36.2	36.7
t4	70.0	65.2	38.7	38.2	39.0	42.3	42.2	32.0	43.0	42.5	45.8	43.0	58.8	39.5	39.5	44.7	40.5	43.8	42.1
t5	69.8	65.8	39.8	36.7	44.7	42.0	43.7	32.7	43.2	39.2	44.0	45.2	50.7	39.8	35.0	44.0	44.5	42.0	41.7

ave	65.9	60.9	35.3	35.7	37.9	39.5	39.6	31.4	39.2	38.3	40.6	39.4	49.6	37.9	35.5	40.7	39.6	39.3	38.7
gpt-4o-mini-2024-07-18																			
t1	58.5	57.0	41.3	34.8	48.5	45.8	46.5	31.5	40.7	43.3	46.3	43.3	45.5	42.0	37.5	50.0	48.7	51.0	43.6
t2	66.5	64.2	44.0	33.3	52.5	48.0	49.3	33.0	41.5	43.3	49.0	44.8	56.3	40.8	34.8	49.5	49.0	49.2	44.9
t3	86.2	80.3	52.2	37.2	64.8	57.2	56.8	31.5	51.7	61.3	63.0	56.2	63.7	47.5	43.2	64.5	54.5	62.0	54.2
t4	58.5	57.5	55.7	35.3	52.0	54.5	56.8	31.0	53.5	61.0	57.5	56.3	52.3	51.7	45.2	53.5	52.3	54.8	51.5
t5	77.3	70.3	52.2	34.5	56.8	58.2	55.2	33.7	47.0	54.0	56.7	54.0	59.2	48.0	39.5	54.8	53.0	58.0	50.9
ave	69.4	65.9	49.1	35.0	54.9	52.7	52.9	32.1	46.9	52.6	54.5	50.9	55.4	46.0	40.0	54.5	51.5	55.0	49.0
gpt-4-turbo-2024-04-09																			
t1	64.2	69.3	60.3	35.2	62.8	59.8	56.7	29.7	59.0	58.5	63.2	58.3	60.2	49.7	37.3	64.7	58.5	64.5	54.9
t2	64.0	60.5	47.0	34.2	53.5	51.3	52.0	33.7	43.5	44.3	53.5	46.7	53.8	35.0	34.7	50.7	47.2	47.8	45.6
t3	87.3	81.0	62.0	38.2	70.3	66.0	65.8	33.7	67.5	61.0	71.7	67.8	67.8	41.7	38.0	71.5	62.0	66.7	59.5
t4	68.5	66.0	57.2	39.7	60.8	57.2	56.2	34.0	57.8	50.7	59.8	56.8	55.8	43.0	42.2	59.5	54.2	55.7	52.5
t5	85.5	61.8	24.7	36.2	59.0	48.5	51.3	33.8	58.8	46.7	63.0	61.0	67.3	44.3	36.8	55.3	56.8	55.8	50.0
ave	73.9	67.7	50.2	36.7	61.3	56.6	56.4	33.0	57.3	52.2	62.2	58.1	61.0	42.7	37.8	60.3	55.7	58.1	52.5
gpt-4o-2024-08-06																			
t1	68.5	64.8	61.3	40.5	60.0	62.8	57.3	33.3	59.0	60.8	59.7	59.2	58.0	52.3	41.3	61.8	59.3	58.7	55.3
t2	74.3	71.3	53.2	32.5	55.5	52.2	49.5	30.5	52.7	50.5	55.8	54.7	59.7	47.2	39.8	53.2	55.3	51.5	49.6
t3	89.2	82.3	71.8	45.0	75.2	68.2	68.0	32.7	69.8	71.2	71.3	71.8	71.5	55.8	52.7	72.0	64.5	67.5	64.3
t4	76.0	76.5	69.5	46.5	69.3	65.2	62.8	32.8	67.5	65.5	70.7	68.8	70.0	59.7	50.5	73.0	64.3	70.0	62.9
t5	79.8	75.0	67.2	40.8	70.3	67.5	63.2	33.5	66.0	62.7	68.2	64.0	63.7	53.7	47.0	66.7	62.0	58.8	59.7
ave	77.6	74.0	64.6	41.1	66.1	63.2	60.2	32.6	63.0	62.1	65.1	63.7	64.6	53.7	46.3	65.3	61.1	61.3	58.4
gemini pro 1.5																			
t1	74.2	69.5	68.3	46.5	64.7	64.7	58.2	34.7	60.5	57.3	67.2	62.7	66.0	51.2	38.2	66.0	59.2	61.8	57.9
t2	76.3	66.3	34.2	40.2	46.0	40.3	42.7	32.8	53.0	38.7	55.0	51.0	54.7	44.2	37.0	29.2	56.8	35.3	43.2
t3	88.5	82.0	75.3	47.5	73.3	68.2	64.8	33.2	64.8	61.3	73.7	68.5	68.3	54.5	42.3	69.3	61.8	65.0	62.0
t4	87.7	76.8	67.0	48.3	64.2	62.8	62.8	32.0	60.2	57.3	67.0	61.8	63.7	50.8	41.5	64.5	60.2	65.2	58.1
t5	83.0	76.8	73.2	43.7	65.0	66.2	63.0	32.3	61.7	58.3	67.8	63.3	64.5	54.5	41.7	67.2	60.0	67.5	59.4
avg	81.9	74.3	63.6	45.2	62.6	60.4	58.3	33.0	60.0	54.6	66.1	61.5	63.4	51.0	40.1	59.2	59.6	59.0	56.1

Table 13: **AfriXNLI results for all prompt templates and their averages**. Best prompt in Gray

Model	**eng**	**fra**	**amh**	**ewe**	**hau**	**ibo**	**kin**	**lin**	**lug**	**orm**	**sna**	**sot**	**swa**	**twi**	**wol**	**xho**	**yor**	**zul**	**ave**
mT0-XXL-MT																			
t1	36.4	33.0	27.4	24.8	29.2	28.4	27.2	28.6	26.0	27.4	26.6	30.4	32.2	30.2	24.8	26.8	28.8	28.2	27.9
t2	36.6	31.6	28.6	22.0	27.0	31.0	26.2	29.2	25.0	26.2	26.0	30.0	32.2	29.2	24.4	25.8	30.6	27.4	27.6
t3	38.0	32.2	30.8	22.8	28.2	29.8	25.8	30.8	25.6	25.8	24.6	30.0	30.8	27.6	23.0	25.0	27.4	27.0	27.2
t4	35.6	32.4	26.6	25.0	27.8	30.8	28.2	30.6	26.0	28.2	25.4	31.6	33.2	30.2	22.8	26.4	27.2	27.2	28.0
t5	37.4	32.6	28.8	22.8	27.6	30.4	25.8	31.4	25.2	25.8	24.4	28.4	31.8	28.4	26.2	25.8	28.4	27.8	27.4
ave	36.8	32.4	28.4	23.5	28.0	30.1	26.6	30.1	25.6	26.7	25.4	30.1	32.0	29.1	24.2	26.0	28.5	27.5	27.6
Aya-101																			
t1	41.0	37.2	32.2	27	33.4	34.2	29.6	27.6	27.4	26.4	26.2	33.2	35.2	26.8	24.4	31.6	29.0	29.8	29.6
t2	40.0	36.6	31.6	25.4	33.4	36.8	30.8	27.8	28.0	26.2	28.2	31.8	32.2	26.8	25.2	32	28.4	29.8	29.7
t3	41.8	36.0	32.6	26.0	32.0	34.4	31.0	28.2	27.8	26.2	27.0	32.8	34.2	26.8	25.0	31.8	30.0	28.6	29.7
t4	40.2	36.8	31.6	25.6	33.0	36.8	29.2	27.8	27.4	27.8	28.0	32.0	33.8	26.0	24.6	32.2	28.8	31.2	29.7
t5	42.8	36.8	32.0	25.8	31.2	36.6	31.6	28.8	26.6	26.4	27.6	32.2	33.6	26.6	25.0	31.6	29.0	29.2	29.6
ave	41.2	36.7	32.0	26.0	32.6	35.8	30.4	28.0	27.4	26.6	27.4	32.4	33.8	26.6	24.8	31.8	29.0	29.7	29.7
BLOOMZ 7B																			
t1	34.0	33.8	22.0	24.0	24.2	22.6	28.4	23.2	23.0	24.8	24.8	24.4	25.0	22.4	23.8	23.0	23.2	24.4	24.0
t2	33.8	32.2	19.0	23.6	24.0	24.2	27.0	23.0	21.2	23.6	24.8	24.2	27.2	25.2	26.0	24.4	24.6	22.8	24.1
t3	34.8	32.6	19.2	22.6	25.0	25.6	26.4	21.8	21.2	21.6	23.6	24.2	26.2	23	23.6	22.2	25.0	23.8	23.4
t4	34.4	34.0	21.6	21.6	23.4	25.2	27.0	24.0	22.2	23.0	25.0	23.0	28.0	24.6	25.0	22.6	24.2	22.0	23.9
t5	35.0	33.0	20.0	23.0	24.4	24.2	26.2	22.4	23.2	22.8	22.8	24.4	26.4	23.0	24.2	23.4	24.2	23.2	23.6
ave	34.4	33.1	20.4	23.0	24.2	24.4	27.0	22.9	22.2	23.2	24.2	24.0	26.6	23.6	24.5	23.1	24.2	23.2	23.8
LLaMa 3 8B																			
t1	43.2	38.6	27.4	24.8	26.6	27.4	27.2	27.4	27.4	26.2	25.8	25.4	28.8	25.8	25.0	23.6	27.6	28.2	26.5
t2	54.6	42.6	28.6	25.6	28.6	30.0	30.6	28.4	25.6	26.4	26.6	29.0	31.4	26.0	26.2	24.6	29.4	28.4	27.8
t3	45.0	39.2	27.4	21.8	27	27.8	27.6	27.2	25.8	27.6	26.2	25.6	29.0	24.8	24.4	24.8	28.8	26.4	26.4
t4	55.4	46	29.4	26.4	27.8	30.6	30.4	28.4	25	31.6	26.6	28.6	32	26.6	26.4	23.4	29	27.8	28.1
t5	45.2	41.8	25.6	24.0	26.8	29.0	30.0	28.8	23.8	28.8	26.2	25.8	32.6	23.8	25.6	22.8	29.2	26.8	26.9
ave	48.7	41.6	27.7	24.5	27.4	29.0	29.2	28.0	25.5	28.1	26.3	26.9	30.8	25.4	25.5	23.8	28.8	27.5	27.1
LLaMaX 3 8B																			
t1	48.4	38.8	30.6	27.0	33.4	30.4	30.0	31.4	26.2	28.2	28.4	28.8	35.8	25.4	26.8	29.0	29.0	27.2	29.2
t2	46.8	38.6	28.8	28.4	31.6	32.2	28.2	27.6	25.4	30.8	29.4	28.6	33.6	25.4	27.4	28.4	28.2	25.0	28.7
t3	45.4	38.8	28.2	28.6	32.2	31.4	27.6	28.2	25.0	30.0	30.4	26.6	34.6	24.6	27.0	28.0	28.8	25.6	28.6
t4	49.2	40.4	28.6	27.4	32.6	31.0	28.0	28.2	25.6	28.6	30.4	28.4	35.6	26.0	26.6	27.0	29.8	26.8	28.8
t5	47.2	38.0	30.2	28.0	33.4	32.0	27.8	28.4	25.2	30.0	31.6	29.4	35.6	25.8	27.0	28.0	29.8	26.6	29.3
ave	47.4	38.9	29.3	27.9	32.6	31.4	28.3	28.8	25.5	29.5	30.0	28.4	35.0	25.4	27.0	28.1	29.1	26.2	28.9
Gemma 2 9B																			
t1	69.8	62.6	41.3	29.3	37.9	39.2	31.8	37.0	30.5	32.8	39.7	34.7	48.9	32.1	28.7	33.2	33.2	36.1	35.4
t2	65.5	57.9	38.9	25.3	34.3	34.9	29.6	34.6	27.7	30.1	34.3	33.2	47.8	28.4	26.3	33.1	33.8	35.6	33.0
t3	68.3	61.5	41.1	27.9	38.1	37.5	31.2	36.3	28.7	33.0	39.6	35.9	47.2	32.8	29.1	32.5	34.8	38.5	35.3
t4	65.9	59.3	40.6	29.3	38.8	38.7	32.9	34.2	29.9	33.0	35.0	35.8	46.1	32.2	26.1	33.6	33.7	37.9	34.9
t5	67.6	60.0	39.9	27.1	37.9	36.0	29.4	33.8	28.1	31.8	36.9	34.1	49.1	30.8	28.9	31.8	35.5	35.4	34.2
ave	67.4	60.3	40.4	27.8	37.4	37.3	31.0	35.2	29.0	32.1	37.1	34.7	47.8	31.3	27.8	32.8	34.2	36.7	34.5
Gemma 2 27B																			

t1	75.6	66.4	40.6	32.4	43.2	44.2	40.2	38.2	32.6	33.6	44.6	41.8	56.0	35.6	30.4	42.0	41.0	42.2	39.9
t2	74.2	64.6	40.0	31.6	39.6	43.6	37.4	38.2	34.4	35.8	41.6	40.4	53.2	35.2	31.0	43.4	40.2	42.2	39.2
t3	72.2	63.6	41.0	30.6	40.6	41.2	38.4	38.4	31.8	33.0	41.8	41.0	51.4	34.6	32.0	43.2	38.2	42.0	38.7
t4	72.0	64.8	41.6	32.8	41.2	41.8	38.2	39.0	32.4	37.6	40.0	40.2	52.8	36.0	29.8	42.6	39.0	42.6	39.2
t5	70.4	63.4	39.8	31.0	41.8	42.0	39.2	37.8	33.0	35.0	43.2	40.2	50.0	36.0	33.2	42.0	40.0	41.8	39.1
ave	72.9	64.6	40.6	31.7	41.3	42.6	38.7	38.3	32.8	35.0	42.2	40.7	52.7	35.5	31.3	42.6	39.7	42.2	39.2
LLaMa 3.1 8B																			
t1	62.8	56.0	33.2	30.6	33.6	31.4	29.2	33.0	31.8	32.4	26.8	31.8	39.2	27.0	26.8	27.2	32.4	30.8	31.1
t2	55.6	52.0	33.2	29.2	31.8	33.4	29.0	30.0	30.0	29.0	28.0	31.0	38.0	27.0	25.8	27.0	32.2	31.2	30.4
t3	56.2	54.0	35.0	28.8	34.6	32.6	31.6	30.6	29.4	30.4	29.6	31.8	40.0	27.2	26.6	26.6	32.4	31.8	31.2
t4	55.8	49.8	34.6	29.2	30.6	30.0	29.4	30.4	28.8	30.0	26.6	27.8	35.8	26.4	24.2	25.6	31.4	28.8	29.4
t5	55.8	51.2	33.2	27.6	33.2	29.4	30.0	30.2	28.8	30.2	28.2	31.0	37.2	27.0	27.2	25.8	31.0	30.2	30.0
ave	57.2	52.6	33.8	29.1	32.8	31.4	29.8	30.8	29.8	30.4	27.8	30.7	38.0	26.9	26.1	26.4	31.9	30.6	30.4
LLaMa 3.1 70B																			
t1	76.4	69.4	41.6	32.2	47.6	47.2	38.6	40.0	34.4	35.6	41.6	39.0	55.8	28.4	31.6	34.2	41.4	40.4	39.4
t2	74.4	68.6	41.4	30.2	42.4	47.2	35.4	39.2	33.6	35.2	37.6	35.6	54.6	31.0	32.6	32.6	40.2	39.2	38.0
t3	73.0	66.6	40.4	32.2	43.6	46.2	36.2	37.4	32.2	35.2	40.6	36.6	52.2	33.2	32.4	32.2	38.8	38.8	38.0
t4	73.4	68.8	41.0	27.8	42.8	46.4	36.4	38.2	33.6	36.0	37.8	35.4	56.4	30.4	29.2	30.2	38.4	37.0	37.3
t5	74.2	65.0	39.8	30.2	42.0	45.8	35.2	38.4	33.0	35.2	38.6	38.2	52.4	32.2	33.0	32.6	41.2	40.2	38.0
ave	74.3	67.7	40.8	30.5	43.7	46.6	36.4	38.6	33.4	35.4	39.2	37.0	54.3	31.0	31.8	32.4	40.0	39.1	38.1
CommandR (Aug)																			
t1	62.8	54.6	28.8	26.0	25.6	24.8	29.2	31.8	27.2	28.2	27.6	25.6	31.4	27.2	27.6	24.2	29.4	24.4	27.4
t2	58.4	52.0	28.2	24.8	26.6	24.4	25.6	30.6	24.6	26.0	27.6	24.6	29.6	28.8	24.6	24.2	28.6	29.8	26.8
t3	51.0	48.0	23.0	25.4	26.0	26.0	25.4	28.2	24.0	25.0	25.0	23.8	29.6	28.0	24.8	25.8	26.2	26.6	25.8
t4	61.0	55.8	26.8	30.0	27.8	30.0	29.8	30.0	27.2	26.8	31.4	25.8	34.0	29.2	28.2	24.6	30.2	25.0	28.6
t5	50.0	47.2	24.2	23.0	24.0	22.8	23.6	23.8	23.6	25.0	24.6	22.8	28.8	22.6	22.4	25.4	25.2	27.0	24.3
ave	56.6	51.5	26.2	25.8	26.0	25.6	26.7	28.9	25.3	26.2	27.2	24.5	30.7	27.2	25.5	24.8	27.9	26.6	26.6
gpt-3.5-turbo-0125																			
t1	72.0	66.6	31.2	34.8	38.0	39.8	34.2	40.0	39.0	38.0	39.8	39.0	52.4	37.8	32.8	38.6	37.2	36.2	38.1
t2	68.4	62.2	31.0	32.4	36.0	37.6	36.6	38.0	33.2	33.0	38.6	37.2	48.8	36.0	29.4	34.6	40.0	36.6	36.2
t3	67.4	62.4	33.0	25.0	34.6	37.4	36.2	39.4	34.2	31.8	38.2	37.8	48.0	33.8	26.8	37.4	37.2	36.4	35.5
t4	72.0	63.2	33.6	33.2	37.8	35.6	36.2	37.6	39.6	38.4	37.6	39.6	53.0	34.4	29.2	37.6	35.6	39.2	37.4
t5	65.2	62.2	31.2	30.0	32.6	36.0	32.4	36.2	34.4	32.2	36.4	40.2	48.6	34.4	27.8	33.6	33.8	33.4	34.6
ave	69.0	63.3	32.0	31.1	35.8	37.3	35.1	38.2	36.1	34.7	38.1	38.8	50.2	35.3	29.2	36.4	36.8	36.4	36.3
gpt-4o-mini-2024-07-18																			
t1	82.6	78.2	37.2	28.0	51.4	46.2	52.8	44.0	39.4	46.2	53.6	46.8	64.0	38.0	31.6	50.6	48.4	50.0	45.5
t2	78.8	74.0	35.2	25.0	47.8	45.2	46.8	39.2	33.0	40.2	47.8	44.2	64.0	32.4	27.4	45.0	40.6	42.8	41.0
t3	79.8	76.0	38.4	27.8	46.6	45.8	40.2	43.4	33.6	43.4	48.2	45.2	60.0	35.2	27.8	48.6	41.2	49.2	42.2
t4	39.4	36.0	27.8	25.0	29.0	28.6	26.8	27.0	24.6	26.8	29.0	28.0	36.2	23.8	24.6	26.6	26.8	29.0	27.5
t5	79.8	75.6	40.0	31.4	46.8	43.8	39.6	41.8	36.2	43.4	50.0	43.8	61.4	37.4	30.4	47.6	45.0	47.4	42.9
ave	72.1	68.0	35.7	27.4	44.3	41.9	41.2	39.1	33.4	40.0	45.7	41.6	57.1	33.4	28.4	43.7	40.4	43.7	39.8
gpt-4-turbo-2024-04-09																			
t1	80.4	79.4	44.4	32.4	63.2	61.8	60.4	53.6	49.6	49.4	63.4	63.4	74.4	35.2	35.0	63.4	54.6	63.2	54.2
t2	81.6	75.2	49.2	31.2	60.6	58.0	58.6	46.8	41.2	50.6	61.4	62.4	74.2	33.0	28.4	62.2	53.0	59.6	51.9
t3	83.2	79.8	49.4	30.4	65.0	60.0	59.6	51.4	47.4	51.2	62.0	63.8	75.6	32.2	31.2	66.0	56.2	63.8	54.1
t4	62.0	50.6	28.4	27.8	50.0	41.0	47.6	35.2	35.6	35.0	41.0	44.2	63.4	29.2	26.6	44.2	40.8	40.8	39.4
t5	85.2	79.2	48.2	31.2	63.0	59.6	58.8	50.6	48.2	50.6	64.4	64.4	75.2	34.8	29.6	65.0	58.2	63.6	54.1
ave	78.5	72.8	43.9	30.6	60.4	56.1	57.0	47.5	44.4	47.4	58.4	59.6	72.6	32.9	30.2	60.2	52.6	58.2	50.7
gpt-4o-2024-08-06																			
t1	87.4	83.2	59.8	33.6	67.2	67.2	64.2	61.0	52.8	61.0	67.6	67.4	77.4	43.2	37.8	70.2	61.2	68.2	60.0
t2	87.6	79.6	53.2	31.6	65.6	62.0	60.2	53.4	49.4	55.0	62.2	60.6	75.0	42.4	33.4	66.6	57.0	65.2	55.8
t3	87.6	83.2	56.8	32.0	67.8	65.4	63.8	60.8	51.4	59.6	68.0	65.8	76.8	41.8	34.8	69.8	61.6	68.0	59.0
t4	61.2	52.2	37.0	26.2	46.2	44.8	40.4	37.8	36.8	40.2	41.6	41.0	57.8	33.4	27.2	43.4	40.6	39.6	39.6
t5	88.0	84.4	57.4	34.8	66.8	64.0	65.0	60.8	52.0	61.6	66.8	68.8	76.2	44.4	35.0	70.2	61.4	68.6	59.6
ave	82.4	76.5	52.8	31.6	62.7	60.7	58.7	54.8	48.5	55.5	61.2	60.7	72.6	41.0	33.6	64.0	56.4	61.9	54.8
gemini-pro 1.5																			
t1	82.0	81.8	68.0	39.0	71.2	70.4	65.0	55.2	53.0	55.8	66.8	67.6	78.4	48.2	32.0	69.2	57.4	66.0	60.2
t2	70.2	63.4	53.6	37.0	51.6	55.6	47.8	45.8	42.8	44.0	51.8	52.2	56.2	42.8	33.6	50.0	47.4	48.2	47.5
t3	79.0	65.0	56.8	38.6	57.6	61.2	55.4	49.6	47.8	47.8	52.6	58.8	63.0	46.6	33.2	56.4	52.6	56.0	52.1
t4	83.4	73.8	56.2	34.2	55.4	59.4	51.2	44.4	42.4	41.2	54.0	54.6	66.0	37.6	27.6	54.2	46.0	50.8	48.5
t5	88.8	81.4	60.6	44.2	67.0	70.4	63.6	59.0	54.6	57.4	65.6	68.2	77.6	46.2	32.4	66.8	61.4	63.6	59.9
ave	80.7	73.1	59.0	38.6	60.6	63.4	56.6	50.8	48.1	49.2	58.2	60.3	68.2	44.3	31.8	59.3	53.0	56.9	53.6

Table 14: **AfriMMLU results for all prompt templates and their averages**. Best prompt in Gray

Model	**eng**	**fra**	**amh**	**ewe**	**hau**	**ibo**	**kin**	**lin**	**lug**	**orm**	**sna**	**sot**	**swa**	**twi**	**vai**	**wol**	**xho**	**yor**	**zul**	**ave**
mT0-XXL-MT																				
t1	3.6	4.8	4.4	1.6	4.8	1.2	3.6	2	4	0.8	3.2	4	3.6	1.2	0.8	2	4.8	2	3.2	2.8
t2	4.0	4.4	3.2	3.2	3.2	0.4	3.2	1.2	3.2	1.2	2.4	3.6	4.4	1.6	1.6	2.0	2.8	0.8	2.8	2.4
t3	3.2	4.0	3.6	2.0	3.2	0.4	3.2	2.0	3.2	0.8	4.0	2.8	3.2	1.6	1.6	2.4	3.2	1.6	2.8	2.4
t4	3.6	4.4	3.6	2.8	4.0	0.8	3.2	0.8	3.6	1.2	2.0	3.2	4.0	2.0	1.6	2.0	3.2	2.8	2.4	2.5
t5	3.6	4.4	3.6	2.8	3.2	0.4	2.8	2.0	3.2	0.8	2.4	3.6	4.0	1.6	0.8	0.8	4.0	2.0	2.4	2.4
ave	3.6	4.4	3.7	2.5	3.7	0.6	3.2	1.6	3.4	1.0	2.8	3.4	3.8	1.6	1.3	1.8	3.6	1.8	2.7	2.5
Aya-101																				
t1	10.8	9.6	6.8	3.2	7.2	3.2	3.6	4.4	2.4	4.8	6.8	10.8	6	1.6	0.8	1.2	4.4	3.2	3.6	4.4
t2	11.6	10.4	7.6	2.8	4.4	2.8	5.2	4	2.4	3.2	6	7.2	8.4	1.6	1.6	0.8	4	4.4	3.6	4.1
t3	10.0	9.2	7.2	3.6	7.2	3.2	6.4	5.2	1.6	3.2	4.8	8.0	6.8	1.6	0.8	0.8	4.0	3.6	3.6	4.2

t4	10.4	8.4	8.0	2.8	5.2	2.8	5.2	3.6	5.2	5.6	6.0	8.0	5.6	1.6	2.0	0.4	4.4	3.2	4.8	4.4
t5	10.0	8.4	6.0	3.2	6.4	3.6	5.6	3.6	2.8	3.2	6.4	6.0	6.8	2.8	2.0	2.4	4.0	3.6	4.8	4.3
ave	10.6	9.2	7.1	3.1	6.1	3.1	5.2	4.2	2.9	4.0	6.0	8.0	6.7	1.8	1.4	1.1	4.2	3.6	4.1	4.3
BLOOMZ 7B																				
t1	2.4	2.4	1.2	0.8	1.6	0.8	1.2	1.2	2.0	0.4	0.4	1.2	2.0	0.8	1.2	1.6	1.2	1.2	0.4	1.1
t2	2.8	0.8	1.6	0.8	0.4	0.8	0.0	1.2	1.6	1.6	2.0	2.8	1.6	1.2	1.2	2.0	1.2	2.4	0.8	1.4
t3	2.8	3.2	1.2	0.4	2.4	0.0	0.8	0.8	1.2	1.6	2.4	2.4	2.0	1.2	1.6	1.2	1.6	1.2	0.8	1.3
t4	2.8	2.0	1.6	2.4	2.8	1.6	1.6	1.6	2.4	1.6	0.8	1.6	0.8	0.8	1.2	2.4	1.6	2.0	2.0	1.7
t5	2.4	0.8	0.8	1.6	1.6	1.6	2.0	2.0	2.0	2.4	1.6	1.2	1.6	1.6	1.6	2.0	1.6	2.4	2.0	1.7
ave	2.6	1.8	1.3	1.2	1.8	1.0	1.1	1.4	1.8	1.5	1.4	1.8	1.6	1.1	1.4	1.8	1.4	1.8	1.2	1.5
LLaMa 3 8B																				
t1	29.2	33.2	2	2.8	6.4	5.2	5.2	2.4	5.6	1.6	2	1.6	10	3.6	2.8	3.2	5.2	2.4	3.6	3.9
t2	34.8	32	3.2	2	6.4	6	4.8	3.2	4.4	2.4	1.6	1.2	12.4	3.2	0.4	1.2	4	3.2	2.8	3.7
t3	59.2	44	2	2.4	6.8	4	3.2	2.8	3.6	2.4	2.4	2.8	16.8	1.6	1.2	3.2	0.4	4	3.2	3.7
t4	60.8	60.4	2	3.2	8.8	6	2.8	4.4	4.8	3.6	2.8	4	26.4	2	2	2	3.6	4.4	3.6	5.1
t5	43.2	28.4	2.8	2	6	6.4	4.8	4	4.4	2.4	2.8	2.8	20.8	3.2	1.2	2.8	3.2	2.8	4.8	4.5
ave	45.4	39.6	2.4	2.5	6.9	5.5	4.2	3.4	4.6	2.5	2.3	2.5	17.3	2.7	1.5	2.5	3.3	3.36	3.6	4.2
LLaMa 3.1 8B																				
t1	41.6	39.2	2.0	4.4	6.8	3.6	4.4	2.4	2.0	3.2	6.4	2.8	28.8	2.4	1.2	2.4	4.4	4.4	3.6	5.0
t2	30.8	46.0	3.6	3.2	8.4	3.6	3.2	4.0	5.6	3.2	4.4	5.6	30.4	2.8	1.2	4.0	1.2	1.6	5.2	5.4
t3	38.4	34	2.8	2.4	7.6	4.8	4	2.8	3.6	2.4	6	4	24.4	4.4	0.4	2	1.6	2	1.2	4.5
t4	56.8	54	4.4	2.4	10	4.4	7.2	3.2	7.6	3.6	4.8	6.4	41.2	4	1.2	2.8	2.4	5.6	5.2	6.8
t5	47.6	55.6	2.4	2.4	9.2	5.6	7.6	3.6	3.6	4.4	5.6	3.2	30.4	5.2	2.0	1.6	2.4	5.6	5.2	5.9
ave	43.0	45.8	3.0	3.0	8.4	4.4	5.3	3.2	4.5	3.4	5.4	4.4	31.0	3.8	1.2	2.6	2.4	3.8	4.1	5.5
LLaMaX 3 8B																				
t1	12.4	9.6	4.0	1.2	4.4	2.8	0.4	2.0	3.2	1.2	3.6	1.6	2.8	1.2	2.8	2.0	1.6	4.0	6.0	2.6
t2	10.0	9.6	3.6	2.4	4.4	2.4	3.2	3.2	5.6	1.6	2.4	4.4	8.4	0.8	2.0	2.4	4.8	2.0	5.2	3.5
t3	16.0	10.0	6.0	2.0	5.6	4.4	5.2	4.0	4.4	1.2	4.4	2.4	6.0	1.2	2.4	2.4	6.8	2.8	6.4	4.0
t4	15.2	12.0	3.6	1.6	9.6	5.2	4.4	2.8	5.2	1.6	6.0	3.2	10.8	1.6	4.0	3.2	5.6	4.8	7.2	4.7
t5	8.8	9.6	4.0	0.4	6.8	2.4	2.8	1.6	6.0	2.8	4.0	3.2	10.0	1.2	2.0	3.2	4.0	2.0	6.4	3.7
ave.	12.5	10.2	4.2	1.5	6.2	3.4	3.2	2.7	4.9	1.7	4.1	3.0	7.6	1.2	2.6	2.6	4.6	3.1	6.2	3.7
Gemma 2 9B																				
t1	44.0	43.2	26.0	3.2	26.4	9.6	16.0	9.2	11.6	8.0	20.4	13.2	46.4	6.4	0.4	2.8	16.0	8.4	14.8	14.0
t2	68.8	61.2	26.4	4.8	35.2	11.6	22.4	11.6	17.6	9.6	26.0	22.0	61.6	9.2	0.4	3.6	20.8	13.6	21.2	18.7
t3	43.6	45.6	26.8	2.4	26.4	10.0	17.2	9.6	12.4	4.8	20.4	16.4	50.0	5.6	0.0	3.2	18.4	9.6	20.8	14.9
t4	54.0	52.8	21.2	4.8	17.6	4.4	13.2	8.4	8.8	4.0	16.8	9.2	46.4	5.2	0.8	2.8	12.8	7.6	13.2	11.6
t5	7.6	16.0	2.4	1.2	12.4	3.2	10.8	6.8	10.4	5.6	23.6	17.2	44.0	5.6	0.0	1.6	14.8	10.8	20.8	11.2
ave.	43.6	43.8	20.6	3.3	23.6	7.8	15.9	9.1	12.2	6.4	21.4	15.6	49.7	6.4	0.3	2.8	16.6	10.0	18.2	14.1
Gemma 2 27B																				
t1	80.4	64.4	24.4	4.4	43.6	16.8	28.0	12.0	19.2	10.0	31.2	30.0	64.4	10.0	0.8	6.0	25.6	20.8	32.4	22.3
t2	83.6	75.2	39.2	6.0	46.8	24.8	31.2	16.8	22.4	14.8	34.4	32.8	71.6	8.4	3.2	5.2	30.0	23.6	35.2	26.3
t3	78.0	68.4	38.0	4.8	48.0	22.4	28.8	16.4	19.6	9.6	32.0	30.8	68.8	10.8	2.0	4.8	32.8	23.2	32.8	25.0
t4	85.6	80.0	33.6	7.6	49.6	24.0	32.4	18.0	23.6	12.8	35.2	38.4	73.6	12.4	3.2	5.6	32.0	22.4	34.4	27.0
t5	78.8	64.0	36.8	5.2	48.8	24.4	36.0	16.8	23.2	10.8	37.2	32.4	64.4	11.6	4.0	6.8	30.4	22.8	34.4	26.2
ave.	81.3	70.4	34.4	5.6	47.4	22.5	31.3	16.0	21.6	11.6	34.0	32.9	68.6	10.6	2.6	5.7	30.2	22.6	33.8	25.4
LLaMa 3.1 70B																				
t1	85.6	70.4	12.0	5.2	42.8	24.4	24.8	9.6	22.8	10.4	17.2	24.8	63.6	10.8	1.6	4.4	14.0	15.6	24.0	19.3
t2	77.6	71.2	7.2	6.4	47.2	32.0	25.2	10.8	20.4	9.6	17.2	24.4	59.6	6.8	1.2	3.2	10.8	18.8	20.8	18.9
t3	83.6	72.0	6.4	5.2	42.4	28.8	23.2	13.2	20.4	10.4	16.0	28.0	58.4	7.2	1.2	3.6	14.0	16.4	21.2	18.6
t4	86.8	76.4	17.6	8.0	48.8	37.2	26.4	11.2	24.4	10.8	21.2	32.8	68.0	14.4	0.4	3.2	18.8	23.6	27.2	23.2
t5	86.0	78.4	22.0	5.2	45.2	31.2	30.8	13.6	20.0	12.8	21.2	28.0	63.6	7.6	1.6	6.0	16.8	18.8	24.8	21.7
ave	83.9	73.7	13.0	6.0	45.3	30.7	26.1	11.7	21.6	10.8	18.6	27.6	62.6	9.4	1.2	4.1	14.9	18.6	23.6	20.3
CommandR (Aug)																				
t1	72.8	58.8	2.4	3.2	3.6	1.6	2.8	7.6	4.4	3.6	4.4	3.6	16.8	3.6		4.8	2.8	4.0	2.0	4.5
t2	46.8	18.8	3.2	2.8	2.0	2.4	2.8	3.2	2.4	2.4	3.2	4.0	8.0	5.2		2.4	3.2	4.0	2.4	3.4
t3	74.4	64.4	4.4	2.4	6.0	2.0	5.2	8.4	5.6	3.2	4.0	6.8	20.8	3.6		2.4	5.6	5.2	4.8	5.7
t4	12.0	9.2	6.8	2.8	1.6	2.4	4.4	5.6	3.2	1.6	3.2	2.8	6.0	2.4		3.2	3.6	2.8	3.2	3.5
t5	52.0	19.2	2.0	3.2	2.8	2.0	3.6	3.6	4.4	2.4	3.2	3.6	10.8	3.6		3.2	2.8	3.6	3.6	3.7
ave.	51.6	34.1	3.8	2.9	3.2	2.1	3.8	5.7	4.0	2.6	3.6	4.2	12.5	3.7		3.2	3.6	3.9	3.2	4.1
gpt-3.5-turbo-0125																				
t1	70.4	52.4	1.2	4.0	8.8	2.4	8.4	7.6	4.4	5.2	8.4	4.4	54.4	5.6	2.4	3.6	5.2	8.0	8.8	8.4
t2	73.6	56.8	2.0	3.2	10.0	2.8	11.6	7.2	9.2	7.2	9.2	7.6	59.2	4.8	3.6	3.2	9.2	8.8	13.6	10.1
t3	65.6	59.6	5.2	2.4	6.8	2.8	12.8	8.0	7.6	7.2	11.2	8.8	52.4	3.6	2.0	2.8	9.2	11.2	9.6	9.6
t4	48.4	35.2	4.4	3.6	6.4	1.6	7.2	7.6	6.4	3.2	6.4	3.2	31.2	4.8	2.8	3.6	4.0	3.6	7.6	6.3
t5	69.6	56.0	4.0	4.8	10.0	2.0	11.6	8.0	8.8	6.0	9.6	5.2	52.4	4.0	4.4	3.6	7.6	6.0	11.6	9.4
ave.	65.5	52.0	3.4	3.6	8.4	2.3	10.3	7.7	7.3	5.8	9.0	5.8	49.9	4.6	3.0	3.4	7.0	7.5	10.2	8.8
gpt-4o-mini-2024-07-18																				
t1	80.0	71.6	5.6	4.8	50.4	33.2	45.2	18.4	18.0	40.4	43.2	32.4	63.6	11.2	2.0	7.2	32.0	38.4	37.2	28.4
t2	85.2	71.6	31.6	6.0	56.0	33.6	48.0	25.6	29.2	39.2	44.8	36.8	70.8	15.6	2.8	7.6	32.4	45.6	43.6	33.5
t3	78.0	70.0	34.0	6.0	52.8	34.8	41.2	20.8	27.2	40.0	39.6	38.0	67.2	17.2	2.4	5.6	33.6	43.2	42.4	32.1
t4	86.0	69.2	24.4	5.2	53.2	26.4	43.6	19.2	22.0	37.6	41.6	40.0	71.6	14.0	3.2	4.4	32.0	38.8	41.6	30.5
t5	82.0	72.0	25.2	8.4	54.4	34.8	50.0	28.0	31.2	42.8	47.6	34.4	73.6	16.8	4.0	6.0	34.4	40.4	44.8	33.9
ave.	82.2	70.9	24.2	6.1	53.4	32.6	45.6	22.4	25.5	40.0	43.4	36.3	69.4	15.0	2.9	6.2	32.9	41.3	41.9	31.7
gpt-4-turbo-2024-04-09																				
t1	77.2	61.2	7.6	6.8	56.0	38.0	54.4	36.8	37.6	42.4	55.6	52.4	74.0	7.6	3.6	8.0	45.2	42.8	47.6	36.3
t2	86.0	69.2	40.4	8.0	64.4	48.4	58.0	38.8	46.0	47.6	58.0	56.4	77.6	10.4	2.0	9.6	49.6	56.8	52.4	42.6
t3	76.8	65.2	11.2	6.8	56.4	37.2	54.0	39.6	39.6	44.0	58.0	52.0	75.6	8.8	0.4	6.8	42.8	48.8	47.6	37.0
t4	89.2	70.0	28.4	4.0	54.8	37.6	52.4	29.6	38.4	38.0	53.2	44.4	66.4	9.2	3.2	8.4	38.4	42.8	41.6	34.8
t5	78.8	65.6	11.6	10.8	59.2	41.6	59.6	42.4	40.0	48.8	60.8	55.6	76.8	11.6	4.0	11.6	48.0	54.8	53.2	40.6

ave.	81.6	66.2	19.8	7.3	58.2	40.6	55.7	37.4	40.3	44.2	57.1	52.2	74.1	9.5	2.6	8.9	44.8	49.2	48.5	38.3
gpt-4o-2024-08-06																				
t1	83.2	65.6	49.2	8.0	64.4	54.8	60.0	42.0	48.0	57.6	58.4	56.0	74.8	27.2	3.6	23.6	47.6	64.8	54.0	46.7
t2	84.0	68.8	57.6	8.8	64.8	57.6	60.4	51.2	51.6	61.2	58.4	60.8	78.8	31.2	4.8	28.0	52.4	62.0	57.2	49.8
t3	81.6	68.4	58.0	6.8	62.4	55.6	59.2	51.6	50.8	54.4	57.6	60.4	74.8	29.2	2.4	24.4	49.6	63.6	54.4	48.0
t4	88.8	69.2	55.6	11.2	66.8	56.8	64.4	48.4	50.4	56.0	62.0	60.8	80.0	29.2	3.6	23.2	51.2	63.6	58.0	49.5
t5	81.2	68.0	56.0	10.0	65.2	58.4	61.6	53.6	52.8	56.4	61.6	59.6	74.8	28.8	2.4	32.8	49.2	62.4	57.6	49.6
ave.	83.8	68.0	55.3	9.0	64.7	56.6	61.1	49.4	50.7	57.1	59.6	59.5	76.6	29.1	3.4	26.4	50.0	63.3	56.2	48.7
gemini pro 1.5																				
t1	83.2	63.6	73.6	34.4	68.8	62.4	65.6	47.6	48.0	49.2	62.0	59.2	80.0	30.8	1.6	10.8	52.0	62.4	54.4	50.8
t2	76.8	76.8	59.6	35.2	69.2	58.0	64.0	54.8	51.6	47.6	62.4	62.0	83.6	32.0	4.4	11.2	55.2	58.4	56.8	50.9
t3	81.6	74.4	70.0	38.0	66.8	58.8	65.6	54.0	53.6	51.2	61.2	60.4	77.6	27.6	1.2	12.8	53.6	63.2	56.0	51.3
t4	53.6	48.0	40.4	14.4	38.4	26.8	36.0	26.8	29.6	24.4	30.0	33.2	47.2	14.8	2.0	6.4	29.2	38.8	30.4	27.6
t5	82.8	67.2	67.6	40.8	65.6	63.2	63.6	51.2	54.4	57.2	64.4	61.2	76.4	38.0	3.2	12.8	50.8	61.6	57.2	52.3
ave.	75.6	66.0	62.2	32.6	61.8	53.8	59.0	46.9	47.4	45.9	56.0	55.2	73.0	28.6	2.5	10.8	48.2	56.9	51.0	46.6

Table 15: **AfriMGSM results for all prompt templates and their averages**. Best prompt in Gray

Model	**eng**	**fra**	**amh**	**ewe**	**hau**	**ibo**	**kin**	**lin**	**lug**	**orm**	**sna**	**sot**	**swa**	**twi**	**wol**	**xho**	**yor**	**zul**	**ave**
In-language																			
afro-xlmr-large-76L	88.2	83.3	78.5	58.3	73.3	70.0	65.8	33.3	68.0	69.3	70.8	70.8	73.3	59.5	51.8	73.0	63.2	72.5	65.7
mT0-XXL-MT (t2)	63.5	61.5	58.3	38.5	57.5	56.5	51.3	33.5	54.2	47.2	54.8	56.0	55.5	48.8	39.7	56.3	52.3	55.0	51.0
Aya-101 (t4)	67.0	59.7	64.2	43.2	57.0	55.5	54.3	33.5	51.7	51.5	55.7	52.2	56.5	47.0	36.7	55.2	54.5	55.3	51.5
BLOOMZ 7B (t2)	60.3	56.0	36.8	35.7	36.5	44.7	38.5	33.8	41.5	35.2	43.3	40.5	45.8	37.5	36.0	39.8	45.2	40.3	39.4
LLaMa 3 8B (t2)	43.9	47.4	40.2	34.3	35.6	39.1	32.9	32.7	33.3	35.4	35.5	35.9	34.1	34.3	33.3	36.8	35.17	38.5	35.4
LLaMa 3.1 8B (t2)	53.0	50.2	38.8	37.7	36.5	38.5	38.3	32.3	37.5	35.0	35.8	34.8	43.8	37.0	32.3	33.8	38.5	34.0	36.6
LLaMaX 3 8B (t1)	53.0	49.3	40.2	35.8	44.8	42.5	37.7	29.8	42.7	40.5	47.0	39.7	45.7	36.3	37.3	46.2	44.0	42.0	40.8
Gemma 2 9B (t2)	55.3	50.7	43.2	35.3	47.0	40.7	40.2	32.8	38.8	37.8	42.3	40.2	46.3	37.2	35.2	42.5	41.3	44.5	40.3
Gemma 2 27B (t2)	67.8	63.3	47.0	36.8	49.7	46.2	40.5	32.0	41.7	35.8	46.0	43.5	57.0	36.0	36.7	45.0	42.5	48.0	42.8
LLaMa 3.1 70B (t2)	57.3	50.7	43.2	34.3	42.8	42.3	36.5	32.8	37.5	34.7	35.5	38.3	44.0	36.0	34.7	39.3	39.0	37.0	38.0
CommandR (Aug) (t3)	79.9	75.7	45.7	38.9	45.1	42.4	45.6	34.5	39.0	44.8	48.9	45.1	55.3	41.8	34.5	45.9	45.9	40.8	43.4
gpt-3.5-turbo-0125 (t4)	70.0	65.2	38.7	38.2	39.0	42.3	42.2	32.0	43.0	42.5	45.8	43.0	58.8	39.5	39.5	44.7	40.5	43.8	42.1
gpt-4o-mini-2024-07-18 (t3)	86.2	80.3	52.2	37.2	64.8	57.2	56.8	31.5	51.7	61.3	63.0	56.2	63.7	47.5	43.2	64.5	54.5	62.0	54.2
gpt-4-turbo-2024-04-09 (t3)	87.3	81.0	62.0	38.2	70.3	66.0	65.8	33.7	67.5	61.0	71.7	67.8	67.8	41.7	38.0	71.5	62.0	66.7	59.5
gpt-4o-2024-08-06 (t3)	89.2	82.3	71.8	45.0	75.2	68.2	68.0	32.7	69.8	71.2	71.3	71.8	71.5	55.8	52.7	72.0	64.5	67.5	64.3
Claude OPUS (t1)	85.7	74.7	61.3	54.5	61.5	45.0	64.5	30.5	63.7	50.2	57.0	68.3	70.5	56.0	50.0	68.8	63.7	63.7	58.1
Gemini-1.5-pro(t3)	88.5	82.0	75.3	47.5	73.3	68.2	64.8	33.2	64.8	61.3	73.7	68.5	68.3	54.5	42.3	69.3	61.8	65.0	62.0
Translate-test																			
afro-xlmr-large-76L		83	73.7	54.3	67.2	66	63	32.8	65.7	65.8	71.2	70.2	73	56.8	47.5	74.2	63.7	72	63.6
mT0-XXL-MT (t2)		59.8	54.5	45.3	52.7	50.0	49.8	34.5	48.2	50.2	55.0	53.0	56.8	46.2	42.7	54.8	50.7	54.3	49.9
Aya-101 (t4)		61.2	60.7	40.8	53.8	52.3	50.3	33.0	48.7	51.3	54.0	56.0	54.0	44.5	39.3	56.8	51.0	57.0	50.2
BLOOMZ 7B (t2)		56.8	52.8	42.7	51.3	48.5	47.3	34.2	46.5	48.0	52.0	51.3	51.2	44.7	40.2	52.8	47.5	51.3	47.6
LLaMa 3 8B (t2)		42.7	38.7	37.7	39.5	38.0	35.0	33.2	40.7	40.2	40.0	40.8	39.5	36.2	35.3	40.3	36.2	39.8	38.2
LLaMa 3.1 8B (t2)		49.8	49.2	42.1	45.3	43.6	42.8	35.0	43.5	44.5	45.3	46.9	46.8	40.7	37.4	45.8	42.8	45.8	43.6
LLaMaX 3 8B (t1)		33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Gemma 2 9B (t2)		50.7	47.5	41.3	43.3	44.7	42.5	32.7	47.2	44.2	43.8	42.7	46.7	42.2	39.2	45.7	43.5	45.7	43.3
Gemma 2 27B (t2)		60.8	55.8	45.5	48.5	49.7	47.7	31.5	51.3	52.3	51.7	50.2	55.7	47.2	40.2	54.8	50.2	51.5	49.0
LLaMa 3.1 70B (t2)		58.8	42.8	39.3	45.2	39.8	36.2	32.7	44.3	46.0	51.8	50.8	52.5	38.8	34.5	44.8	41.8	42.8	42.8
CommandR (Aug) (t3)		75.1	67.7	52.1	57.0	55.9	58.9	32.0	59.6	56.0	62.0	64.1	66.7	45.5	46.8	63.7	60.1	64.0	57.0
gpt-3.5-turbo-0125 (t4)		56.2	40.3	41.7	48.8	48.8	46.7	32.5	49.8	45.3	50.8	51.8	45.0	41.7	39.0	51.5	44.5	50.2	45.5
gpt-4o-mini-2024-07-18 (t3)		75.8	68.0	49.8	57.0	59.7	56.5	32.2	56.3	58.0	63.0	63.2	66.2	49.5	43.0	65.2	57.2	62.8	56.7
gpt-4-turbo-2024-04-09 (t3)		77.3	69.8	45.3	59.0	59.7	57.2	33.2	57.3	60.3	63.2	64.5	64.7	49.7	42.0	67.2	56.0	63.7	57.0
gpt-4o-2024-08-06 (t3)		73.8	63.8	42.8	54.8	53.3	52.8	32.3	54.3	55.5	58.2	58.0	58.5	45.7	38.8	58.2	53.5	52.8	52.1
Claude OPUS (t1)		56.0	65.7	45.7	76.3	59.0	55.0	32.0	57.7	55.8	61.5	62.5	63.8	46.3	41.8	63.5	55.3	59.8	56.4
Gemini-1.5-pro (t1)		71.5	60.0	41.0	50.3	50.8	48.2	31.8	49.0	52.0	56.3	54.7	57.8	45.7	38.2	58.0	47.5	57.3	49.9

Table 16: **AfriXNLI results for in-language and translate-test**. We make use of the best prompt

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
in-language																			
mT0-XXL-MT (t1)	36.4	33.0	27.4	24.8	29.2	28.4	27.2	28.6	26.0	27.4	26.6	30.4	32.2	30.2	24.8	26.8	28.8	28.2	27.9
Aya-101 (t2)	40.0	36.6	31.6	25.4	33.4	36.8	30.8	27.8	28.0	26.2	28.2	31.8	32.2	26.8	25.2	32.0	28.4	29.8	29.7
BLOOMZ 7B (t2)	33.8	32.2	19.0	23.6	24.0	24.2	27.0	23.0	21.2	23.6	24.8	24.2	27.2	25.2	26.0	24.4	24.6	22.8	24.1
LLaMa 3 8B (t4)	55.4	46.0	29.4	26.4	27.8	30.6	30.4	28.4	25.0	31.6	26.6	28.6	32.0	26.6	26.4	23.4	29	27.8	28.1
LLaMa 3.1 8B (t1)	62.8	56.0	33.2	30.6	33.6	31.4	29.2	33.0	31.8	32.4	26.8	31.8	39.2	27.0	26.8	27.2	32.4	30.8	31.1
LLaMaX 3 8B (t5)	47.2	38.0	30.2	28.0	33.4	32.0	27.8	28.4	25.2	30.0	31.6	29.4	35.6	25.8	27.0	28.0	29.8	26.6	29.3
Gemma 2 9B (t1)	69.8	62.6	41.3	29.3	37.9	39.2	31.8	37.0	30.5	32.8	39.7	34.7	48.9	32.1	28.7	33.2	33.2	36.1	35.4
Gemma 2 27B (t1)	75.6	66.4	40.6	32.4	43.2	44.2	40.2	38.2	32.6	33.6	44.6	41.8	56.0	35.6	30.4	42.0	41.0	42.2	39.9
LLaMa 3.1 70B (t1)	76.4	69.4	41.6	32.2	47.6	47.2	38.6	40.0	34.4	35.6	41.6	39.0	55.8	28.4	31.6	34.2	41.4	40.4	39.4
CommandR (Aug) (t4)	60.8	49.0	28.4	27.8	25.8	27.2	28.0	28.6	29.8	28.0	28.0	26.6	29.0	25.8	28.2	28.4	27.2	27.4	27.8
gpt-3.5-turbo-0125 (t1)	72.0	66.6	31.2	34.8	38.0	39.8	34.2	40.0	39.0	38.0	39.8	39.0	52.4	37.8	32.8	38.6	37.2	36.2	38.1
gpt-4o-mini-2024-07-18 (t1)	82.6	78.2	37.2	28.0	51.4	46.2	52.8	44.0	39.4	46.2	53.6	46.8	64.0	38.0	31.6	50.6	48.4	50.0	45.5
gpt-4-turbo-2024-04-09 (t1)	80.4	79.4	44.4	32.4	63.2	61.8	60.4	53.6	49.6	49.4	63.4	63.4	74.4	35.2	35.0	63.4	54.6	63.2	54.2
gpt-4o-2024-08-06 (t1)	87.4	83.2	59.8	33.6	67.2	67.2	64.2	61.0	52.8	61.0	67.6	67.4	77.4	43.2	37.8	70.2	61.2	68.2	60.0
Claude OPUS (t1)	74.6	64.4	57.6	33.6	39.4	43.6	42.2	43.6	41.0	40.4	43.4	47.0	55.0	38.2	33.2	42.6	43.6	43.8	43.0
Gemini-1.5-pro (t1)	82.0	81.8	68.0	39.0	71.2	70.4	65.0	55.2	53.0	55.8	66.8	67.6	78.4	48.2	32.0	69.2	57.4	66.0	60.2
Translate-test																			
mT0-XXL-MT (t1)		32.2	30.0	27.2	28.8	27.4	29.8	28.0	26.0	27.2	29.8	30.4	30.4	25.4	27.0	30.2	29.8	26.4	28.4
Aya-101 (t2)		37.8	32.4	28.6	31.0	31.6	31.8	33.6	27.2	28.6	32.4	34.2	31.2	31.8	26.6	30.8	34.2	32.2	31.1
BLOOMZ 7B (t2)		31.0	28.6	24.4	31.0	24.6	29.0	26.4	27.2	28.6	29.2	28.0	30.6	30.8	24.6	25.2	30.2	27.4	27.9
LLaMa 3 8B (t4)		39.8	34.6	30.6	31.8	31.2	30.8	33.0	29.0	29.2	30.8	34.8	34.2	31.2	28.4	33.0	32.6	33.2	31.8
LLaMa 3.1 8B (t1)		51.0	46.0	36.2	40.0	43.6	43.2	42.8	33.4	39.8	40.8	45.6	48.2	33.8	31.2	44.2	43.6	44.4	41.1
LLaMaX 3 8B (t5)		39.4	37.6	32.4	35.2	37.6	35.0	33.8	33.2	36.4	35.8	41.0	36.2	34.2	26.2	36.2	33.4	38.8	35.2
Gemma 2 9B (t1)		59.4	47.2	41.2	45.2	47.8	46.4	48.0	38.4	44.0	43.8	51.0	54.2	37.2	30.2	49.2	46.2	45.2	44.7
Gemma 2 27B (t1)		62.4	57.6	40.4	50.0	50.0	50.6	47.0	42.8	46.6	49.8	55.6	59.8	39.8	31.2	55.2	52.6	51.6	48.8
LLaMa 3.1 70B (t1)		67.4	55.6	44.8	50.6	55.8	55.6	53.8	46.8	49.4	53.6	59.6	63.0	41.2	32.8	55.4	49.6	52.8	51.3
CommandR (Aug) (t4)		53.6	42.2	38.4	42.2	42.4	44.6	40.4	34.2	38.0	41.4	48.8	46.6	34.2	28.6	44.8	40.6	44.8	40.8
gpt-3.5-turbo-0125 (t1)		61.4	50.4	41.4	49.6	47.8	48.0	44.8	40.4	47.2	51.4	52.8	55.2	41.0	33.4	49.8	47.8	47.6	46.8
gpt-4o-mini-2024-07-18 (t1)		70.0	58.0	40.2	51.6	52.2	54.6	49.6	42.8	50.6	53.8	61.0	61.0	38.2	33.0	55.6	49.4	51.0	50.2
gpt-4-turbo-2024-04-09 (t1)		73.2	60.8	42.0	49.6	57.0	57.6	51.4	45.2	51.8	52.4	59.2	61.8	41.6	33.4	60.2	54.0	54.8	52.1
gpt-4o-2024-08-06 (t1)		76.4	62.8	43.8	54.0	57.4	58.2	54.4	46.4	54.8	55.6	63.2	67.4	44.0	32.0	62.2	52.4	57.4	54.1
Claude OPUS (t1)		63.2	52.4	39.8	46.2	48.8	49.2	46.2	40.2	45.6	48.0	52.4	75.3	37.6	31.0	51.2	47.4	50.2	47.6
Gemini-1.5-pro (t1)		73.4	61.4	43.0	56.0	57.4	55.6	54.6	44.4	53.8	57.0	59.0	66.4	40.6	28.4	57.8	56.0	58.4	53.1

Table 17: **AfriMMLU results for in-language and translate-test**. We make use of the best prompt

Model	eng	fra	vai	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
in-language																				
mT0-XXL-MT (t1)	3.6	4.8	0.8	4.4	1.6	4.8	1.2	3.6	2.0	4.0	0.8	3.2	4.0	3.6	1.2	2.0	4.8	2.0	3.2	2.9
Aya-101 (t1)	10.8	9.6	0.8	6.8	3.2	7.2	3.2	3.6	4.4	2.4	4.8	6.8	10.8	6.0	1.6	1.2	4.4	3.2	3.6	4.6
BLOOMZ 7B (t4)	2.8	2.0	1.2	1.6	2.4	2.8	1.6	1.6	1.6	2.4	1.6	0.8	1.6	0.8	0.8	2.4	1.6	2.0	2.0	1.7
LLaMa 3 8B (t3)	59.2	44.0	1.2	2.0	2.4	6.8	4.0	3.2	2.8	3.6	2.4	2.4	2.8	16.8	1.6	3.2	0.4	4	3.2	3.9
LLaMa 3.1 8B (t4)	56.8	54.0	1.2	4.4	2.4	10.0	4.4	7.2	3.2	7.6	3.6	4.8	6.4	41.2	4.0	2.8	2.4	5.6	5.2	7.2
LLaMaX 3 8B (t4)	15.2	12.0	4.0	3.6	1.6	9.6	5.2	4.4	2.8	5.2	1.6	6.0	3.2	10.8	1.6	3.2	5.6	4.8	7.2	4.8
Gemma 2 9B (t2)	68.8	61.2	0.4	26.4	4.8	35.2	11.6	22.4	11.6	17.6	9.6	26.0	22.0	61.6	9.2	3.6	20.8	13.6	21.2	19.8
Gemma 2 27B (t4) - few-shot	85.6	80.0	3.2	33.6	7.6	49.6	24.0	32.4	18.0	23.6	12.8	35.2	38.4	73.6	12.4	5.6	32.0	22.4	34.4	28.5
LLaMa 3.1 70B (t4) -tt	86.8	76.4	0.4	17.6	8.0	48.8	37.2	26.4	11.2	24.4	10.8	21.2	32.8	68.0	14.4	3.2	18.8	23.6	27.2	24.6
CommandR (Aug) (t3)	74.4	64.4	1.2	4.4	2.4	6.0	2.0	5.2	8.4	5.6	3.2	4.0	6.8	20.8	3.6	2.4	5.6	5.2	4.8	5.7
gpt-3.5-turbo-0125 (t2)	73.6	56.8	3.6	2.0	3.2	10.0	2.8	11.6	7.2	9.2	7.2	9.2	7.6	59.2	4.8	3.2	9.2	8.8	13.6	10.6
gpt-4o-mini-2024-07-18 (t2)	85.2	71.6	2.8	31.6	6.0	56.0	33.6	48.0	25.6	29.2	39.2	44.8	36.8	70.8	15.6	7.6	32.4	45.6	43.6	35.4
gpt-4-turbo-2024-04-09 (t2)	86.0	69.2	2.0	40.4	8.0	64.4	48.4	58.0	38.8	46.0	47.6	58.0	56.4	77.6	10.4	9.6	49.6	56.8	52.4	45.2
gpt-4o-2024-08-06 (t2)	84.0	68.8	4.8	57.6	8.8	64.8	57.6	60.4	51.2	51.6	61.2	58.4	60.8	78.8	31.2	28.0	52.4	62.0	57.2	52.6
Claude OPUS (t1)	59.6	52.8		36.0	22.0	27.6	17.2	28.0	24.0	28.4	20.8	25.2	32.0	42.0	19.2	12.4	24.8	28.8	26.4	25.9
Gemini-1.5-Pro (t5)	82.8	67.2	3.2	67.6	40.8	65.6	63.2	63.6	51.2	54.4	57.2	64.4	61.2	76.4	38.0	12.8	50.8	61.6	57.2	52.3
Translate-test																				
mT0-XXL-MT (t1)		2.4		2.0	2.4	0.8	1.2	3.2	2.4	2.8	3.2	2.0	2.4	3.6	2.0	2.4	3.2	3.6	3.2	2.5
Aya-101 (t1)		8.4		8.4	6.4	7.6	6.0	10.0	6.4	6.8	7.6	6.8	10.4	9.2	6.0	8.4	8.8	8.8	8.0	7.9
BLOOMZ 7B (t4)		1.6		2.8	1.6	1.6	0.8	1.2	2.0	2.4	2.0	2.0	2.0	3.6	1.6	2.4	0.8	2.0	1.2	1.9
LLaMa 3 8B (t3)		52.4		39.2	22.8	35.2	31.6	41.6	35.2	26.0	36.0	36.8	40.0	48.4	22.8	14.4	32.4	35.6	36.0	33.4
LLaMa 3.1 8B (t4)		44.0		34.8	20.8	36.0	24.4	38.4	31.2	24.0	32.0	29.6	36.4	44.4	19.2	14.0	32.8	31.6	32.0	30.1
LLaMaX 3 8B (t4)		10.4		13.2	7.2	10.8	8.4	10.0	9.6	8.4	10.8	8.0	7.6	12.0	4.4	6.0	8.8	11.6	10.0	9.2
Gemma 2 9B (t2)		64.0		46.4	26.0	44.0	37.2	50.0	37.6	30.0	35.6	44.0	49.6	60.0	22.8	14.8	41.6	45.6	45.2	39.4
Gemma 2 27B (t4)		70.8		53.2	30.0	54.0	44.0	55.2	47.2	34.4	46.0	48.0	54.4	69.6	29.2	21.6	48.4	48.4	54.0	46.1
LLaMa 3.1 70B (t4) -tt		73.6		54.8	30.4	48.0	43.2	52.8	48.0	35.6	44.0	46.8	55.2	72.0	26.4	20.4	49.6	48.4	54.0	45.6
CommandR (Aug) (t3)		64.0		44.8	25.6	43.2	36.0	45.2	38.4	29.2	37.2	37.2	47.2	60.0	27.2	14.4	41.2	44.0	42.0	38.3
gpt-3.5-turbo-0125 (t2)		58.8		40.8	22.0	40.0	32.0	44.0	40.4	26.8	38.0	40.4	46.8	55.2	24.4	16.8	39.2	41.6	44.8	37.1
gpt-4o-mini-2024-07-18 (t2)		68.0		49.2	28.4	47.6	36.4	51.6	42.0	30.8	43.2	44.4	52.8	66.8	26.4	18.0	44.4	46.4	48.0	42.3
gpt-4-turbo-2024-04-09 (t2)		73.6		49.6	24.4	49.6	40.8	54.0	44.0	31.2	42.8	45.6	52.8	68.0	28.8	20.0	47.6	47.6	49.2	43.5
gpt-4o-2024-08-06 (t2)		70.0		48.4	23.6	46.8	39.2	51.6	44.8	35.2	42.0	42.4	54.8	68.0	26.4	16.0	45.6	47.2	49.2	42.6
Claude OPUS (t1)		47.2		38.8	23.2	33.2	22.0	41.2	36.0	26.4	32.4	36.0	40.8	48.4	19.6	16.8	35.6	36.4	36.4	32.7
Gemini-1.5-Pro (t5)		70.0		51.6	25.2	48.8	42.8	48.4	43.2	36.0	41.6	44.4	52.8	64.4	28.0	16.4	45.6	45.2	48.8	44.3

Table 18: **AfriMGSM results for in-language and translate-test**. We make use of the best prompt

Table 19: **Cross-lingual transfer results on AfriXNLI**: We fine-tuned various multilingual encoders on English training data, and evaluated on other languages. Best result per language in **bold**

Model	eng	fra	amh	ewe	hau	ibo	kin	lin	lug	orm	sna	sot	swa	twi	wol	xho	yor	zul	ave
XLM-R-large	**90.5**	**84.3**	75.8	37.2	68.2	39.3	40.5	32.8	37.7	59.2	41.2	39.7	74.2	39.0	43.8	63.5	38.5	62.2	49.6
Serengeti	77.3	60.7	54.8	51.3	61.0	56.0	55.2	36.2	55.3	46.3	58.3	55.0	66.2	42.7	43.8	56.3	53.5	57.8	53.1
Afro-XLMR-base	81.5	78.5	71.3	36.3	68.8	59.8	57.7	**37.8**	44.2	56.7	59.8	61.5	67.0	40.8	41.0	61.0	52.2	62.7	54.9
Afro-XLMR-large	86.5	82.3	77.2	39.7	75.2	69.8	64.5	35.3	57.8	69.7	68.0	69.2	**74.3**	39.5	39.8	69.0	61.0	72.2	61.4
Afro-XLMR-76L-large	88.2	83.3	**78.5**	**58.3**	**73.3**	**70.0**	65.8	33.3	**68.0**	**69.3**	**70.8**	**70.8**	73.3	**59.5**	**51.8**	**73.0**	**63.2**	**72.5**	**65.7**
GPT-4o	86.2	78.7	66.7	48.3	69.2	68.2	**66.8**	31.2	67.2	66.2	69.8	68.3	72.5	53.2	49.5	72.5	63.5	70.0	62.7