Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments

Tuka Alhanai[1*], **Adam Kasumovic**[1*], **Mohammad M. Ghassemi**[1*],
Aven Zitzelberger[1], **Jessica M. Lundin**[2], **Guillaume Chabot-Couture**[2]
[1]Ghamut Corporation, MI, USA
[2]Bill and Melinda Gates Foundation, WA, USA
{tuka, adam.kasumovic, ghassemi, aven.zitzelberger}@ghamut.com
{jessica.lundin, Guillaume.Chabot-Couture}@gatesfoundation.org

Abstract

Large Language Models (LLMs) have shown remarkable performance across various tasks, yet significant disparities remain for non-English languages, and especially native African languages. This paper addresses these disparities by creating approximately 1 million human-translated words of new benchmark data in 8 low-resource African languages, covering a population of over 160 million speakers of: Amharic, Bambara, Igbo, Sepedi (Northern Sotho), Shona, Sesotho (Southern Sotho), Setswana, and Tsonga. Our benchmarks are translations of Winogrande and three sections of MMLU: college medicine, clinical knowledge, and virology. Using the translated benchmarks, we report previously unknown performance gaps between state-of-the-art (SOTA) LLMs in English and African languages. Finally, using results from over 400 fine-tuned models, we explore several methods to reduce the LLM performance gap, including high-quality dataset fine-tuning (using an LLM-as-an-Annotator), cross-lingual transfer, and cultural appropriateness adjustments. Key findings include average mono-lingual improvements of 5.6% with fine-tuning (with 5.4% average mono-lingual improvements when using high-quality data over low-quality data), 2.9% average gains from cross-lingual transfer, and a 3.0% out-of-the-box performance boost on culturally appropriate questions. The publicly available benchmarks, translations, and code from this study support further research and development aimed at creating more inclusive and effective language technologies.

Code — https://github.com/InstituteforDiseaseModeling/Bridging-the-Gap-Low-Resource-African-Languages
Extended version — https://arxiv.org/abs/2412.12417

Introduction

For many tasks, Large Language Models (LLMs) perform on-par with or approaching human performance. Furthermore, LLM capabilities are improving: the performance gap between state-of-the-art LLMs (e.g. GPT-4) and humans for many benchmarks is much smaller than the gap between previous LLM generations (e.g. GPT 3.5) and humans (Achiam et al. 2023; Bandarkar et al. 2024; Sakaguchi et al. 2021; Lin et al. 2021; Hendrycks et al. 2021b). Despite impressive advancements, LLMs are significantly less capable when assessed in non-English languages. When assessing LLMs in native African languages, which are predominantly low-resource (Joshi et al. 2020), the gap between human and LLM performance is notable (when known), but generally remains unknown because many standard benchmarks do not exist in native African languages.

The performance discrepancy between LLMs in English and African languages is not just a technical challenge; it is a significant issue of equity. All 2,123 native African languages are low-resource, including the 31 languages with more than 10 million speakers (Hammerstrom 2015; Joshi et al. 2020)[1]. Naturally, lower language resource levels result in poorer-performing LLMs. This is particularly tragic because LLMs are least reliable for language speakers who have the most to gain. Of the world's poor, 66% live in Africa (Galal 2024), 66% of those don't have access to the Internet (ITU 2023), and 80% do not speak English (CIA 2024). Thus, even in the unlikely event that the world's poorest could access the Internet and afford the costs of a state-of-the-art LLM, their ability to read and write English would prevent them from making use of the tools. Ultimately, the differences in LLM performance between languages result in a "rich-get-richer" effect: LLMs are more helpful to (English-speaking) people who are better off, and who may then provide better content to train better LLMs.

One approach to bridging the LLM performance gap is to translate all non-English language queries into English, query an English LLM, and backtranslate the response. This approach is only viable if the cumulative errors from translation and backtranslation are smaller than the errors from using non-English language LLMs alone. Previous studies report that errors from leading machine translation tools (i.e. Google Translate) can be substantial in African languages (Bapna et al. 2022; Benjamin 2019); however, the extent to which translation errors impact the *substantive* reasoning capabilities of multilingual LLMs versus their *style* ("translationese"), remains unclear. Insofar as the *substantive* errors are minimal, machine translation may serve as a workaround

[*]These authors contributed equally.

[1]It was assumed that if a resource level rating in (Joshi et al. 2020) for a language was not given, then it was low-resource.

when: (i) the goal is to leverage an LLM to answer questions based on facts, and (ii) when the LLM being used has limited non-English experience (e.g. Phi 3 (Abdin et al. 2024)).

In summary, existing LLMs under-perform in native African languages, and it is unknown if translation technologies introduce too much error to serve as a viable workaround. Thus, more tools, resources, and studies are needed to help the research community understand (1) how large is the LLM performance gap in African languages and (2) what can be done to close the performance gap, once known?

Aims

Our work has three specific aims, listed below:

- **Aim 1 - Benchmark Translation**: We translate the popular multiple choice reasoning benchmark Winogrande, as well as three clinical sections of MMLU (college medicine, clinical knowledge, and virology) into 8 low-resourced and under-studied African languages (Amharic, Bambara, Igbo, Sepedi, Shona, Sesotho, Setswana, and Tsonga), allowing assessment of the (unknown) capabilities of LLMs in African languages.
- **Aim 2 - Performance Assessment:** We apply several state-of-the-art (SOTA) LLMs to the newly translated benchmarks from Aim 1 and measure the extent of the performance gap between English and each of the African languages on the benchmarks. To assess the viability of machine translation, we compare the performance of SOTA LLMs on machine-translated benchmarks versus human-translated benchmarks. We also assessed the performance gap between culturally appropriate and inappropriate benchmark questions. This quantitative assessment highlights the areas and languages where LLM improvements are most needed.
- **Aim 3 - Performance Enhancement**: We explore various fine-tuning strategies to determine their impact on closing the LLM performance gap in African languages. This includes adjusting the fine-tuning data used based on the data domain, language, data quality, and the volume of training samples. Understanding how fine-tuning characteristics impact LLM performance will inform prospective data collection efforts for the community at large.

The benchmarks, translations, and all the code needed to recreate the results herein may be found online and are made publicly available under the MIT license.

Related Work

Benchmarks in African Languages

Recent advances in natural language processing have seen a growing interest in assessing language modeling performance in African languages. This interest has resulted in benchmarks in several tasks, including language identification with AfroLID (Adebara et al. 2022), machine translation with FLORES-200 (NLLB Team et al. 2024), and natural language inference with XTREME (Hu et al. 2020). More recently, benchmarks have emerged for African languages to assess LLM *reasoning* using multiple choice questions. In particular, the manually curated reading comprehension benchmark Belebele provides the most extensive coverage of African languages (25 languages; 115 languages in total) (Bandarkar et al. 2024). In addition to Belebele, Irokobench provides manual translations of reasoning tasks into 15 African languages (Adelani et al. 2024), while Winogrande-MMLU-Clinical-ZA also provides manual translations of reasoning tasks into 3 African languages (BMGF 2024). Given the lack of reasoning benchmarks available for African languages, particularly benchmarks that have been manually translated or otherwise sourced from human-written text (i.e. not machine-translated or AI-generated), we aim to translate two established reasoning and domain-knowledge benchmarks, Winogrande (Sakaguchi et al. 2021) and MMLU (Hendrycks et al. 2021b,a), into 8 African languages. The translations of these two benchmarks provide valuable additions to existing African language datasets, enabling the proper evaluation of LLMs in African languages in popular LLM evaluation tasks. This effort helps pave the way for developing LLMs that perform as well in African languages as they do in English, as the availability of African language benchmarks allows developers to continually refine and enhance their LLMs for use in African languages.

Impact of Culture on LLM Performance

When utilizing data from different languages, cultural factors have been shown to affect the accuracy, relevance, and sensitivity of LLM output. Researchers have focused on the "toxicity" of model outputs (i.e. the output of racist, violent, or harmful content) and have released several benchmarks to measure toxicity in LLMs (Wen et al. 2023). However, the impact of cultural nuances has been relatively less explored, in which content that in one context (language) is mundane or inoffensive (e.g. *"a child dislikes broccoli"*) may be considered culturally strange, incoherent, or disrespectful in another context (language) (Liu et al. 2021). Cultural nuances have been shown to impact both the quality of translations (Yao et al. 2023) as well as the model's understanding and generation of responses (Putri et al. 2024). Strategies to identify culture-specific references generally require ground-truth labels generated by human annotators, such as a list of offensive language-specific words (NLLB Team et al. 2024), a list of universally acceptable words (Borin, Comrie, and Saxena 2013), or predefined dimensions of culture (Arora, Kaffee, and Augenstein 2023). Word lists identify explicit expressions in the data and thus have the advantage of scaling to large amounts of data; however, word lists do not capture data that are *implicitly* inappropriate. Therefore, several efforts have been made to annotate implicit expressions of cultural (in)appropriateness (Xu et al. 2021; Hartvigsen et al. 2022). In this study, we evaluate the impact of implicit cultural appropriateness on LLM performance (via human annotators), when translating seemingly inoffensive data (i.e. Winogrande) from English. We also share the cultural annotation data to enrich the translated benchmark.

Impact of Fine-tuning on Cross-lingual Transfer

Previous studies report that cross-lingual instruction tuning improves performance by nearly as much as monolingual instruction tuning (Shaham et al. 2024). Given that even the most widely spoken African languages are low-resource, cross-lingual tuning may provide a way to "boost" the amount of data (and thus the performance) of LLMs in low-resource languages (Beukman and Fokam 2023; Whitehouse, Choudhury, and Aji 2023). However, the impact of cross-lingual tuning is unknown for most African languages (Hu et al. 2020; Liang et al. 2020; Conneau et al. 2018). Hence, we also compare the effects of mono- and cross-lingual tuning using the translated benchmarks.

Impact of Data Quality on LLM Performance

The scarcity of data is a significant barrier to improving LLM performance (Villalobos et al. 2024). In an effort to reduce data barriers, automated methods have been presented to identify high *quality* data samples within an existing corpus (Longpre et al. 2024). Recent methods have focused on using LLMs to assess the quality of data (Tan et al. 2024), which includes annotating data based on the robustness of question-answer pairs (QA pairs) to variations in the question (Chen and Mueller 2024), rating data quality on a Likert scale (Zhou et al. 2024), evaluating the correctness of an answer (Cole et al. 2023), or generating a binary "preference" score when selecting between two texts, such as the LLM-as-a-Judge framework (Zheng et al. 2023). In this study, we apply an LLM-as-an-Annotator of data quality across two translated benchmarks in eight African languages.

Methods

Benchmark Translation (Aim 1)

We translated the popular multiple choice reasoning benchmark Winogrande, as well as three clinical sections of MMLU (college medicine, clinical knowledge, and virology) into 8 low-resourced and under-studied African languages (Amharic, Bambara, Igbo, Sepedi, Shona, Sesotho, Setswana, and Tsonga). These benchmarks were selected because they are multiple choice (and thus easy to compare across languages) and widely used (300+ citations/year). A summary of the translation process is described below, and we provide additional information on the procedures, translator profiles, remuneration approach, and other details in Appendix Section A.

Winogrande Translation Process The translation of Winogrande (3,674 QA pairs, 73,742 words) occurred in three steps: (i) a *translator* translated QA-pairs from English to the target language, (ii) an independent *validator* checked each of the translations and corrected any identified translation errors, and (iii) two independent *evaluators* assessed the final quality of the validated/corrected translation.

All individuals were recruited from Upwork.com; we recruited the most qualified available individuals (*translators, validators, and evaluators*) for each language and always paid above the average wage ($7.30 USD) in South Africa (Statistics South Africa 2020). All *translators* and *validators* were allowed to use a machine translation tool, but had to manually validate the output and correct errors. The *evaluators* were presented with the *validators'* translations (which were corrections of the *translators'* translations) and asked to rate the *quality* of the translation according to three options: (i) "Good translation" ("good"), (ii) "Incorrect, but someone could understand the idea" ("understandable"), and (iii) "Completely wrong" ("wrong") (Benjamin 2019; Bapna et al. 2022). To aid the research community, we have made the initial translations, corrections, and evaluations publicly available along with the translated benchmarks. Examples of the translation and validation survey forms are available in Figures A.1 and A.2.

MMLU-Clinical Translation Process Unlike Winogrande, specialized domain-knowledge was required to translate our selected MMLU sections: *virology*[2] (189 QA pairs / 5,452 words), *clinical knowledge* (299 QA pairs / 8,744 words), and *college medicine* (200 QA pairs / 12,911 words); thus, we hired a professional translation firm (Translated.com) to perform the translations of the three MMLU sections. The professional translation firm guaranteed that only translators with prior translation experience as well as subject domain translation experience were used. The translation firm was paid $11,232.29 USD to translate the three MMLU sections over a 15-day period.

Evaluation of LLM Performance (Aim 2)

We applied several state-of-the-art LLMs to the newly translated benchmarks from Aim 1 and measured the performance gap between English and each of the African languages on the benchmarks. We also assessed the performance gap between culturally appropriate and inappropriate benchmark questions in Winogrande. Additional details on the models assessed and the means by which cultural appropriateness labels were generated are provided below.

The LLM Performance Gap We performed evaluations of SOTA LLMs on the translated benchmarks. We also compared LLM performance to English as a reference, as well as pre-existing benchmark translations for Afrikaans, Zulu, and Xhosa. The LLMs evaluated represented models that were either private: gpt-3.5-turbo-1106 (GPT-3.5) (OpenAI 2023), gpt-4-turbo-2024-04-09 (GPT-4) (Achiam et al. 2023), gpt-4o-2024-05-13 (GPT-4o) (OpenAI 2024); public: Llama 3 70B / 8B instruction-tuned (Dubey et al. 2024); small for edge-devices: Phi 3 Mini 4K instruction-tuned (Abdin et al. 2024); or specialized multilingual: Aya 23 (Aryabumi et al. 2024), Aya 101 (Üstün et al. 2024), BLOOMZ 7b1 (Muennighoff et al. 2023). See Appendix B for model hyperparameters. For both translated benchmarks, `Winogrande` (binary choice co-reference resolution task) and the three clinical sections of `MMLU` (multiple choice medical domain knowledge task), 5-shot accuracy was reported, mimicking (Achiam et al. 2023). `Belebele` (multiple choice reading comprehension task) results were also reported to serve as an additional benchmark (which also

[2]The virology section for Afrikaans, Zulu, Xhosa were also translated, extending the work of (BMGF 2024).

covers the languages in this study), and for which 0-shot accuracy was used (Bandarkar et al. 2024). See Figures A.4-A.6 for the evaluation prompts used for each benchmark.

Measuring Impact of Cultural Appropriateness Although our selected MMLU sections have a cross-cultural focus (i.e. health-related), Winogrande may contain a set of questions that could be considered strange, incoherent, or disrespectful in some cultural contexts (e.g. "*Jessica thought Sandstorm was the greatest song ever written but Patricia hated it. [Patricia / Jessica] bought a ticket to the jazz concert.*").

We assessed how LLM performance in African languages changed when assessed on culturally appropriate vs. culturally inappropriate questions. More specifically, we generated annotations of translation *appropriateness* using the same *evaluators* of Winogrande translation *quality* (described earlier). The *evaluators* were provided with the translated QA pair and asked "*For a typical native speaker in a typical conversational context (casual or professional), could the translated sentence be considered strange, incoherent, or disrespectful?"*, with the following options: *(i) "No, the sentence is typical", (ii) "Maybe, I'm not sure", (iii) "Yes, the sentence is strange, incoherent, or disrespectful", or (iv) "I don't understand the sentence*" (Sap et al. 2020). Of the QA pairs labeled as a "good" or "understandable" translation (i.e. QA pairs with at least decent translation *quality*), a QA pair was considered "culturally appropriate" if both *evaluators* labeled it as "No, the sentence is typical"; otherwise, it was considered "culturally inappropriate". To clarify, a "culturally appropriate" QA pair requires labels for both decent translation *quality* and translation *appropriateness*. Following our definition, 74.2% of translated QA pairs were labeled culturally appropriate, while 20.6% of translated QA pairs were labeled culturally inappropriate. An example of the task can be seen in Figure A.3 and the cultural annotation results are shown in Table A.1. In addition, QA pair examples of translation *quality* and *appropriateness* annotation combinations are shown in Tables A.2-A.9[3]. We provide a detailed discussion of the inter-annotator agreements, as well as the Cohen's Kappa and Fleiss' Kappa scores in Section C of the Appendix.

To determine the effect of cultural appropriateness on LLM performance, the GPT-family models were evaluated out-of-the-box on the Winogrande test set, split by the collected human annotations of appropriateness in each language. Additionally, to account for the possibility of the human annotations capturing appropriateness in English and not just the target language, the same splits of data for each language were evaluated in English to be used as a baseline. Then, if we observe a greater lift achieved in the target language than in English for a given split of the data, we can conclude that the human annotations do indeed reflect a difference in culture-specific appropriateness and also impact the LLM's performance.

[3]Languages that use non-Latin characters (Amharic, Bambara, Igbo) cannot be rendered; however, all annotations are available in the GitHub repository associated with this work.

Enhancement of LLM Performance (Aim 3)

After measuring the LLM performance gap in African languages (Aim 2), we explored various fine-tuning strategies to reduce it. This included adjusting the fine-tuning data based on domain (Winogrande vs. MMLU college medicine), language (mono-lingual vs. cross-lingual contexts), data quality (low vs. high), and the volume of training samples (25%, 50%, 75%, 100%). Additional details on the specific experiments follow. All experiments in this section were performed with Llama 3 70B, which was: (i) the best performing open-source model, (ii) was possible to fine-tune, and (iii) was fiscally feasible. Moreover, all experiments in this section used the same fine-tuning prompts, which can be found in Figures A.7-A.8.

Fine-tuning with Varying Languages and Domains We determined the effects of mono- vs. cross-lingual fine-tuning on LLM performance by comparing performance gains of Lllama 3 70B on our benchmarks after tuning. More specifically, we report the mono- and cross-lingual performance of Llama 3 70B after fine-tuning using two of the translated benchmarks (Winogrande small train split, or MMLU "college medicine" section) and evaluating on four test sets (Winogrande test split, MMLU "clinical knowledge" section, MMLU "virology" section, and Belebele).

Fine-tuning with Varying Data Quality and Quantity To determine the effects of data quality and quantity on LLM performance in low-resource settings, we utilized GPT-4o to generate quality scores for each QA pair in our fine-tuning datasets (i.e. LLM-as-an-Annotator). More specifically, for each QA pair in the fine-tuning dataset, GPT-4o was prompted to provide a score of 1 to 10 based on the usefulness of a QA pair for fine-tuning an LLM to improve its performance on a target evaluation benchmark (1 is the least useful, while 10 is the most useful. See Figure A.9). The LLM-as-an-Annotator was run three times to capture potential variability in outputs. The fine-tuning dataset was then divided into tertiles according to the average of three quality scoring runs. The tertile with the highest scores was designated as the "high quality" set, while the tertile with the lowest scores was labeled as the "low quality" set. To evaluate the impact of data volume on fine-tuning LLM performance, the quality sets were randomly sampled at increments of 0% (i.e. no tuning), 25%, 50%, 75%, and 100%.

Results

In this section, we provide results of five experiments that support our three aims: (Aim 1) benchmark translation, (Aim 2) evaluation of LLM performance, and (Aim 3) enhancement of LLM performance. More specifically, the subsections below provide assessments of: (1) benchmark translation fidelity, (2) "out-of-the-box" LLM performance on the translated benchmarks, (3) "out-of-the-box" LLM performance on the culturally "appropriate" vs. "inappropriate" subsets, (4) fine-tuned LLM performance using mono- and cross-lingual data, and (5) fine-tuned LLM performance using varying data quality and quantity.

Benchmark Translation Fidelity

MMLU translations (by Translated.com) took 15 days to complete and cost $11,232.29 USD. Winogrande translation (by Upwork.com *translators*) took between 3 days (Setswana) and 9 days (Sesotho) and had a cumulative cost of $15,126.40 USD. Winogrande translation verification (by Upwork.com *validators*) took between 4 days (Setswana) and 9 days (Shona) and had a cumulative cost of $9,179.37 USD. Winogrande translation assessment surveys for quality and appropriateness (by Upwork.com *evaluators*) took between 3 days (Shona) and 8 days (Bambara) and had a cumulative cost of $4,680.72 USD.

As seen in Figure A.10, corrections to the original translations varied by language: from 4.9% (for Sesotho) to 65.3% (for Shona) of the QA pairs. As seen in Table A.10, 94.7% of the validated / corrected translations were considered a good / understandable translation ("Good translation" or "Incorrect, but someone could understand the idea") by at least one *evaluator*, only 5.3% were considered wrong ("Completely wrong") by either *evaluator*, and a mere 0.2% were considered wrong ("Completely wrong") by both *evaluators*. Thus, we have reason to believe that (while not perfect) the translations are sufficient to measure the LLM performance gap between English and the African languages.

Measuring the LLM Performance Gap

Here, we provide baseline performance results for SOTA LLM models on three benchmarks: `Winogrande`, the three clinical sections of `MMLU`, and `Belebele`. The average 5-shot (0-shot for Belebele) accuracy scores across all languages and benchmarks are reported in Table 1, with English for reference. The best performing LLM (GPT-4o) had a performance gap ranging from 12.0% to 19.9% absolute between English and the average of the 11 African languages. For individual languages, a breakdown of the baseline performance of SOTA LLM models on the same three benchmarks can be seen in Table 2. We observed that among African languages, GPT-4o consistently performed the best in Afrikaans and the worst in Bambara, with performance gaps between the two ranging from 22.9% (MMLU Virology) to 56.1% (Belebele) absolute across benchmarks, suggesting that there is considerable variance in the performance of SOTA models across individual African languages.

Model	**Bele**	**Wino**	**MMLU**		
			CM	CK	Vir.
Baseline Performance (English language)					
GPT-4o	**95.9**	**83.9**	**84.4**	**89.8**	**60.2**
Average Performance (all 11 African languages)					
GPT-4o	76.0	64.8	66.6	70.6	48.2
GPT-4	69.6	60.9	56.2	60.7	46.0
Aya 101	58.4	50.5	35.7	36.1	32.0
Llama 3 70B	41.2	50.6	35.9	40.6	32.3
Aya 23	38.8	51.2	34.3	34.9	28.4
GPT-3.5	36.2	51.2	34.6	37.0	32.8
Llama 3 8B	36.3	50.4	31.9	35.3	27.3
Bloomz 7B	34.2	49.1	28.9	31.0	25.8
Phi 3 3B	32.2	50.7	30.3	32.4	27.8
Random	*25.0*	*50.0*	*25.0*	*25.0*	*25.0*
Performance Gap (English - African languages)					
GPT-4o	19.9	19.1	17.8	19.2	12.0

Table 1: **LLM Performance Gap Between English and African Languages**. The table displays SOTA out-of-the-box model performance averaged across 11 African languages. The best performing model (GPT-4o) yields between 12.0% and 19.9% absolute difference in performance between English and the average of 11 African languages. Bele: Belebele, Wino: Winogrande, CM: College Medicine, CK: Clinical Knowledge, Vir.: Virology

Additionally, we provide baseline performance results for SOTA models when using machine-translated versions of the benchmarks (see Tables A.11-A.15 for a breakdown of performance by language and benchmark). The performance difference between machine-translated queries and directly using an LLM in non-English languages was not always significant. Hence, our findings suggest that native language LLMs may not be needed in some contexts (see Appendix Section D for a more detailed discussion).

In Figure A.11, we present correlations between LLM performance across language pairs. English was the least correlated with the other languages. The seven Bantu languages and Igbo (Volta-Niger) had the highest correlation values (see Figure A.12 for language families). Bambara and Amharic were least correlated with the other languages, reflecting some combination of the different grammar paradigms of the Mande and Semitic language families, the difference in data quality, and (Amharic) script characters.

Impact of "Appropriateness" on Performance Gap

Here, we assess the impact of cultural appropriateness on LLM performance; we evaluated the best-performing model (GPT-4o) out-of-the box on the Winogrande test set, splitting the dataset by the human annotations in each language, and observing the lift obtained for the "appropriate" QA pairs compared to the "inappropriate" QA pairs (see Figure 1). When comparing the performance lift of GPT-4o in English on the same annotations, we see that 7/11 African languages have a higher lift in the target language, indicating that GPT-4o performs better on culturally appropriate QA pairs for those languages (that is, the lift is not due to the question itself being odd, independent of language; see Figure A.14). The performance of GPT-4o on appropriate over inappropriate Winogrande subsets ranged from -0.6% (Xhosa) to +12.6% (Shona), with an average difference of +3.3% across all languages. The performance lift of GPT-4o when using English evaluations as a baseline ranged from -2.3% (Igbo) to +15.6% (Sepedi), with an average difference of +3.0% across all languages. A breakdown of the performance by language can be found in Appendix Tables A.16 to A.19.

	en	af	zu	xh	am	bm	ig	nso	sn	st	tn	ts
	Winogrande											
GPT-4o	83.9	79.7	68.3	65.9	59.4	50.2	60.7	64.1	69.5	67.4	64.7	62.6
GPT-4	83.5	77.0	64.2	62.3	51.0	50.7	58.7	58.8	65.6	63.8	59.9	57.7
GPT-3.5	59.6	55.0	50.8	52.2	51.3	50.4	51.9	50.2	51.6	49.2	51.5	49.6
Llama 3 70B IT	61.2	51.0	50.4	50.8	50.8	50.5	50.5	50.5	50.4	50.4	50.5	50.4
Llama 3 8B IT	52.0	50.7	50.4	50.4	50.3	50.4	50.4	50.4	50.4	50.4	50.4	50.4
Phi 3 Mini 4K IT	64.7	51.6	50.2	51.3	50.2	50.0	51.0	49.7	51.9	49.5	50.9	50.9
Aya 23 35B	68.7	56.5	49.6	51.8	51.3	50.8	50.6	51.4	50.6	50.0	49.8	50.5
Aya 101	49.5	51.1	49.1	51.2	51.2	52.0	50.5	49.0	51.0	50.5	50.8	49.6
BLOOMZ 7b1	48.6	50.3	48.7	48.8	49.3	50.0	48.8	47.9	48.6	49.6	49.2	49.1
	MMLU College Medicine											
GPT-4o	84.4	84.4	72.8	78.0	67.1	38.2	58.4	66.5	76.9	67.1	63.6	60.1
GPT-4	78.0	79.8	61.3	61.3	46.8	30.1	50.9	53.8	68.8	54.9	57.8	53.2
GPT-3.5	63.6	56.1	32.9	37.6	24.3	30.6	28.9	34.1	35.8	32.4	32.9	34.7
Llama 3 70B IT	76.9	68.2	35.8	40.5	36.9	24.3	33.5	27.2	38.2	28.3	34.1	29.5
Llama 3 8B IT	60.1	44.5	31.8	37.0	16.7	30.1	24.9	32.4	38.7	34.7	27.2	32.9
Phi 3 Mini 4K IT	66.5	39.9	30.6	28.9	28.0	31.2	28.9	28.3	28.3	27.7	34.1	30.1
Aya 23 35B	62.4	49.1	35.8	32.9	34.7	28.3	31.8	35.8	32.4	36.4	34.7	24.9
Aya 101	42.8	40.5	35.3	32.9	35.8	31.2	31.2	34.7	42.2	38.2	34.7	35.8
BLOOMZ 7b1	36.4	34.1	30.1	28.3	26.0	28.9	26.6	27.2	30.1	29.5	27.7	29.5
	MMLU Clinical Knowledge											
GPT-4o	89.8	87.2	79.6	78.5	70.6	40.0	63.0	72.5	80.4	69.4	71.3	64.5
GPT-4	84.2	81.9	70.2	68.3	54.0	34.7	54.7	58.1	67.2	64.2	58.9	55.8
GPT-3.5	72.5	62.6	39.2	37.4	30.2	31.3	34.3	33.2	41.1	34.3	30.6	32.8
Llama 3 70B IT	82.3	71.3	39.2	38.5	33.6	32.5	33.6	37.0	43.4	43.8	39.2	34.0
Llama 3 8B IT	69.1	45.3	36.6	34.7	24.9	32.1	36.2	35.5	39.2	39.2	33.2	31.7
Phi 3 Mini 4K IT	70.6	40.8	29.1	30.6	32.1	27.2	33.6	32.8	33.2	30.9	35.8	29.8
Aya 23 35B	69.4	55.1	32.5	33.6	29.1	33.2	32.5	37.4	32.5	34.3	30.9	32.5
Aya 101	45.3	42.6	38.1	35.5	38.9	28.3	34.7	36.6	44.2	37.0	30.6	30.9
BLOOMZ 7b1	44.9	33.2	30.6	33.2	28.3	32.5	26.8	34.7	30.6	32.1	27.2	31.3
	MMLU Virology											
GPT-4o	60.2	55.4	51.2	50.6	51.2	32.5	44.0	50.0	53.0	47.6	48.8	45.8
GPT-4	59.6	59.0	49.4	50.6	45.8	31.9	44.6	46.4	49.4	45.2	44.0	39.8
GPT-3.5	51.8	42.8	30.7	37.3	29.5	36.1	28.3	33.7	31.3	30.1	30.1	30.7
Llama 3 70B IT	53.6	46.4	28.9	27.1	36.1	35.5	34.9	26.5	31.3	29.5	28.3	31.3
Llama 3 8B IT	51.8	38.6	28.9	27.1	28.3	20.5	24.1	27.7	22.3	25.3	27.7	29.5
Phi 3 Mini 4K IT	48.8	30.1	27.1	25.3	27.1	25.3	32.5	27.7	28.3	24.1	25.3	32.5
Aya 23 35B	49.4	42.2	24.7	26.5	39.2	27.7	19.9	28.3	28.3	26.5	25.9	23.5
Aya 101	33.1	34.9	33.7	31.3	36.1	21.7	30.1	35.5	31.3	34.9	33.1	29.5
BLOOMZ 7b1	38.0	26.5	26.5	24.7	21.1	21.1	26.5	24.1	31.3	27.7	27.7	26.5
	Belebele											
GPT-4o	95.9	94.4	79.7	82.8	78.0	38.3	71.3	77.2	81.0	80.0	77.0	76.8
GPT-4	96.1	93.6	75.9	76.3	61.3	37.9	64.4	64.6	77.6	76.3	67.6	70.0
GPT-3.5	86.2	75.7	33.8	31.8	30.0	30.6	29.2	32.3	34.6	32.1	31.7	36.3
Llama 3 70B IT	94.6	84.8	36.7	36.8	35.1	35.6	35.9	37.1	38.1	35.9	36.4	40.7
Llama 3 8B IT	80.0	69.4	32.6	33.0	30.2	32.0	35.7	32.9	33.9	32.2	34.0	33.4
Phi 3 Mini 4K IT	89.2	52.6	28.3	29.1	28.7	31.7	28.8	31.6	29.1	30.6	29.2	34.3
Aya 23 35B	93.6	83.6	33.2	35.6	29.0	34.6	28.7	34.6	37.8	35.6	36.8	37.4
Aya 101	79.4	77.4	59.3	60.7	67.7	40.7	50.7	59.3	57.8	59.9	57.8	51.0
BLOOMZ 7b1	79.0	36.7	35.8	35.0	24.3	31.6	31.2	34.1	36.7	34.7	35.9	39.7

Table 2: **Results of State-of-the-Art Models on Human-Translated Benchmarks**. The table displays SOTA out-of-the-box model performance similar to Table 1 but with the performance for each individual African language shown instead of averaged.

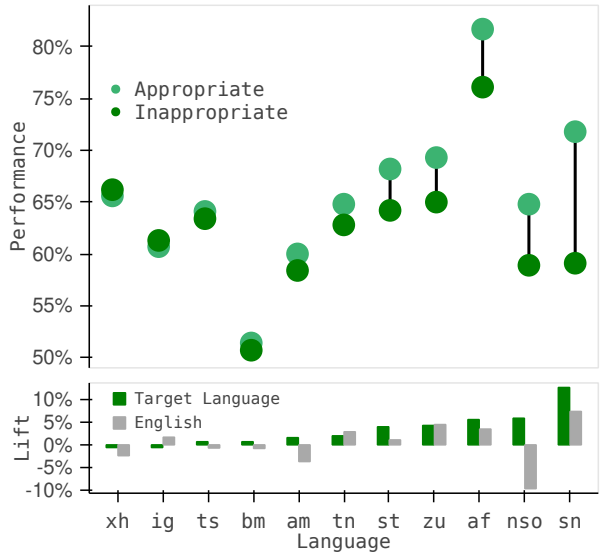

Figure 1: **GPT-4o Winogrande Performance on "appropriate" vs. "inappropriate" Data.** GPT-4o was evaluated on Winogrande (test set) out-of-the-box in each target language and in English. Top plot: the absolute performance on QA pairs considered culturally "appropriate" and "inappropriate" according to native speakers. Bottom plot: performance lifts for each language (green) and in English (grey), using the same annotations. QA Pair was defined as "appropriate" when either annotator marked the cultural appropriateness of the question as "typical". Only QA pairs where both annotators reported that the translation quality was "good" or "understandable" were considered. See Table A.19 for a breakdown of performance by language. See Figure A.13 distributions when repeated random samples of the same size as the appropriate and inappropriate counts for each target language are drawn.

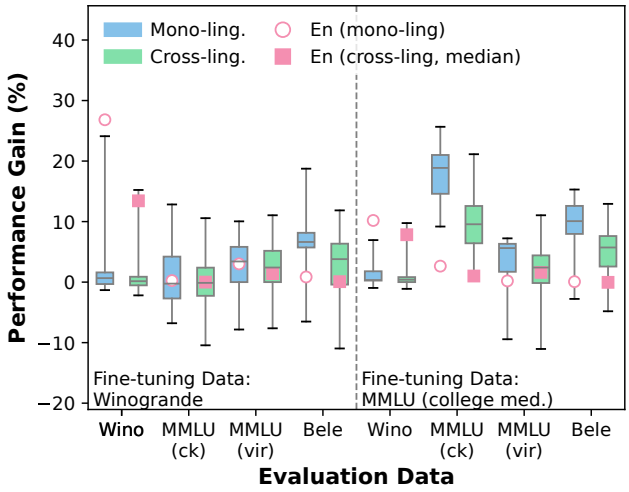

Figure 2: **Mono- and Cross-lingual LLM Performance Gains.** The figure displays boxplots of performance gains when fine-tuning with either the translated Winogrande train set (left) or MMLU college medicine section (right). The fine-tuned models were evaluated across 4 datasests (x-axis) for mono-lingual gains (blue) across 11 African languages, and cross-lingual gains (green) across 110 African language pairs. The most significant gains were with models fine-tuned with MMLU college medicine and evaluated on MMLU clinical knowledge. Wino: Winogrande, ck: clinical knowledge, vir: virology, Bele: Belebele. En: English.

Mono- and Cross-lingual Evaluations

Here, we assess how mono- and cross-lingual fine-tuning impacts LLM performance on our selected benchmarks. The results are shown in Figure 2 (see more detailed breakdowns in Appendix Tables A.20 to A.27). Across the 11 languages, the average mono-lingual gain was 5.6%[4]. Mono-lingual performance gains when fine-tuning with MMLU college medicine were greatest when evaluating on MMLU clinical knowledge (17.4% on average), followed by Belebele (9.4% on average), MMLU virology (3.5% on average), and Winogrande (1.2% on average). When fine-tuning with Winogrande, the greatest gains were observed when evaluating LLMs on Belebele (6.4% on average), followed by Winogrande (2.7% on average), MMLU virology (2.6% on average), and MMLU clinical knowledge (1.2% on average).

Cross-lingual evaluations yielded a similar trend as mono-lingual results; that is, mono-lingual gains (5.6% on average) also yielded corresponding cross-lingual gains (2.9% on average)[5], although to a lesser degree (see Figure 2). At the language-level, when mono-lingual gains for a given language was observed, then most models trained in other languages provided gains (i.e. cross-lingual transfer).

Data Quality and Quantity Evaluations

Here, we evaluate the impact of both the quality and quantity of fine-tuning data on African language LLM performance. Results across the 11 African languages for the highest performing mono-lingual fine-tuning experiments (fine-tuning on MMLU college medicine and evaluating on MMLU clinical knowledge) are shown in Figure 3. There were increasing performance gains with increasing data sizes of 2.3% on average when data sizes were doubled (from 33 to 66 samples), in either quality sets. When data were split between high-quality and low-quality data, the average gain from the data quality split was 5.4% on average. The complete breakdown of results by language for fine-tuning on Winogrande and MMLU, and evaluating on the four benchmarks is listed in Appendix Tables A.28 to A.31.

Discussion

This study aimed to measure (and explore means to address) the performance gap of Large Language Models (LLMs)

[4]Average individual gains of all fine-tuned models above the baselines in the *same* language as the fine-tuning language.

[5]Average individual gains of all fine-tuned models above the baselines in languages *other than* the fine-tuning language.

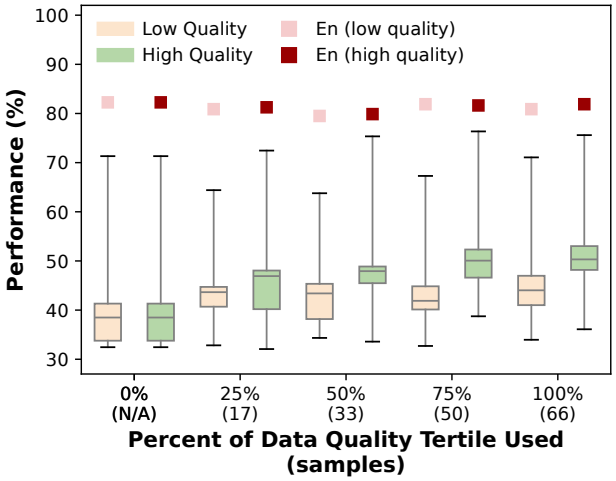

Figure 3: **LLM Performance Across Quality and Quantity Combinations.** The figure displays LLM performance when fine-tuning by data quality and quantity, using MMLU college medicine and evaluating on MMLU clinical knowledge (which had the greatest mono-lingual gains from Figure 2). The quality of samples was rated using GPT-4o LLM-as-an-Annotator scores. The lowest tertile and highest tertile were defined as low (yellow) and high (green) quality samples, respectively, and were used to fine-tune Llama 3 70B IT. Boxplots display performance across 11 African languages. English (En) is provided as a reference (red). Overall, the use of high-quality fine-tuning data over low-quality fine-tuning data improved performance for African languages. See Table A.29 for a breakdown by language.

in English and African languages by translating popular benchmarks, assessing performance on those benchmarks, and exploring fine-tuning strategies that close the gap. The performance gap is not only a technical challenge but also a matter of equity, as many native African languages are low-resource, affecting the accessibility and effectiveness of LLMs for over 160 million speakers[6].

The creation of benchmarks in low-resourced African languages is a critical step toward achieving equitable advancements in natural language processing. By translating popular benchmarks such as Winogrande and sections of the MMLU into eight under-studied African languages, we provide essential tools for evaluating and improving LLM performance in these languages. Our work not only highlights existing performance gaps, but also lays the groundwork for future research and development aimed at improving language technologies for native African language speakers. The benchmarks we translated may enable a more accurate assessment of LLM capabilities and drive progress toward more inclusive and effective language models.

Our study revealed a significant LLM performance gap between English and African languages: 12.0%-19.9% absolute. This performance gap has profound implications, as it exacerbates the digital divide and limits the accessibility and utility of LLMs for millions of speakers of low-resource languages. Addressing this gap is crucial for ensuring that advances in AI benefit all language communities equitably, thereby promoting greater inclusivity.

Moreover, our study also revealed a wide LLM performance gap between individual African languages, with Afrikaans (the highest-performing language) performing between 22.9%-56.1% (absolute) better than Bambara (the lowest-performing language). This disparity is likely due to Afrikaans and English both belonging to the Germanic language family (indicating greater similarity of Afrikaans to English) and Afrikaans having the highest resource level (indicating greater data availability for pre-training LLMs) across all African languages included in our study (Joshi et al. 2020). In contrast, Bambara is part of the Mande language family (see Figure A.12), which does not have a single language that is considered high-resource (or even at the resource level of Afrikaans) (Hammerstrom 2015; Joshi et al. 2020). This implies that Bambara and its relatives have considerably poorer data availability for pre-training LLMs than Afrikaans. Taken together, these findings underscore the importance of measuring the LLM performance gap of individual African languages to guide and prioritize efforts in creating language resources effectively (i.e. by targeting languages with lower performance).

Models showed up to 15.6% better performance on culturally appropriate questions, indicating that cultural relevance significantly influences model accuracy. This underscores the importance of incorporating cultural context in model training and evaluation, particularly for diverse and low-resource languages. Ignoring cultural nuances can lead to biases and inaccuracies, further marginalizing underrepresented language communities.

Fine-tuning was shown to be effective in enhancing model performance. Across all languages and benchmarks, an average improvement of 5.6% was observed for mono-lingual experiments (evaluating in the same language as fine-tuning). The quality of the dataset significantly influences performance, with higher quality data leading to improvements of up to 14.5%, averaging 5.4% over lower quality data of the same size. In addition, the alignment of the training domain with the target domain yielded the strongest gains. For example, fine-tuning a model with *college medicine* data resulted in notable improvements in *clinical knowledge* tasks: 17.4% on average.

In scenarios where target language data are scarce, utilizing data from related languages may help. Cross-lingual transfer methods provided gains of up to 21.1%, with an average improvement of 2.9% across all languages and benchmarks. Our results highlight the potential benefit of using linguistically similar resources to enhance model performance in low-resource languages.

Limitations

Our study has several limitations that should be addressed or extended in future work:

1. **Choice of Fine-tuning Model:** Although Llama 3 was

[6]Ethical Statement of this work is viewable in the Appendix.

the best open-source solution at the time of this study, it does not outperform GPT-4o out-of-the-box, even after fine-tuning. Future studies should consider evaluating other models (e.g. Llama 3.1, released on July 24th, 2024).

2. **Fine-Tuning Scope:** Our fine-tuning experiments were conducted on individual languages in isolation. We did not explore the potential gains from tuning LLMs using data from multiple languages. Future research should investigate the effects of grouping African or related high-resource languages to enhance performance.
3. **Benchmark Relevance:** The translation of established LLM benchmarks, while valuable, may not fully capture the depth and breadth of African-language specific use-cases. Future benchmark creation efforts should consider generating content that directly supports and aligns with use-cases most relevant to African language speakers.
4. **Variability in Translation Quality:** Recruiting experienced translators for all our chosen languages proved difficult. The number of speakers available on Upwork was limited for many languages, most notably Xhosa, Sesotho, Shona, Setswana, Bambara, Sepedi, and Tsonga (see Figure A.15 for more details). In some cases, it was not possible to find workers with prior experience. We also encountered variability in the dialects spoken by the workers, a factor which was challenging to control given the aforementioned lack of available speakers.
5. **Language Coverage and Scalability:** Although we were able to cover 11 diverse African languages, there are naturally many other African languages presumably with LLM performance gaps which are not covered in this work. We believe that our presented framework is scalable, but acknowledge that the most costly part, the human translation of text, is a significant barrier. To alleviate this, we suggest incorporating a step to automatically identify translated text that may require additional human verification, rather than verifying all translations multiple times (which we performed to ensure that translations were high-fidelity). This should make it easier to extend our work to other languages.

Conclusion

This study takes steps toward addressing the performance gap in Large Language Models (LLMs) for African languages. As part of this work, we created approximately 1 million human-translated words of new benchmark data in 8 African languages, covering a population of 160 million speakers. This effort involved the translation of established benchmarks and the fine-tuning of more than 400+ models. The benchmarks, translations, and all the code needed to recreate the results of this study are publicly available, supporting ongoing efforts to improve LLM performance in African languages and beyond. Future work should continue to explore and address the challenges identified, ensuring that advancements in AI benefit all language communities equitably.

Acknowledgments

All research-related costs (experiments, translation costs, etc.) were generously provided by the Bill and Melinda Gates Foundation.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adebara, I.; Elmadany, A.; Abdul-Mageed, M.; and Inciarte, A. 2022. AfroLID: A Neural Language Identification Tool for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1958–1981.

Adelani, D. I.; Ojo, J.; Azime, I. A.; Zhuang, J. Y.; Alabi, J. O.; He, X.; Ochieng, M.; Hooker, S.; Bukula, A.; Lee, E.-S. A.; et al. 2024. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. *arXiv preprint arXiv:2406.03368*.

Arora, A.; Kaffee, L.-a.; and Augenstein, I. 2023. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In Dev, S.; Prabhakaran, V.; Adelani, D.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 114–130. Dubrovnik, Croatia: Association for Computational Linguistics.

Aryabumi, V.; Dang, J.; Talupuru, D.; Dash, S.; Cairuz, D.; Lin, H.; Venkitesh, B.; Smith, M.; Marchisio, K.; Ruder, S.; et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Bandarkar, L.; Liang, D.; Muller, B.; Artetxe, M.; Shukla, S. N.; Husa, D.; Goyal, N.; Krishnan, A.; Zettlemoyer, L.; and Khabsa, M. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 749–775. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Bapna, A.; Caswell, I.; Kreutzer, J.; Firat, O.; van Esch, D.; Siddhant, A.; Niu, M.; Baljekar, P.; Garcia, X.; Macherey, W.; et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Benjamin, M. 2019. Empirical Evaluation of Google Translate across 107 Languages. https://www.teachyoubackwards.com/empirical-evaluation/. Accessed: 2024-07-25.

Beukman, M.; and Fokam, M. 2023. Analysing Cross-Lingual Transfer in Low-Resourced African Named Entity Recognition. 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), 199–224.

BMGF. 2024. Expanding Reasoning Benchmarks in Low-Resourced African Languages: Winogrande and Clinical MMLU in Afrikaans, Xhosa, and Zulu. https://github.com/InstituteforDiseaseModeling/winogrande-mmlu-clinical-za.

Borin, L.; Comrie, B.; and Saxena, A. 2013. The Intercontinental Dictionary Series: A rich and principled database for language comparison.

Chen, J.; and Mueller, J. 2024. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*.

CIA. 2024. The World Factbook — World. https://www.cia.gov/the-world-factbook/countries/world/#people-and-society. Accessed: 2024-07-29.

Cole, J.; Zhang, M.; Gillick, D.; Eisenschlos, J.; Dhingra, B.; and Eisenstein, J. 2023. Selectively Answering Ambiguous Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 530–543.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475. Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Galal, S. 2024. Africa: Share of global poverty by country 2024. https://www.statista.com/statistics/1228553/extreme-poverty-as-share-of-global-population-in-africa-by-country. Accessed: 2024-07-29.

Hammerstrom, H. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *LANGUAGE*, 91(3).

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3309–3326.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021a. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021b. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, 4411–4421. PMLR.

ITU. 2023. Facts and Figures 2023. https://www.itu.int/itu-d/reports/statistics/facts-figures-2023/. Accessed: 2024-07-29.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 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*, 6282. Association for Computational Linguistics.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; et al. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6008–6018.

Lin, X. V.; Mihaylov, T.; Artetxe, M.; Wang, T.; Chen, S.; Simig, D.; Ott, M.; Goyal, N.; Bhosale, S.; Du, J.; et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. arXiv:2109.13238.

Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; et al. 2024. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, and Toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3245–3276.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Le Scao, T.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; et al. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15991–16111.

NLLB Team; et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018): 841.

OpenAI. 2023. GPT-3.5 Turbo fine-tuning and API updates. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/. Accessed: 2024-07-26.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-07-26.

Putri, R. A.; Haznitrama, F. G.; Adhista, D.; and Oh, A. 2024. Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese. arXiv:2402.17302.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5477–5490.

Shaham, U.; Herzig, J.; Aharoni, R.; Szpektor, I.; Tsarfaty, R.; and Eyal, M. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Statistics South Africa. 2020. Labour Market Dynamics in South Africa. Technical report, Statistics South Africa, Private Bag X44, Pretoria 0001.

Tan, Z.; Beigi, A.; Wang, S.; Guo, R.; Bhattacharjee, A.; Jiang, B.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Üstün, A.; Aryabumi, V.; Yong, Z.-X.; Ko, W.-Y.; D'souza, D.; Onilude, G.; Bhandari, N.; Singh, S.; Ooi, H.-L.; Kayid, A.; et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Villalobos, P.; Ho, A.; Sevilla, J.; Besiroglu, T.; Heim, L.; and Hobbhahn, M. 2024. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

Wen, J.; Ke, P.; Sun, H.; Zhang, Z.; Li, C.; Bai, J.; and Huang, M. 2023. Unveiling the Implicit Toxicity in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1322–1338.

Whitehouse, C.; Choudhury, M.; and Aji, A. 2023. LLM-powered Data Augmentation for Enhanced Cross-lingual Performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 671–686.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2950–2968.

Yao, B.; Jiang, M.; Yang, D.; and Hu, J. 2023. Benchmarking LLM-based Machine Translation on Cultural Awareness.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li13, D.; Xing35, E. P.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.