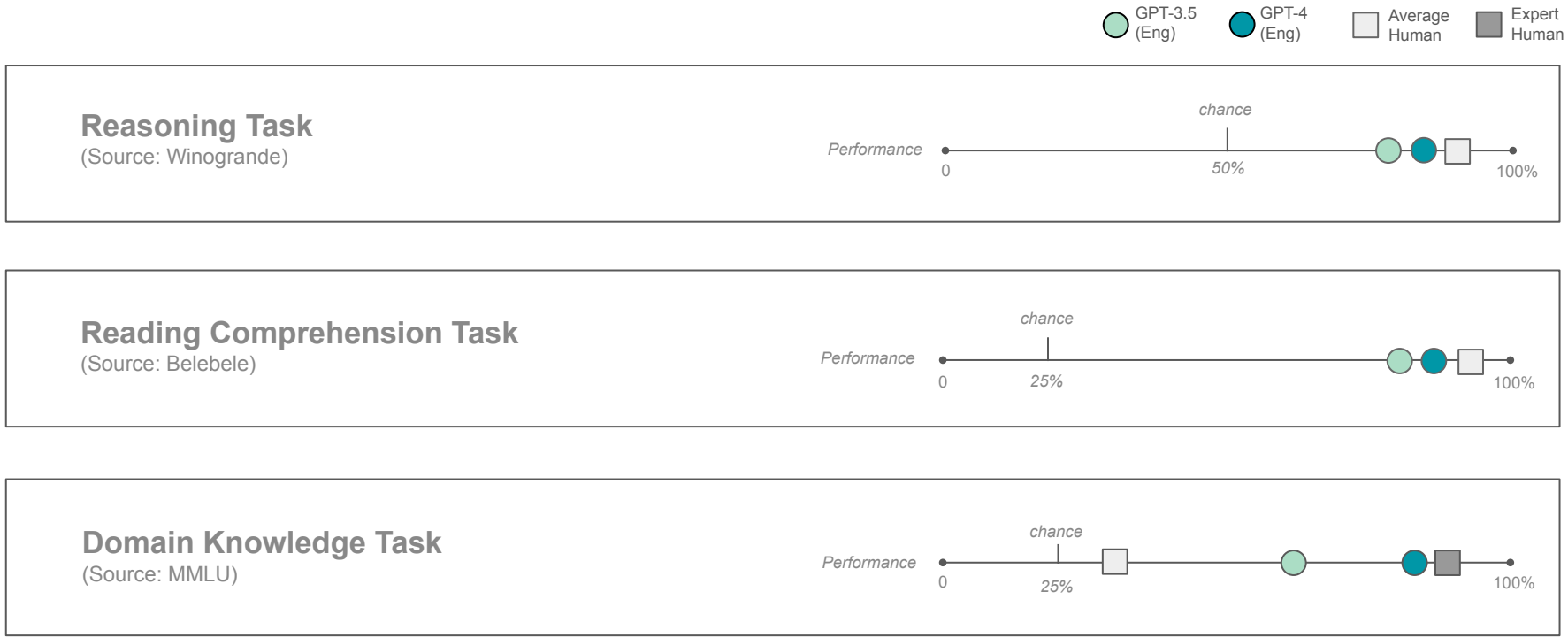
Toward Reducing the LLM Performance Gap for African Languages

Motivation: LLMs perform near human-levels in English, but not other languages

The performance gap diminishes LLM impact for the globally disadvantaged - most of whom don’t speak English.

For many tasks, LLMs perform on-par with or approaching human performance

Furthermore, LLM capabilities are improving: the performance gap between GPT-4 and **humans** is quickly shrinking.



Performance reported out-of-the-box from the literature (GPT-4 Technical Report, Belebele Report, Winogrande Paper, Lin 2022, Hendrycks 2021)

But LLMs are significantly less capable when assessed outside of English

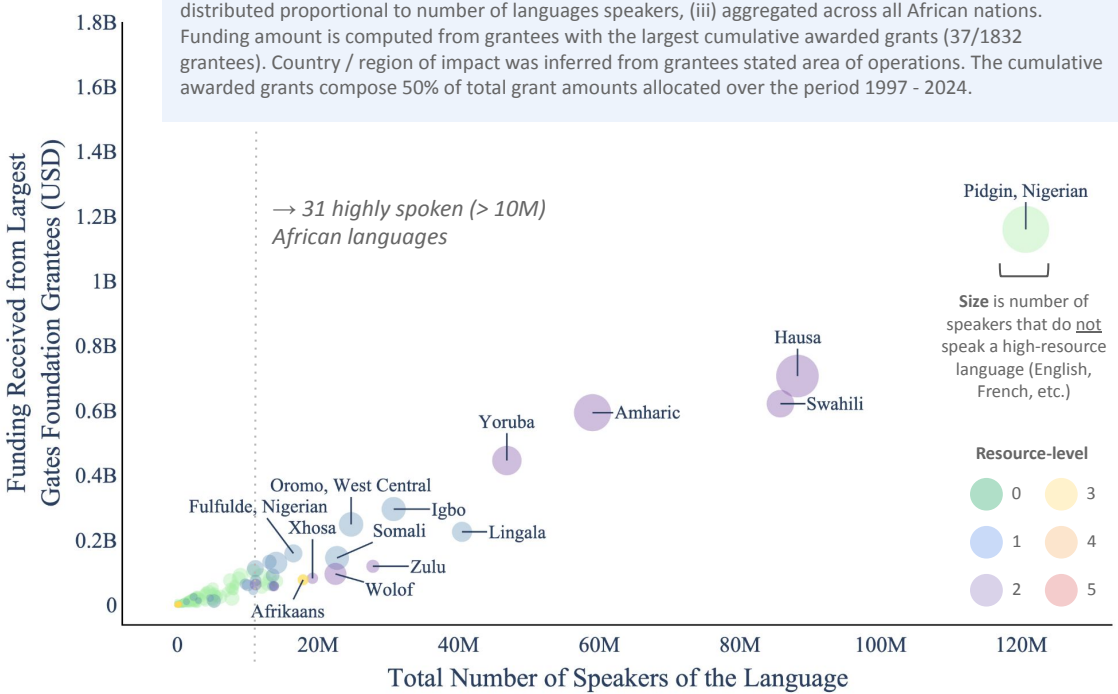
LLMs performance declines on African languages; worse, LLM performance is unknown (?) for standard benchmarks.

Performance reported out-of-the-box from the literature (GPT-4 Technical Report, Belebele Report, Winogrande Paper, Lin 2022, Hendrycks 2021)

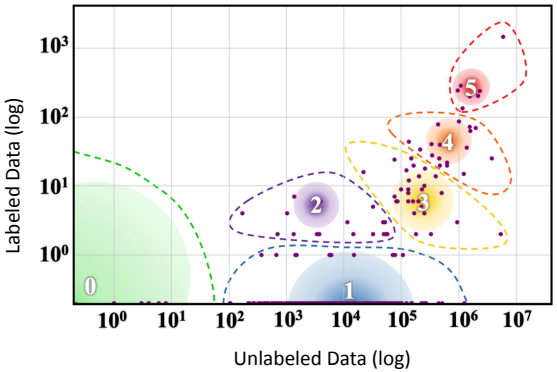
All 2,123 Native African languages are low-resource; 31 have 10M+ speakers

A language is low-resource (LR) when labeled & unlabeled data is limited; limited data → reduced LLM performance.

Method: Estimate uses (i) publicly disclosed BMGF awarded grants assigned for Africa region, (ii) distributed proportional to number of languages speakers, (iii) aggregated across all African nations. Funding amount is computed from grantees with the largest cumulative awarded grants (37/1832 grantees). Country / region of impact was inferred from grantees stated area of operations. The cumulative awarded grants compose 50% of total grant amounts allocated over the period 1997 - 2024.



Language Resource-level Rating

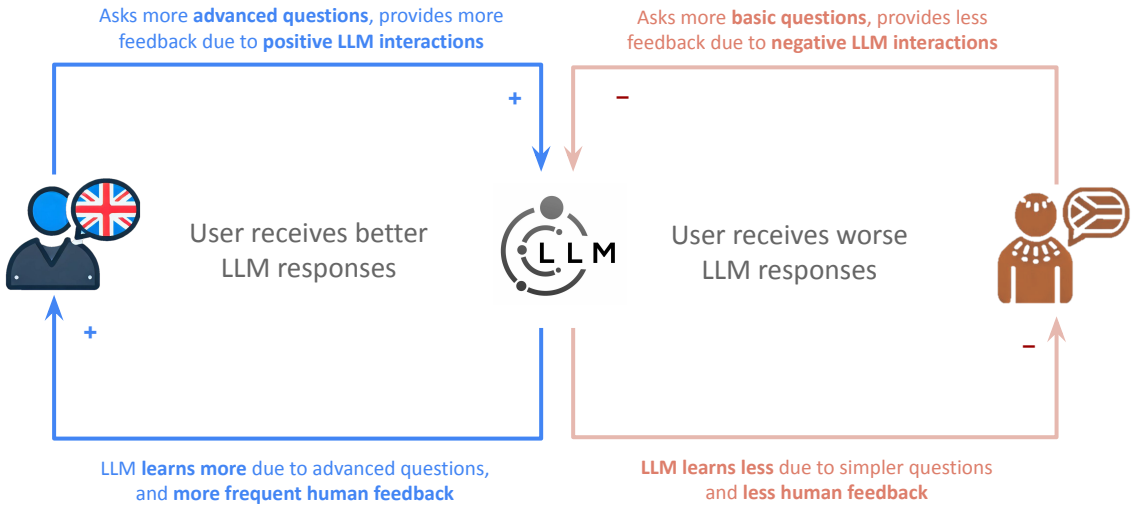


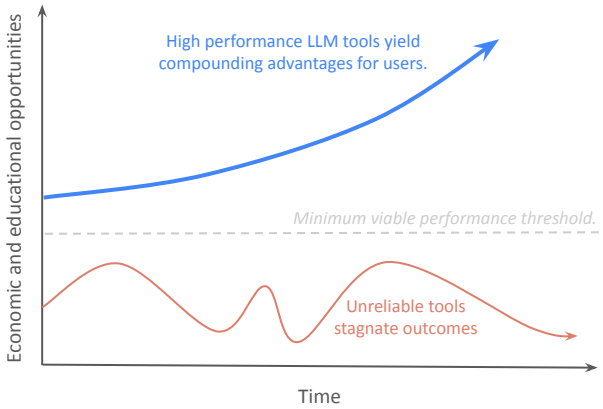
Joshi et al. rated **languages** on a **scale from 0 to 5** based on how much **unlabeled** and **labeled** text **data** they had available on the internet. A **high-resource language** has a rating of **4 or 5** (e.g. Portuguese, English) and a **low-resource language** has a rating of **3 or lower;** all native African languages are within the category of low-resource languages.

Tragically, LLMs are *least reliable* for language speakers that have *most to gain*

Of the world's poor: ~66% live in Africa[1], ~66% of those don't have access to the internet[2], and 80% don't speak English[3]

The differences in LLM performance across languages may lead to a "rich-get-richer" effect:
LLMs are more helpful to people who are better off, and those who are better off may then provide more content to train better LLMs.



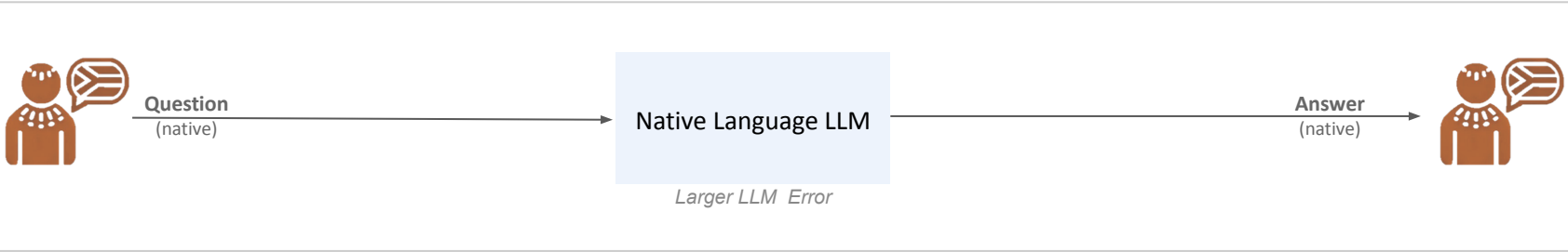


1. Statistica, 2024 2. International Telecommunications Union, 2023 3. CIA World Factbook, 2024

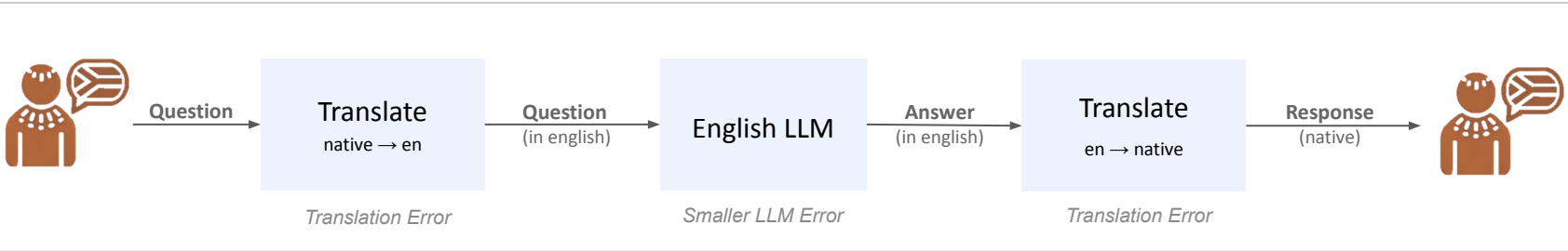
One possible solution: translate into English then query an English LLM

Solution is feasible if errors from translation & backtranslation are less than errors from native language LLMs alone.

Native LLM Option: *Native Language LLM Error*

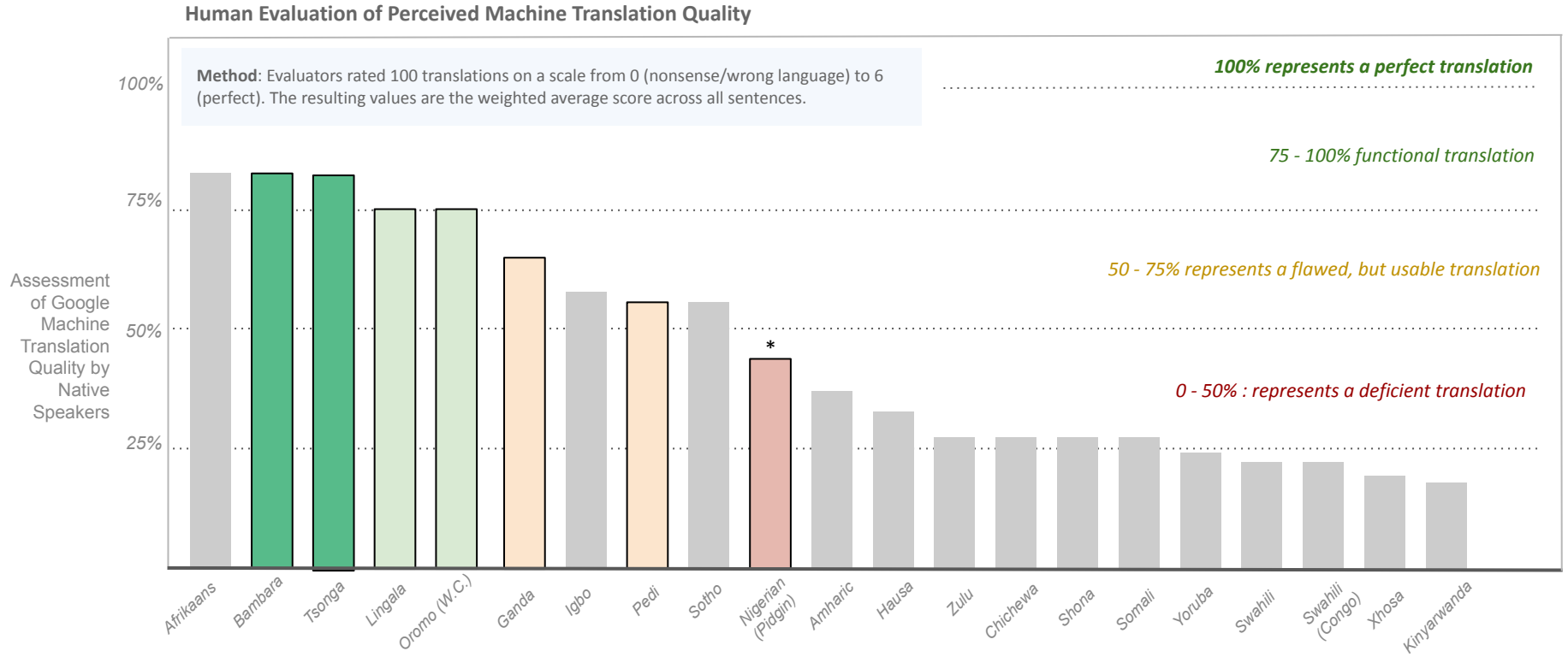


Back Translation Option: *Translation Error* x *English Language LLM Error* x *Backtranslation Error*



Back-translation isn't an option: too many languages (10/31) are unsupported

Of those supported, more than half (11/20) have translation performance less than 50% per human evaluators



(Martin Benjamin, Kamusi Project International, 2018) (Bapna et al. 2022: Table 22)

*Nigerian Pidgin is not supported by Google Translate despite performance being reported.

Question: How large is the LLM performance gap? What can be done to close it?

We focus our investigation on African languages, where the gaps are high when known - but gaps are also unknown.

Our specific aims: create a benchmark, measure the gap, improve the gap

We explored how data domain, quality, quantity and cultural appropriateness impact LLM performance

Specific aims:

1. **Translate two "gold standard" benchmarks** into 11 highly-spoken, low-resource African languages.

2. **Measure the performance gap** between English and the 11 African languages using state-of-the-art LLMs.

3. **Assess how the performance gap is reduced** when fine-tuning LLMs when varying Data:
   a. Cultural appropriateness: Inappropriate vs. appropriate
   b. Domain: Mono-lingual vs. Cross-lingual
   c. Quality: Anticipated utility for fine-tuning
   d. Quantity: Number of fine-tuning samples

*Winogrande and MMLU (Clinical Knowledge, College Medicine and Virology)

Focused on 11 *highly spoken* (>10M) geographically diverse African languages

Languages are spoken in East Africa, West Africa, and Southern Africa covering a total of 230M speakers.

Language	Language Family Group	Number of Speakers (Source: Ethnologue)
Amharic	Afro-Asiatic Semitic	59,037,320
Igbo	Niger-Congo Volta-Niger	30,761,000
Zulu	Niger-Congo Bantu	27,804,600
Xhosa	Niger-Congo Bantu	19,216,300
Afrikaans	Indo-European Germanic	17,886,580
Bambara (Bamanankan)	Niger-Congo Mande	14,054,200
Setswana	Niger-Congo Bantu	13,745,730
Sepedi (Northern Sotho)	Niger-Congo Bantu	13,731,000
Sesotho (Southern Sotho)	Niger-Congo Bantu	13,524,700
Shona	Niger-Congo Bantu	10,783,700
Tsonga	Niger-Congo Bantu	10,003,500
Total		**230,548,630**

11 African languages were the focus of this effort. They are spoken in three geographic regions in Africa (East, West, and Southern) by ~230M people (1/6th of the African population), and represent 5 language groups.

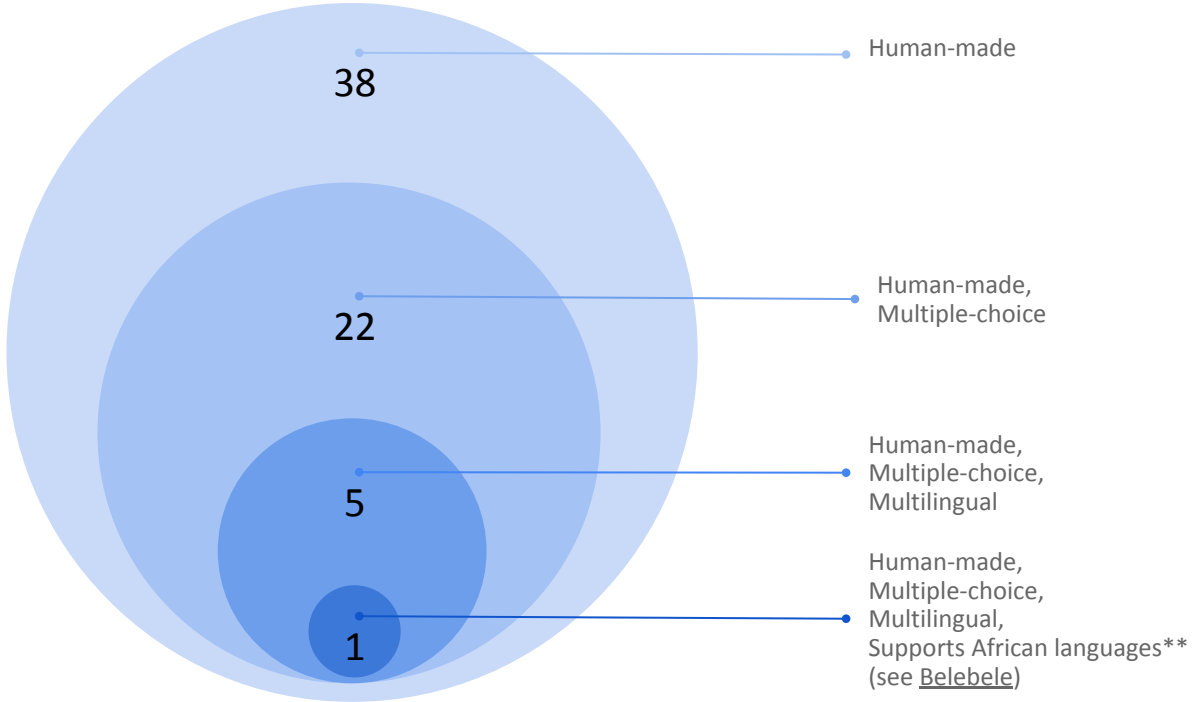
Aim 1: Translate 2 “gold-standard” benchmarks into 11 African Languages (AL)

Benchmarks allow us and the world to assess the (unknown) reasoning capabilities, and health-knowledge of AL LLMs

Motivation: High-quality, multilingual, African language benchmarks are rare

Our review revealed **only one** multiple-choice, multilingual, African language benchmark: Belebele

43 benchmarks were reported in performance assessments of state-of-the-art LLMs* or were publicly available for one of the African languages of interest**.



*Candidate LLMs were BLOOMZ, Mixtral, Mistral, Claude 3, Gemini, Llama 3, GPT-3.5, GPT-4, and PaLM 2. These LLMs are candidates either for LLM-as-a-Judge or as base models for instruction-tuning.

**Zulu, Xhosa, Afrikaans, Pedi (aka Sepedi or Northern Sotho), Tswana (aka Setswana), Sotho (aka Sesotho), Tsonga, Shona, Bambara, Amharic, and Igbo.

We translated *Winogrande*[1] & *MMLU*[2] into our 11 selected African languages

Both benchmarks were (i) multiple choice - for ease of assessment, (ii) widely used - over 300 citations per year.

	Winogrande: Reasoning Task	**MMLU:** Clinical Knowledge Task
What is the benchmark?	A human-generated, human-annotated Q&A benchmark corpus that evaluates a model's general reasoning capabilities with a "fill-in-the-blank" style question and two options.	A collection of multiple-choice Q&A on a diverse set of topics derived from textbooks. The health-focused components were selected since they has the unique advantage of covering a topic (i.e. health) that is broadly and consistently relevant.
What is the data like?	**Example Q**: The children could only count the number of balls in one jar and not the beans in the other because the _ were too numerous. **Ans:** (A) beans, (B) balls.	**Example Q**: Which of the following is true of hepatomegaly? **Ans:** (A) "Emphysema is a cause", (B) "The liver enlarges downwards from the left hypochondrium", (C) "The presence of jaundice, spider naevi and purpura suggest alcohol as a cause", (D)"The liver is usually resonant to percussion".
Who translated it, and why?	Recruited 42 translators from Upwork.com, since no domain knowledge was required. Translators were prioritized based on the following metrics: (1) self-reported proficiency in English, (2) self-reported proficiency in the target language, (3) presence of a "Top Rated" or "Top Rated Plus" badge on Upwork.com, (4) prior job success rate on Upwork.com, (5) prior income earned reported on Upwork.com, and (6) any history of translation jobs in the target language.	A professional translation service (Translated.com) was hired given the specialized domain knowledge required to complete the task. The translation service assigned translators that had prior translation experience as well as subject experience.

1. The "small", version of Winogrande was translated.

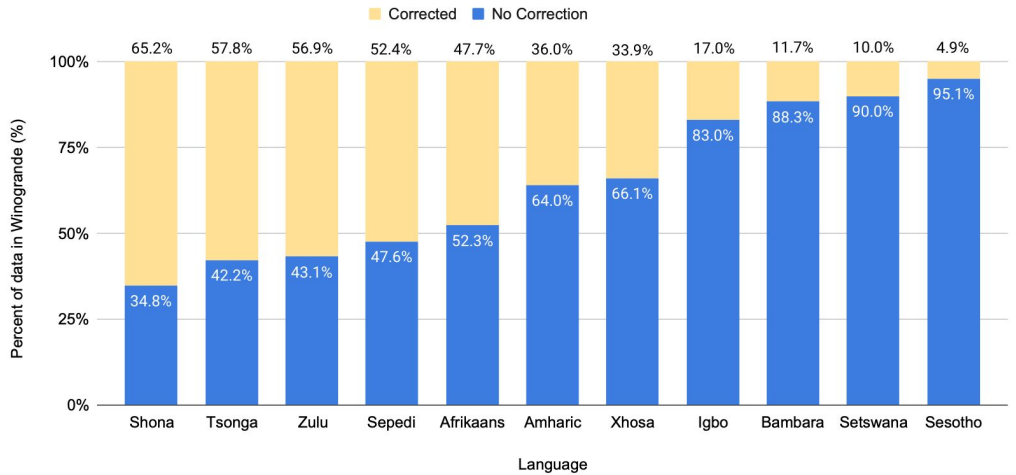
2. "college medicine", "clinical knowledge" and "virology" sections were translated.

We used independent workers for translation and validation of *Winogrande*

~95% of translation were considered “good” by either assessor, ~5% were flagged as “wrong” by either assessor.

Translator: Translated questions from English to target.

Validator: Checked and corrected translation errors.



Two Assessors: Assessed* final translation quality

Either said translation were wrong:

4.7% of validated translations.

Both said translation were wrong:

0.2% of validated translations.

** Assessment Options were multiple choice and included: “Good translation”, “Incorrect, but someone could understand the idea” and “Completely wrong”*

Aim 2: Measure LLM performance gap between English and African Languages

Understanding the extent and reasons for the gap is important to determine how to best close it.

Our benchmarks revealed a sizable LLM performance gap on African languages

The best-in-class LLM (GPT-4o) had a performance gap ranging from 12 - 20% absolute (25 - 53% relative).

Model (parameters)	**Model Type**	**Belebele**	**Winogrande**	**MMLU** (College Medicine)	**MMLU** (Clinical Knowledge)	**MMLU** (Virology)
GPT 4o (Unknown) - English	**Private**	**95.9%**	**83.9%**	**84.4%**	**89.8%**	**60.2%**
				Average Performance (all 11 African languages)		
GPT 4o (Unknown)	Private	76.0%	64.8%	66.6%	70.6%	48.2%
GPT-4 (~1.7T)	Private	69.6%	60.9%	56.2%	60.7%	46.0%
Aya 101 (13B)	Open-source	58.4%	50.5%	35.7%	36.1%	32.0%
Llama 3 IT (70B)	Open-source**	41.2%	50.6%	35.9%	40.6%	32.3%
Aya 23 (35B)	Open-source	38.8%	51.2%	34.3%	34.9%	28.4%
GPT-3.5 (176B)	Private	36.2%	51.2%	34.6%	37.0%	32.8%
Llama 3 IT (8B)	Open-source	36.3%	50.4%	31.9%	35.3%	27.3%
Bloomz (7B)	Open-Source	34.2%	49.1%	28.9%	31.0%	25.8%
Phi-3 4K Mini Instruct (3B)	Open-Source	32.2%	50.7%	30.3%	32.4%	27.8%
Random	-	*25.0%*	*50.0%*	*25.0%*	*25.0%*	*25.0%*
				Performance Gap (absolute difference)		
GPT 4o *(English - African Langs.)*	**-**	**19.9%**	**19.1%**	**17.8%**	**19.2%**	**12.0%**

*** Llama 3 70B IT was the best performing open-source model that was possible to fine-tune (hence why we utilize it for our fine-tuning experiments).*

For comprehensive results tables, see here: https://docs.google.com/spreadsheets/d/1EMEle3ksXcWiht6p082xer_L_LEtcXD1FH6U9dJMf_8/edit?gid=1908518724#gid=1908518724

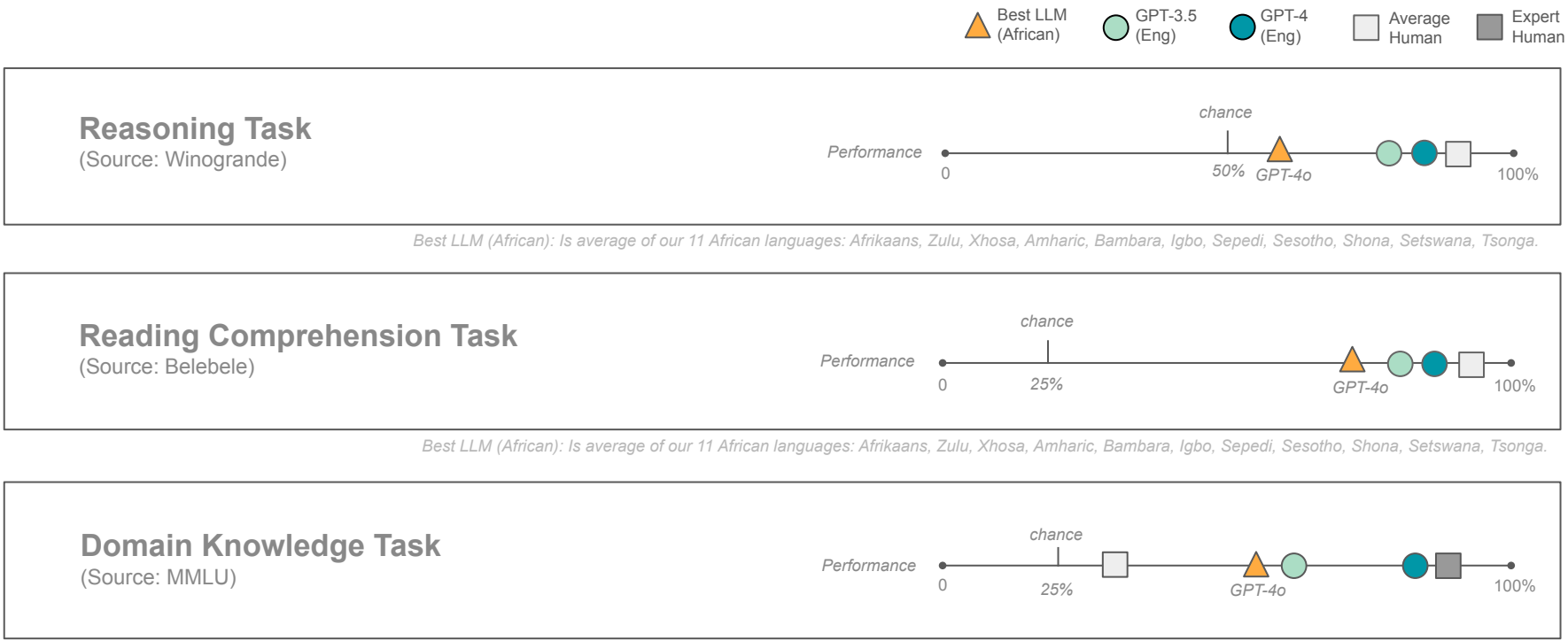
Indeed, LLMs are significantly less capable when assessed outside of English

Our analysis reveals the unknown performance of LLMs on popular evaluation benchmarks.

Performance reported out-of-the-box from the literature (GPT-4 Technical Report, Belebele Report, Winogrande Paper, Lin 2022, Hendrycks 2021)

Indeed, LLMs are significantly less capable when assessed outside of English

Our analysis reveals the unknown performance of LLMs on popular evaluation benchmarks.



Performance reported out-of-the-box from the literature (GPT-4 Technical Report, Belebele Report, Winogrande Paper, Lin 2022, Hendrycks 2021)

Aim 3: Assess how the performance gap is reduced when fine tuning models

Understanding how data characteristics impact LLM performance will inform prospective data collection efforts.

Aim 3 (a): How does cultural appropriateness impact LLM performance?

Aim 3 (a): How does cultural appropriateness impact LLM performance?

LLMs perform up to 13% better (3.3% average) on culturally appropriate questions compared to inappropriate questions.

Two annotators assessed translation quality and cultural "appropriateness"

Data was inappropriate when either annotator marked it as "strange, incoherent, or disrespectful"

Step 1: what is the quality of the translation?

- Good translation
- Incorrect, but someone could understand the idea
- Completely wrong

Step 2: is it **strange**, **incoherent,** or **disrespectful**?

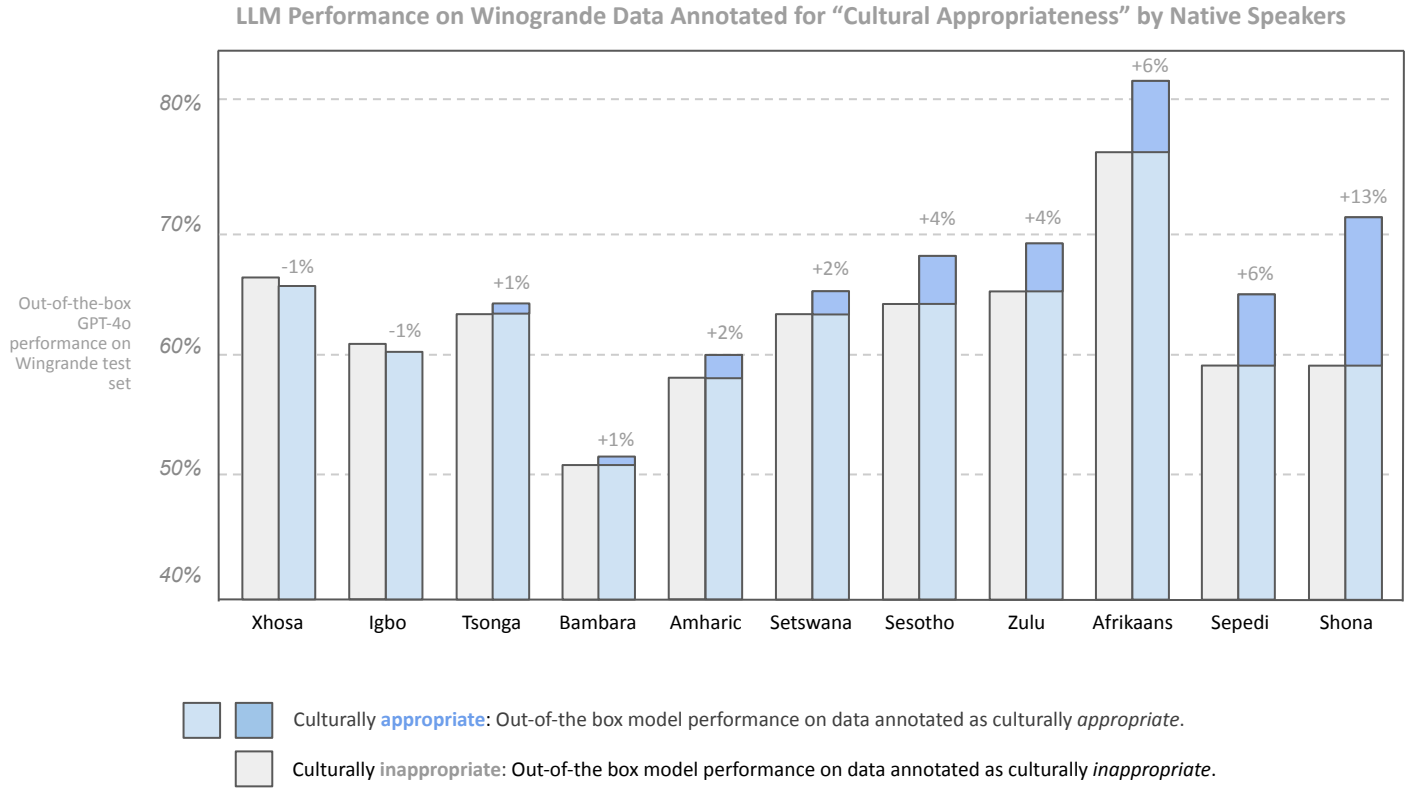
- No, the sentence is typical
- Maybe, I don't know
- Yes, the sentence is strange, incoherent, or disrespectful

Step 3: Data was "inappropriate" when (i) **EITHER** annotator marked it as "strange, incoherent, or disrespectful", and (ii) **BOTH** annotators said "good translation".

Annotator 1			Annotator 2			Result
quality	*inappropriate?*		*quality*	*inappropriate?*		
Good translation	No, the sentence is typical	+	Good translation	No, the sentence is typical	=	Appropriate
Good translation	Yes, the sentence is strange, incoherent, or disrespectful	+	Good translation	No, the sentence is typical	=	Inappropriate
Completely wrong	Maybe, I don't know	+	Good translation	Yes, the sentence is strange, incoherent, or disrespectful	=	Inconclusive

Cultural appropriateness affects LLM performance across languages differently

GPT-4o performance lift on appropriate vs. inappropriate Winogrande items ranged from -1% (Xhosa) to +13% (Shona)



Method: Two annotators were presented with English question-answer text of a reasoning benchmark (Winogrande) and corresponding translations, in their native language. They were asked if the translated sentence could be considered "strange, incoherent, or disrespectful" (i.e. culturally inappropriate).

Model: GPT-4o out-of-the-box was used to evaluate the test set, split by "appropriate" and "inappropriate" annotations.

Observation: Performance lift was observed across most languages (9 / 11) of at least 1.0% and up to 13%. Average lift was 3.3% across all languages combined.

For more detail, please see appendix slides.

All numbers are rounded to the nearest whole integer

Aim 3b: How does cross-lingual instruction tuning impact LLM performance?

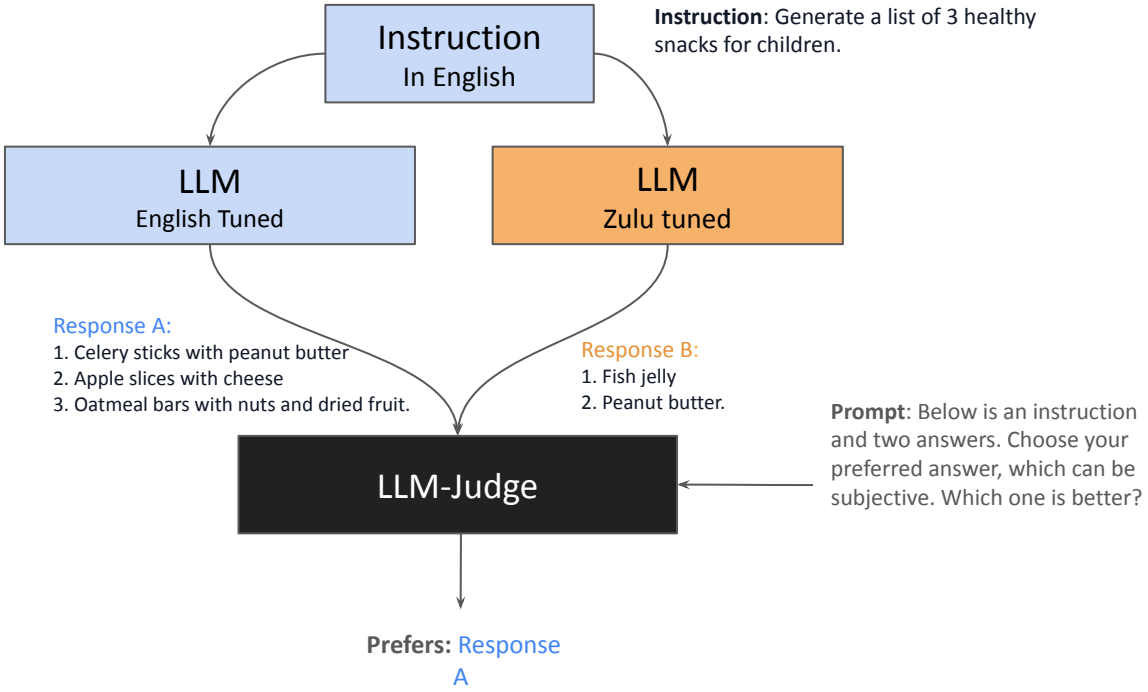
Aim 3b: How does cross-lingual instruction tuning impact LLM performance?

Across all languages & benchmarks: average mono- and cross-lingual gains of 5.6% and 2.9% were observed, respectively.

1. Shaham et. al 2024

Authors at Google[1]: Used LLM-as-a-Judge to assess cross-lingual capabilities

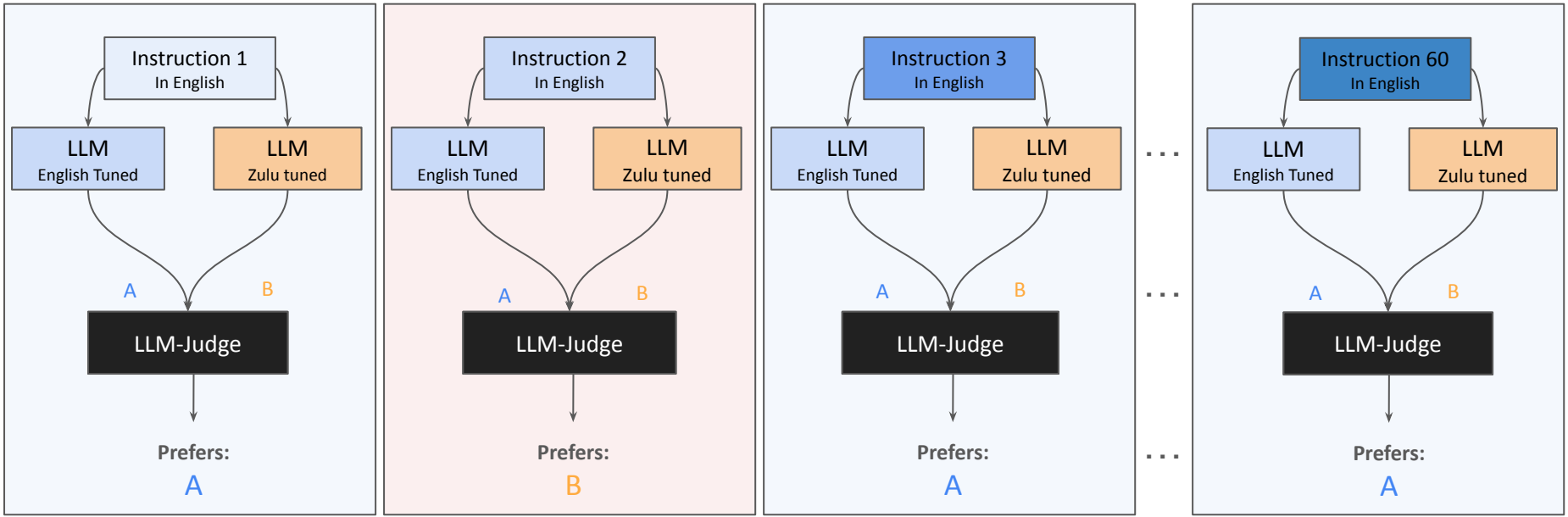
LLM-as-a-Judge is considered attractive because it can be scalably deployed to assess dynamically-generated data.



For a more detailed discussion on LLM as a judge, see Zheng et al. 2023

Example: LLM-Judge compares English instruction-following skills of two models

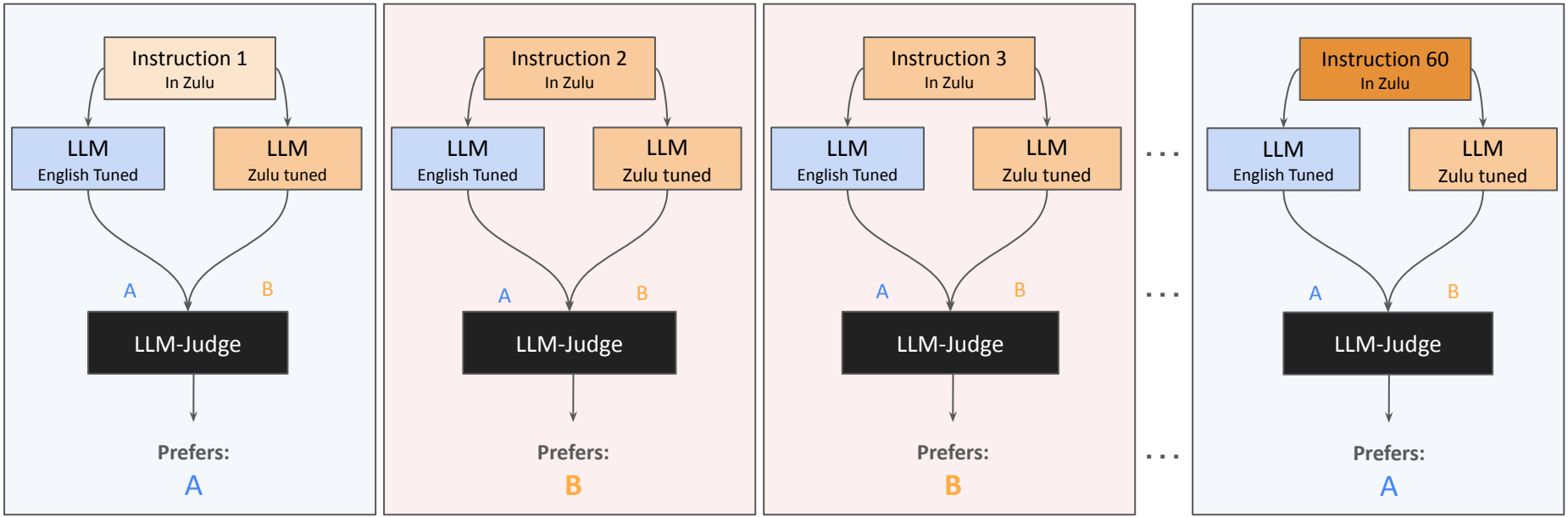
Naturally, we expect an English-tuned model to be strongly preferred to model trained in another language (e.g. Zulu)



When comparing performance on English Instructions, the LLM judge preferred an English-tuned model 55 / 60 times.

Example: LLM-Judge compares Zulu instruction-following skills of two models

If an English-tuned model was preferred as much or less than one tuned in another language → cross-lingual transfer.



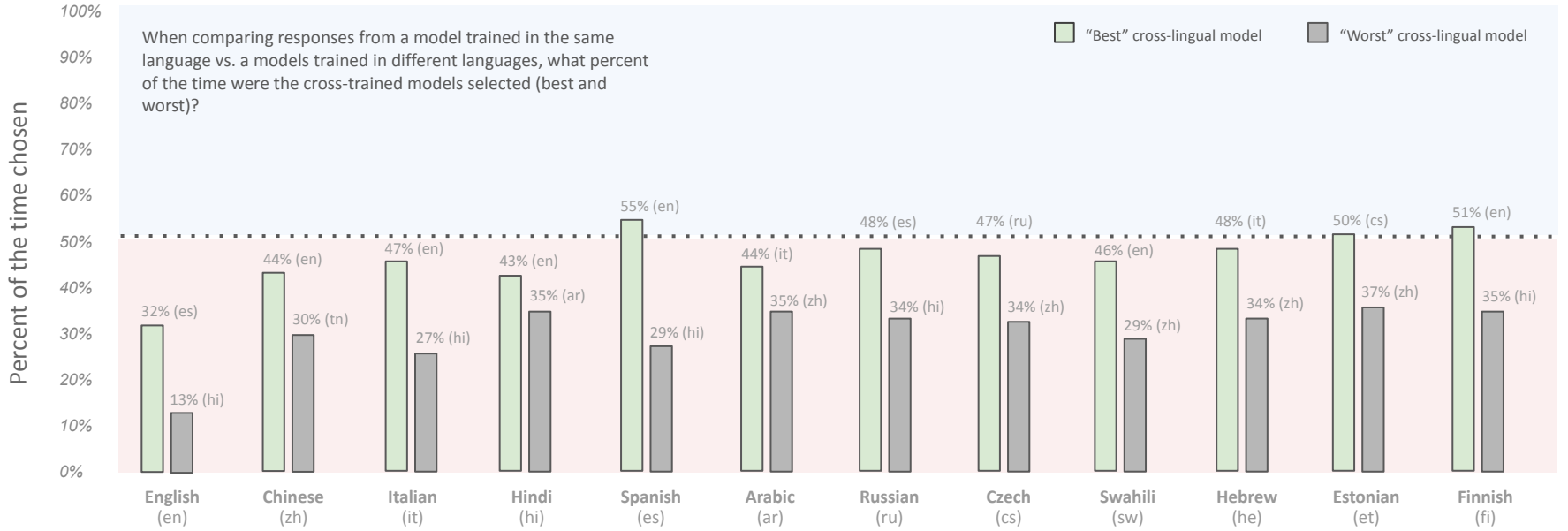
When comparing performance on Zulu instructions, the LLM judge preferred an English-tuned model 30 / 60 times.

1. Shaham et. al 2024

Authors at Google[1]: Cross-lingual instruction tuning improved LLM performance

Models tuned in other languages were near indistinguishable from models in the eval language by LLM judge

Method: Reported in Google[1]. Palm-2 S was fine-tuned using 4,640 samples from a "chatbot" dataset (LIMA/OpenAssistant) that were machine translated. From the 11 trained models, query responses were then generated in the evaluation language (using AlpacaFarm dataset) and the best response (between model pairs) was selected by an LLM-judge (Palm-2 L). *Note that Palm-2 models are not publicly available.*



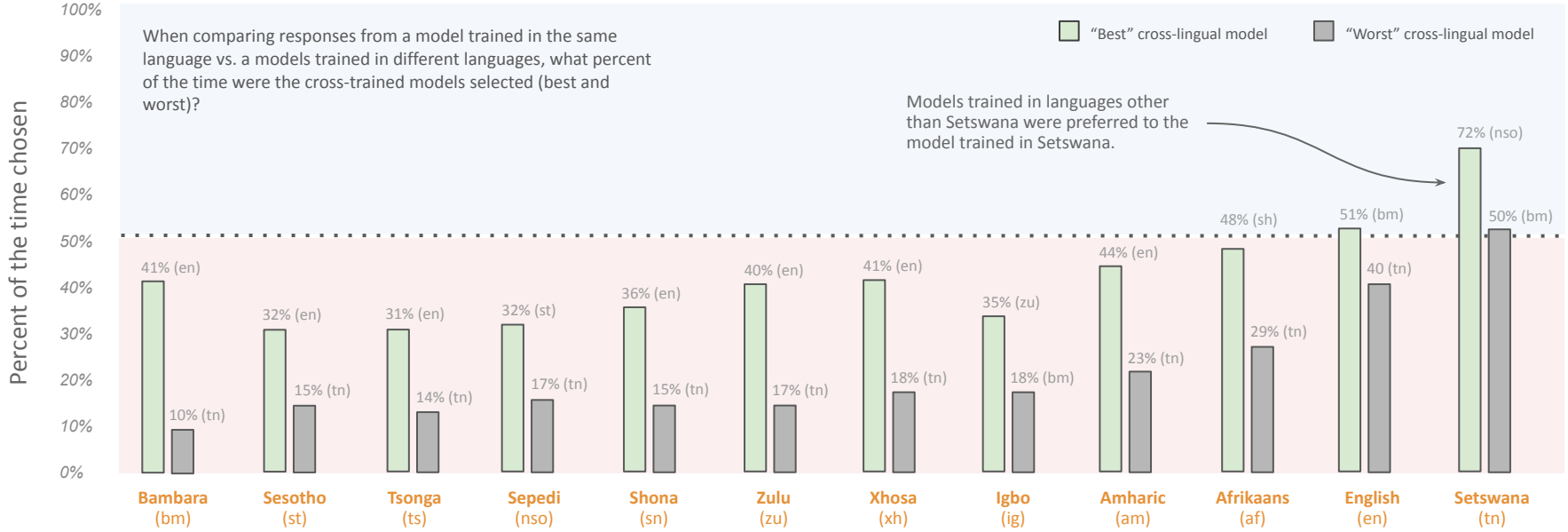
All numbers are rounded to the nearest whole integer

1. Shaham et. al 2024

We recreated Google's[1] results on our 11 African langs. and found similar results

Models tuned in other languages were often deemed useful to models tuned in the eval language by an LLM judge

Method: Based on Google[1], GPT-3.5 models were fine-tuned using 4,640 samples from a "chatbot" dataset (LIMA/OpenAssistant) that were machine translated into the African languages. From the 11 trained models, query responses were then generated in the evaluation language (using AlpacaFarm dataset) and the best response (between model pairs) was selected by an LLM-judge (GPT-4o).



For comprehensive results tables, see here.

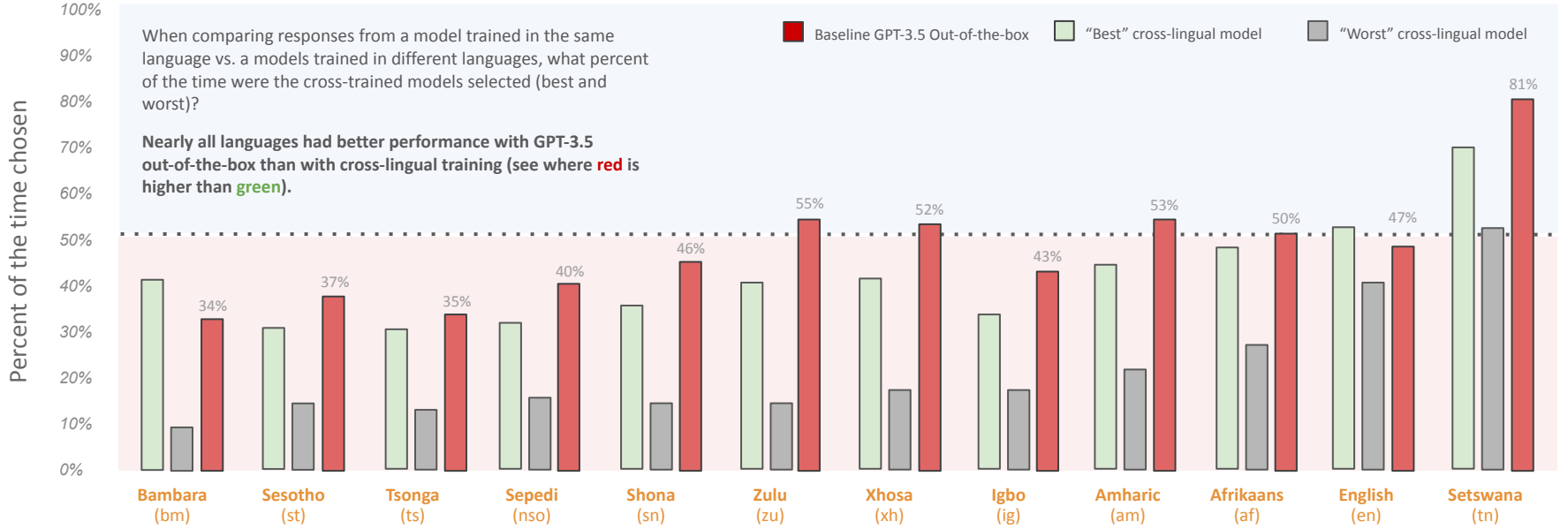
All numbers are rounded to the nearest whole integer

1. Shaham et. al 2024

But the results are less impressive when an appropriate baselines are included

LLM-as-a-Judge (GPT 4o) overwhelming prefered out-of-the-box models over cross-lingually tuned models.

Method: Based on Google[1], GPT-3.5 models were fine-tuned using 4,640 samples from a "chatbot" dataset (LIMA/OpenAssistant) that were machine translated into the African languages. From the 11 trained models, query responses were then generated in the evaluation language (using AlpacaFarm dataset) and the best response (between model pairs) was selected by an LLM-judge (GPT-4o).



For comprehensive results tables, see here.

All numbers are rounded to the nearest whole integer

Possible Reason: LLM-as-a-judge is not “ground-truth”; it’s an LLM’s opinion

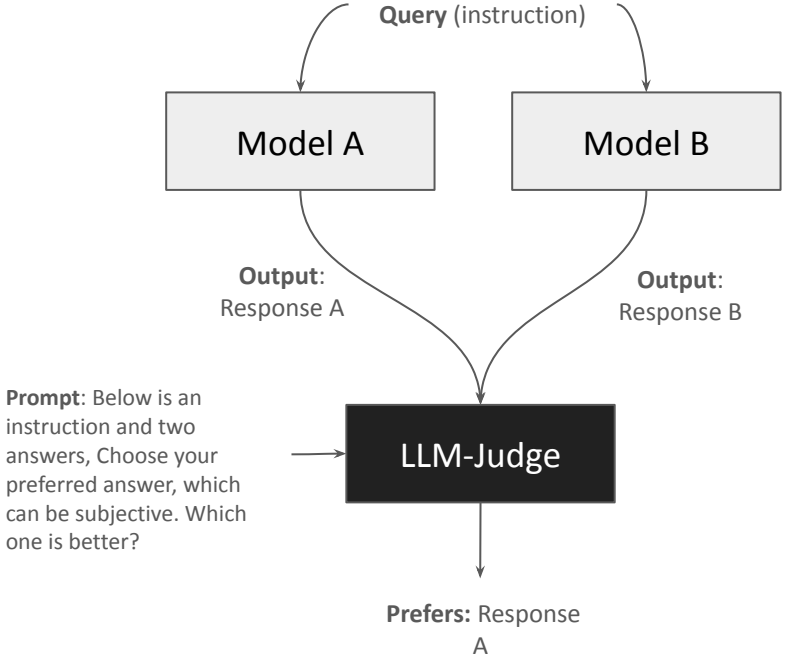
We don’t know if cross-lingual instruction didn’t help, or if LLM judges simply prefer out-of-the-box models.

Advantages of LLM-as-a-judge**:**

- Scalable: LLM is fast, automated, flexible and inexpensive (relative to humans).

Limitations of LLM-as-a-judge**:**

- Relative measure: LLM-as-a-judge does not indicate the absolute quality of the LLM only the relative quality (i.e. LLM-as-a-judge could assess between two equally bad or equally good models).
- Requires pre-training language exposure: The LLM’s judgement may be deficient in domains where pre-training was limited (e.g. Highly Spoken Low-Resource African Languages).



For a more detailed discussion, see Zheng et al. 2023

Our Solution: Remove subjectivity from the assessment by using benchmarks

The multiple-choice benchmarks we created, overcome the limitations of LLM-as-a-judge.

	Winogrande: Reasoning Task	**MMLU:** Clinical Knowledge Task
What is the data like?	**Example Q**: The children could only count the number of balls in one jar and not the beans in the other because the _ were too numerous. **Ans:** (A) beans, (B) balls.	**Example Q**: Which of the following is true of hepatomegaly? **Ans:** (A) "Emphysema is a cause", (B) "The liver enlarges downwards from the left hypochondrium", (C) "The presence of jaundice, spider naevi and purpura suggest alcohol as a cause", (D)"The liver is usually resonant to percussion".

	MMLU: Virology Task	**Belebele:** Reading Comprehension Task
What is the data like?	**Example Q:** Why are parvoviruses a highly impactful parasite? **Ans:** (A) Because they have no nucleic acid, (B) They require a helper virus, (C) Only replicate in dividing cells, (D) Can integrate into host chromosomes.	**Example Q:** Asynchronous communication … allows students … to work at their own pace … flexible working hours … access to information at all times. Which of the following is not a benefit of asynchronous communication for students? **Ans:** (A) The use of internet as a resource, (B) Face-to-face access to instructors at any time of day, (C) Flexible working hours, (D) Pace control.

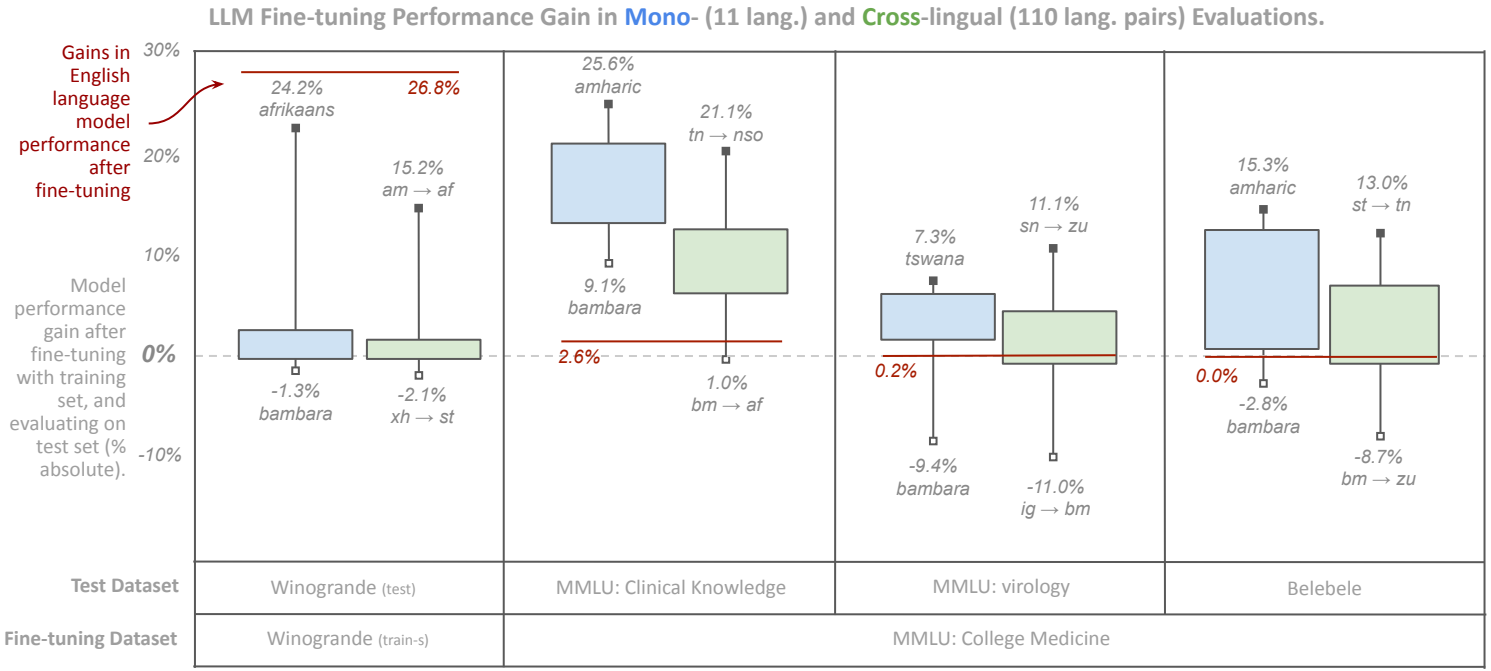
Notice: All experiments from here onward use Llama3 70B IT

The results presented from here onward may be different if another model is used.

Although Llama 3 70B IT did not have the best out-of-the-box performance, it was selected because it: (i) was the best performing *open-source* model, (ii) was possible to fine-tune, and (iii) was fiscally feasible given the total number of experiments needed.

The benchmarks demonstrate that cross-lingual tuning **does** improve models

Across all languages and benchmarks, we saw average mono- and cross-lingual gains of 5.6% & 2.9% respectively



Training: Llama 3 70B IT was used for fine-tuning each language separately. A given fine-tuned model was evaluated on the same language (i.e. mono-lingual) or across all other languages (i.e. cross-lingual) to assess performance gains.

Observation: Mono- and cross-lingual gains are generally observed (> 0%). Greatest gains are observed within closely matching domains; fine-tuning with MMLU College Medicine and evaluating on MMLU Clinical Knowledge yielded at least 9.1% mono-lingual gain, and at least 1.0% cross-lingual gain.

For comprehensive results tables, see here.

Aim 3 (c, d): How does data quantity and quality impact LLM performance?

Aim 3 (c, d): How does data quantity and quality impact LLM performance?

When data is scarce (as is the case in African Languages), LLM performance is more sensitive to quality than quantity.

We used an “LLM-as-an-Annotator” to score the quality of data for fine-tuning

A variation of “LLM-as-a-Judge”, “LLM-as-an-Annotator” provides a scalable method for scoring quality of data samples

STEP 1: A data sample (question-answer pair) from the fine-tuning dataset is passed to the LLM-as-an-Annotater.

STEP 2: LLM-as-an-Annotator is prompted to score the usefulness of the data sample for fine-tuning on a target task, on a scale of 1-10.

STEP 3: With the resulting scores across all data samples, the top third are labeled as high quality, while the bottom third are labeled low quality; data split is used to evaluate LLMs.

Question: The complete resynthesis of phosphocreatine after very high intensity exercise normally takes:

Choices:
(A) 10 seconds,
(B) 30 seconds,
(C) 1 minute,
(D) 4 minutes. (Answer: D)

Prompt: Rate the following fine-tuning dataset sample on a scale of 1-10.

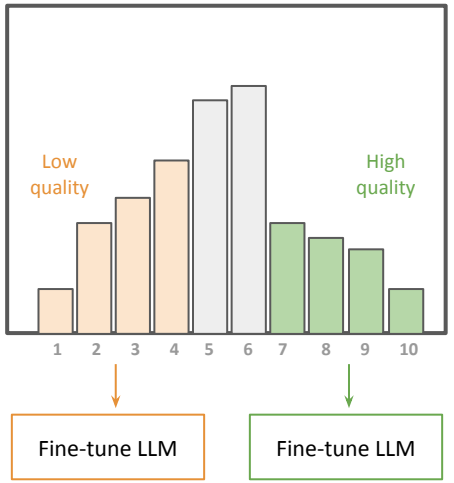
Your rating should reflect the anticipated usefulness of the sample if included in the fine-tuning dataset for improving my LLM's performance on the clinical knowledge section of MMLU (with 1 implying least useful and 10 implying most useful).

{Question-Answer Data Sample}

Question-Answer
In English

LLM-as-an-Annotator

Score: 7 (out of 10)



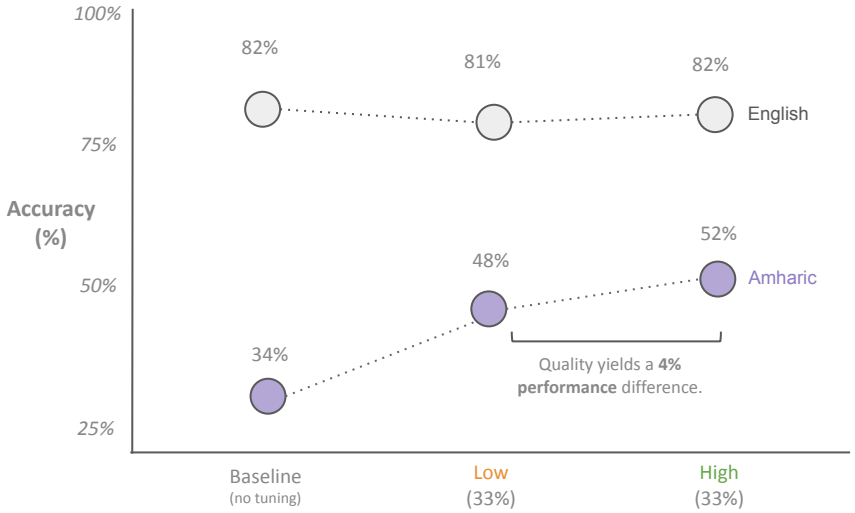
LLM models fine-tuned on different data split by quality are compared to evaluate impact on LLM performance.

Zheng et al. 2023

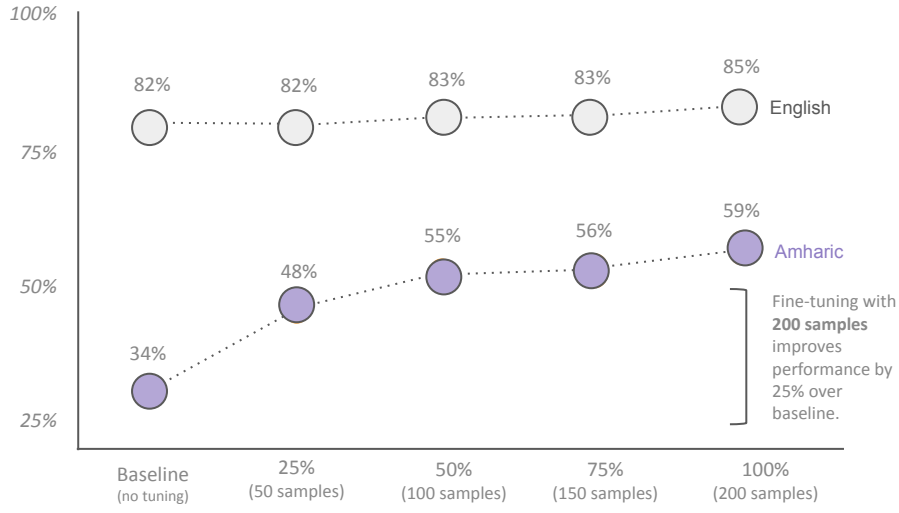
We found that both data quality, as well as quantity, influence performance

In Amharic: 200 samples improves performance 25%; high quality data improves perf. by 4% over low quality data

Data Quality: GPT-4o rated each MMLU college medicine sample in Amharic for its usefulness in improving MMLU clinical knowledge in Amharic, scoring 1-10. Samples were divided into "Low quality" (bottom third) and "High quality" (top third). Llama 3 70B IT was fine-tuned on these datasets, and its performance on MMLU clinical knowledge in Amharic is shown below.



Data Quantity: We took the MMLU college medicine section in Amharic (200 samples) and randomly sampled 25%, 50%, 75%, and 100% (the full dataset) to create four datasets. Llama 3 70B IT was then fine-tuned on these datasets, and its performance on MMLU clinical knowledge in Amharic is provided below.

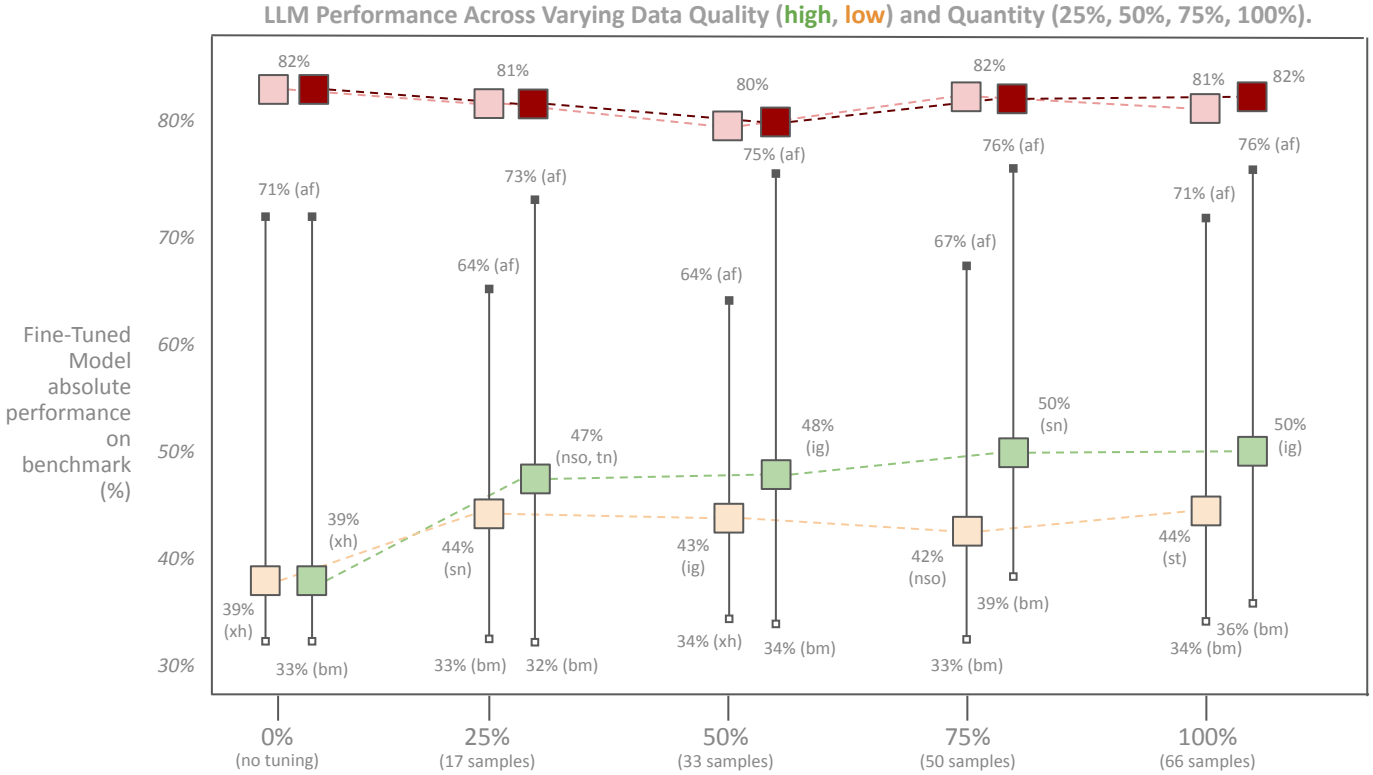


For comprehensive results tables, see here (beginning Row 106 Column C).

All numbers are rounded to the nearest whole integer

But generally, when data is scarce LLM performance is more sensitive to quality

The average gain due to data quality was 5.4%, while the average gains from doubling data (33 → 66 samples) was 2.3%



Data quality: was assessed using GPT-4o model as an annotator to score the quality of a given sample for fine-tuning a model on a target test set (MMLU clinical knowledge).

Training: Llama 3 70B IT was used for fine-tuning. The fine-tuning dataset was MMLU college medicine (200 samples) in English and all 11 African languages.

Trend: Fine-tuning with high quality data generally resulted in higher performance (up to 6.0% on lowest performing models), relative to fine-tuning with low-quality data.

- High quality (English)
- Low quality (English)
- High quality (11 African languages)
- Low quality (11 African languages)

For comprehensive results tables, see here.

All numbers are rounded to the nearest whole integer

Motivation: LLMs perform near human-levels in English, but not other languages

The performance gap diminishes LLM impact for the globally disadvantaged - most of whom don’t speak English.

Question: How large is the LLM performance gap? What can be done to close it?

We focus our investigation on African languages, where the gaps are high when known - but mostly unknown.

Answer: The LLM performance gap for African Languages is 12-20% abs. (25 -53% rel.)

To close the gap, use high quality data, leverage data from related languages, and be mindful of cultural differences.

The performance gap can be closed by:

1. **Fine-tuning using a small, high-quality, task-relevant dataset in the target language:**
   a. Fine-tuning: Across all languages & benchmarks an average monolingual gains of 5.6% was observed.
   b. Data quality: Higher quality data provided up to 14.4% (5.4% average) gains over the same amount of lower quality data.
   c. Domain match: Gains were strongest when the training domain matched the target domain (e.g. College medicine → Clinical knowledge).

2. **Leveraging data from related languages, when data is the target language is limited:**
   a. Cross-lingual transfer: Cross-lingual tuning provided up to 21.1% gains (2.9% average) across all languages / benchmarks.

3. **Being mindful of cultural differences when assessing performance:**
   a. Cultural appropriateness: LLMs perform up to 13% better (3.3% average) on culturally appropriate questions compared to inappropriate questions, and ranged from +1% (Tsonga) to +13% (Shona) on 9/11 languages.

Limitations: performance of Llama 3, scope of fine-tuning, & use-case alignment

Future work should explore other models, and extend the fine-tuning experiments in light of the intended use-case.

1. **Llama 3 is the best open-source solution, but does not outperform GPT-4o out-of-the-box, even after fine-tuning.** Future work should consider extending the experiments presented herein using GPT-4o if an open-source model is not required. We note, that GPT 4o recently became fine-tunable (on July 3rd, 2024).

2. **Our fine-tuning experiments were extensive, but we assessed gains for languages in isolation, not collectively.** Future work should explore the effects of tuning LLMs using data from multiple languages simultaneously; grouping African or related high-resource languages may further improve the performance reported.

3. **Translation of established LLM benchmarks may not perfectly capture the depth / breadth of BMGF use-cases.** While the translated benchmarks offer a previously absent mechanism for measuring LLM performance on 11 African languages, future benchmark creation efforts should consider generating content that supports BMGF use-cases, specifically.

Contributions: Developed novel AI assets impacting 230M speakers globally

Our efforts, dataset, and results were (will be) documented in the form of two publications: NeurIPS 2024, AAAI 2025

2 papers

with 1 under development for AAAI 2025 and one under review by NeurIPS 2024

1.85M words

of new LLM benchmark data in 11 African languages

11 languages

covering a diverse population of Africa

500+ models

trained and evaluated across the entire project

42 translators

managed by our team for translations of Winogrande

230M speakers

of African languages covered by the datasets we created

Appendix

We used Google Sheets to capture Winogrande translations from Upworkers

Native speakers translated the question-answer pairs from the Winogrande small dataset

DESCRIPTION: Below is a table that lists English language Question-Answer pairs (columns C, D, E) and spots for corresponding Amharic Translations of the Question-Answer pairs (columns F, G, H).

TASK: Your task is to translate the English language Question-Answer pairs (columns C, D, E) and place the results in the Amharic Translations of the Question-Answer pairs (columns F, G, H) while *following the translation instructions given directly below exactly*. For background information, you are translating a commonsense reasoning test that features sentences missing a word (denoted with an underscore '_') along with two few-word options, both of which can be substituted into the underscore while keeping the sentence grammatically correct. However, only one of the options makes sense and is correct, but your goal is to simply translate the sentences and both options, not determine which one is correct (we already have the answer key). Your translations must ensure that this commonsense reasoning test can be successfully administered to native speakers of Amharic as well.

TRANSLATION INSTRUCTIONS:
- **Punctation:** Please do not update or alter the punctuation of the original text; if the English language text used a comma or period, ensure that you also use a comma or period in the translation unless it is not grammatically correct.
- **Special Characters:** Please do not update or alter any special characters in your translated text; if the English language text contains one '_' (underscore) character, please ensure that your translation also has only one '_' character, and keep whitespace exactly the same. Do not let the underscores touch letters or other characters (unless it is grammatically correct to do so). For example, in the translation of *"Adam likes milk and John does not, so _ was the only one who bought milk."*, there should likely still be spaces before and after the underscore in the translation. Erroneous punctuation however (e.g. a period that is separated by a space) should be fixed in the translation.
- **English Text:** If you notice grammar or spelling errors in the English text, use your best judgement to make appropriate translations that make sense in Amharic (i.e. please do not carry over any mistakes into the translations). Be sure to read the note above about punctuation mistakes. Do not attempt to edit or fix the English text.
- **Names:** Do not localize names, but ensure they are grammatically correct when substituted for the blank (e.g. "Adam" should not become "Anovuyo" but may become "uAdam" due to proper grammar; note that this is a Xhosa example and may vary for Amharic).
- **Grammatical Correctness and Test Validity (IMPORTANT)**: The translated options (Answer Option 1 and Answer Option 2 in columns G and H), should fit into the Question if substituted for the blank / underscore ('_') (in column F). As an example, in English, the question "The man likes dogs but not cats, so he took home a _ at the pet store." would have options "dog" and "cat", both of which can be substituted into the underscore while keeping the sentence grammatically correct. Clearly, "dog" is correct and "cat" is incorrect. These two features, (1) both options being able to be substituted into the sentence AND (2) exactly one option being the correct one, absolutely must carry over into your translations!

TRANSLATION TASK						ASSESSMENT	
Original English			Amharic Translation			Completion Status	AI Translation Similarity
English Sentence	English Answer Option 1	English Answer Option 2	Amharic Sentence (Insert your translation below)	Amharic Answer Option 1 (Insert your translation below)	Amharic Answer Option 2 (Insert your translation below)	0.00%	0.00%
Ian volunteered to eat Dennis's menudo after already having a bowl because _ despised eating intestine.	Ian	Dennis				✗	No Warning
Ian volunteered to eat Dennis's menudo after already having a bowl because _ enjoyed eating intestine.	Ian	Dennis				✗	No Warning
He never comes to my home, but I always go to his house because the _ is smaller.	home	house				✗	No Warning
He never comes to my home, but I always go to his house because the _ is bigger.	home	house				✗	No Warning

Shown is a screenshot of the survey form used to collect translation data from reviewers on Upwork.com. The English Winogrande items are displayed on the left (blue columns) and the reviewer is instructed to provide the appropriate translations for the question and each answer option, maining punctuation and names (red columns). The translator is also shown their completion status and a warning if their translation is too similar (ROUGE-1 score > 0.9) or too dissimilar (ROUGE-1 score <0.5) to machine translation.

We used Google Sheets to capture Winogrande translations from Upworkers

Native speakers assessed and corrected previous human translations of Winogrande

DESCRIPTION: Below is a table that lists English language Question-Answer pairs (columns D, E, F) and Amharic Translations of the Question-Answer pairs (columns G, H, I).

TASK: Your task is to perform the following checks:
1. Compare the English Sentence (column D) with the Amharic Translation of the Sentence (column G).
2. Compare the English Answer Option 1 (column E) with the Amharic Translation of Answer Option 1 (column H).
3. Compare the English Answer Option 2 (column F) with the Amharic Translation of Answer Option 2 (column I).
4. If any of the Amharic translations are not correct, then mark the quality (column K) as "No - Needs Correction", otherwise mark it as "Yes - Perfect Translation".
5. If any of the Amharic translation are not correct (and marked "No - Needs Correction"), then provide the correct Amharic translation of the English Sentence, Answer Option 1, and Answer Option 2 in columns L, M, and N, respectively.

For background information, you are reviewing/translating a commonsense reasoning test that features sentences missing a word (denoted with an underscore '_') along with two few-word options, both of which can be substituted into the underscore while keeping the sentence grammatically correct. However, only one of the options makes sense and is correct, but your goal is to simply review/translate the sentences and both options, not determine which one is correct (we already have the answer key). The translations must ensure that this commonsense reasoning test can be successfully administered to native speakers of Amharic as well.

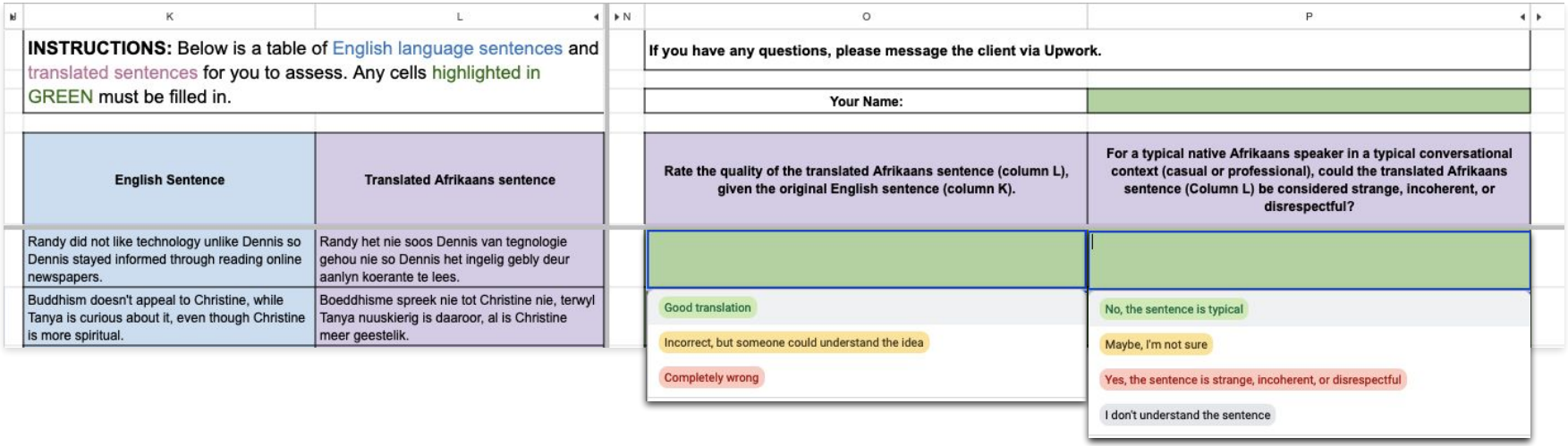
TRANSLATION INSTRUCTIONS:
- **Punctation:** Please do not update or alter the punctuation of the original text; if the English language text used a comma or period, ensure that you also use a comma or period in the translation unless it is not grammatically correct.
- **Special Characters:** Please do not update or alter any special characters in your translated text; if the English language text contains one '_' (underscore) character, please ensure that your translation also has only one '_' character, and keep whitespace exactly the same. Do not let the underscores touch letters or other characters (unless it is grammatically correct to do so). For example, in the translation of "Adam likes milk and John does not, so _ was the only one who bought milk.", there should likely still be spaces before and after the underscore in the translation. Erroneous punctuation however (e.g. a period that is separated by a space) should be fixed in the translation.
- **English Text:** If you notice grammar or spelling errors in the English text, use your best judgement to make appropriate translations that make sense in Amharic (i.e. please do not carry over any mistakes into the translations). Be sure to read the note above about punctuation mistakes. Do not attempt to edit or fix the English text.
- **Names:** Do not localize names, but ensure they are grammatically correct when substituted for the blank (e.g. "Adam" should not become "Anovuyo" but may become "uAdam" due to proper grammar; note that this is a Xhosa example and may vary for Amharic).
- **Grammatical Correctness and Test Validity (IMPORTANT):** The translated options (Answer Option 1 and Answer Option 2 in columns H and I), should fit into the Question if substituted for the blank / underscore ('_') (in column G). As an example, in English, the question "The man likes dogs but not cats, so he took home a _ at the pet store." would have options "dog" and "cat", both of which can be substituted into the underscore while keeping the sentence grammatically correct. Clearly, "dog" is correct and "cat" is incorrect. These two features, (1) both options being able to be substituted into the sentence AND (2) exactly one option being the correct one, absolutely must carry over into your translations!

ENGLISH SOURCE AND AMHARIC TRANSLATION						YOUR ANSWER HERE				ASSESSMENT	
Original English			Amharic Translation			Mark the Translation Quality	Corrected Amharic Translation			Completion Status	Warning
English Sentence	English Option 1	English Option 2	Amharic Sentence	Amharic Option 1	Amharic Option 2	Is the Translation shown in (Col G, H, I in red) correct?	Question	Answer Option 1	Answer Option 2	0.00%	0.00%
Ian volunteered to eat Dennis's menudo after already having a bowl because _ despised eating intestine.	Ian	Dennis	_ አገጀት መብላትን ስለናቀ ኢያን ንድጋዳ ሳህን ከያዘ በኋላ የዴኒስ ሜኑዶን ለመብላት ፈቃደኛ ሆነ.	ኢያን	ዴኒስ					✗	None
Ian volunteered to eat Dennis's menudo after already having a bowl because _ enjoyed eating intestine.	Ian	Dennis	ኢያን አገጀት መብላት ስለወደደ _ ንድጋዳ ሳህን ከያዘ በኋላ የዴኒስ ሜኑዶን ለመብላት ፈቃደኛ ሆነ.	ኢያን	ዴኒስ					✗	None
He never comes to my home, but I always go to his house because the _ is smaller.	home	house	ወጂ መኖሪያዬ ፈጽሞ አደመጣም, ነገር ግን _ ትንሽ ስለሆነ ሁልጊዜ ወጂ ቤቱ አሄዳለሁ.	መኖሪያዬ	ቤቱ	Yes - Perfect Translation / No - Needs Correction				✗	None
He never comes to my home, but I always go to his house because the _ is bigger.	home	house	ወጂ መኖሪያዬ ፈጽሞ አደመጣም, ነገር ግን _ ትልቅ ስለሆነ ሁልጊዜ ወጂ ቤቱ አሄዳለሁ.	መኖሪያዬ	ቤቱ					✗	None

Shown is a screenshot of the survey form used to collect translation review data from reviewers on Upwork.com. The English and human-translated Winogrande items are displayed on the left (blue and red columns) and the reviewer is instructed to assess the given translation as “Perfect translation” or “Needs Correction”. If the translation needs correction, they must provide the full corrected version. The translator is also shown their completion status and a warning if their translation (or the existing translation) is too similar (ROUGE-1 score > 0.9) or too dissimilar (ROUGE-1 score <0.5) to machine translation.

Two raters per benchmark question captured “quality” and “appropriateness”

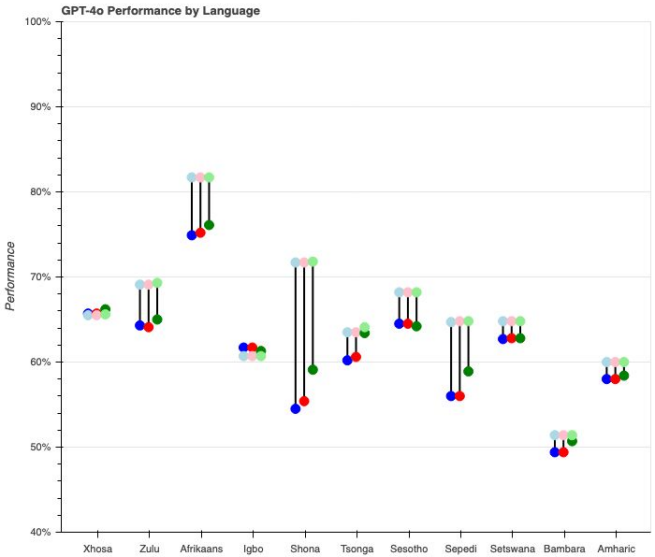
Humans assessed (i) translation quality, and (ii) if the translation was “strange, incoherent or disrespectful”

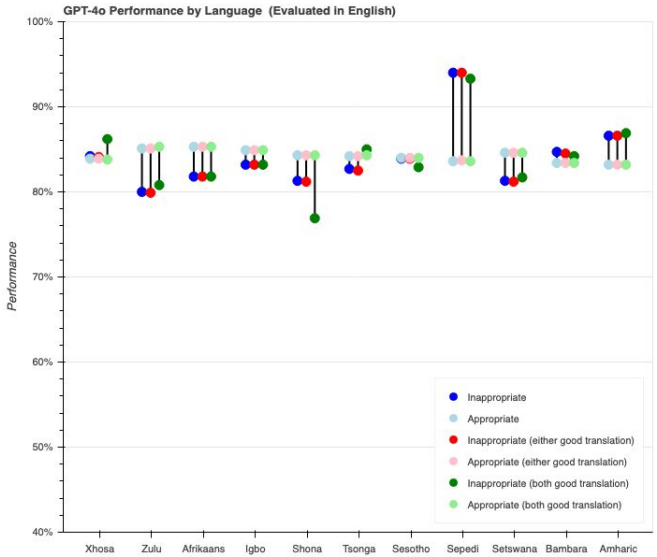


Above is a screenshot of the survey form used to collect cultural appropriateness data from reviewers on Upwork.com. The English and translated Winogrande items are displayed on the left (columns K and L), and the reviewer is instructed to select items from the dropdowns on the right (columns O and P), answering two questions about the translation. First, the reviewer is asked about the quality of translation, rating it as either “Good translation”, “Incorrect, but someone could understand the idea”, or “Completely wrong”. Second, the reviewer is asked whether the translation is strange, incoherent, or disrespectful, rating it as either “No, the sentence is typical”, “Maybe, I’m not sure”, “Yes, the sentence is strange, incoherent, or disrespectful”, or “I don’t understand the sentence”.

Cultural appropriateness affects LLM performance across languages

GPT-4o performance lift on appropriate vs. inappropriate Winogrande items ranged from -2.7% (Igbo) to +19.2% (Sepedi)





	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
all items	0.1%	-0.4%	3.3%	-2.7%	14.2%	1.7%	3.6%	19.2%	-1.1%	3.2%	5.4%
either good translation	0.0%	-0.3%	3.0%	-2.6%	13.2%	1.2%	3.6%	19.2%	-1.4%	3.1%	5.4%
both good translation	1.8%	-0.2%	2.2%	-2.3%	5.3%	1.3%	2.9%	15.7%	-0.8%	1.6%	5.4%

GPT-4o was evaluated in each language and in English on the test split of Winogrande, partitioned by human annotations of cultural appropriateness in each language using 3 different inclusion criteria based on translation quality.

We consider an item to be **inappropriate** if an annotator in **EITHER** round marked the item as **inappropriate.**

Blue: All items included

Red: We consider only items where an annotator in **EITHER** round marked the translation as “good translation” or “understandable”

Green: We consider only items where **BOTH** annotators marked the translation as a “good translation” or “understandable”

To identify lift which is only a result of cultural appropriateness, we subtract the lift when evaluated in English from the lift when evaluated in the target language.

Cultural appropriateness affects LLM performance across languages

We observe a lift in GPT-family performance across languages on appropriate data as determined by human annotators.

A. Performance on "appropriate" subset

Sample Size	1379	1355	1086	809	1548	1422	1461	1711	1433	1019	1370
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	65.5%	69.1%	81.7%	60.7%	71.7%	63.5%	68.2%	64.7%	64.8%	51.4%	60.0%
GPT 4	61.4%	65.3%	80.0%	60.7%	67.4%	58.7%	65.2%	58.9%	60.1%	52.2%	51.7%
GPT 3.5	51.6%	49.7%	56.7%	51.9%	51.1%	50.4%	49.9%	50.8%	50.7%	49.6%	50.4%

B. Performance on "**NOT** appropriate" subset

Sample Size	388	412	681	958	219	345	306	56	334	748	397
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	65.7%	64.3%	74.9%	61.7%	54.5%	60.2%	64.5%	56.0%	62.7%	49.4%	58.0%
GPT 4	60.7%	59.2%	72.3%	57.5%	53.6%	53.3%	58.7%	59.5%	62.1%	49.9%	49.3%
GPT 3.5	53.5%	50.6%	51.9%	50.0%	53.9%	51.3%	52.7%	56.5%	53.9%	51.4%	51.4%

C. English Performance on "appropriate" subset

Sample Size	1379	1355	1086	809	1548	1422	1461	1711	1433	1019	1370
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	83.9%	85.1%	85.3%	84.9%	84.3%	84.2%	84.0%	83.6%	84.6%	83.4%	83.2%
GPT 4	83.1%	83.1%	84.8%	83.0%	83.3%	83.8%	83.5%	82.9%	83.2%	82.1%	82.5%
GPT 3.5	58.7%	58.5%	59.9%	61.0%	59.0%	59.0%	59.4%	58.8%	58.6%	57.0%	58.7%

D. English performance on "**NOT** appropriate" subset

Sample Size	388	412	681	958	219	345	306	56	334	748	397
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	84.2%	80.0%	81.8%	83.2%	81.3%	82.7%	83.9%	94.0%	81.3%	84.7%	86.6%
GPT 4	82.9%	82.8%	80.2%	83.1%	81.4%	80.1%	80.7%	86.3%	82.4%	84.3%	85.1%
GPT 3.5	59.0%	59.7%	57.0%	56.9%	57.4%	57.9%	55.9%	59.5%	59.7%	61.1%	59.2%

E. Increased Target Language performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.2%	4.8%	6.8%	-1.0%	17.2%	3.3%	3.7%	8.8%	2.2%	2.0%	2.0%
GPT 4	0.8%	6.1%	7.8%	3.2%	13.8%	5.4%	6.5%	-0.6%	-2.0%	2.4%	2.4%
GPT 3.5	-1.9%	-1.0%	4.7%	1.8%	-2.7%	-0.9%	-2.8%	-5.8%	-3.2%	-1.8%	-1.0%

F. Increased English performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.3%	5.1%	3.5%	1.7%	3.0%	1.5%	0.1%	-10.4%	3.2%	-1.3%	-3.5%
GPT 4	0.2%	0.3%	4.7%	0.0%	1.8%	3.7%	2.8%	-3.4%	0.7%	-2.2%	-2.6%
GPT 3.5	-0.3%	-1.2%	2.9%	4.1%	1.6%	1.1%	3.5%	-0.8%	-1.1%	-4.1%	-0.5%

G. Increased target languagse performance on "appropriate" subset over English increase

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.2%	-0.4%	3.3%	-2.7%	14.2%	1.7%	3.6%	19.2%	-1.1%	3.2%	5.4%
GPT 4	0.6%	5.9%	3.1%	3.2%	12.0%	1.8%	3.7%	2.7%	-2.7%	4.6%	5.0%
GPT 3.5	-1.6%	0.3%	1.9%	-2.3%	-4.3%	-2.1%	-6.3%	-5.0%	-2.1%	2.3%	-0.5%

All Winogrande items are included. Dataset items are considered “inappropriate” if either annotator marked it as “strange, incoherent, or disrespectful”.

Table A: Per-language GPT-family performance on “appropriate” dataset items evaluated in the target language.

Table B: Per-language GPT-family performance on “inappropriate” dataset items evaluated in the target language.

Table C: Per-language GPT-family performance on “appropriate” dataset items evaluated in English.

Table D: Per-language GPT-family performance on “inappropriate” dataset items evaluated in English.

Table E: Per-language difference in performance between “appropriate” and “inappropriate” dataset items evaluated in the target language (A - B).

Table F: Per-language difference in performance (lift) between “appropriate” and “inappropriate” dataset items evaluated in English (C - D).

Table G: Per-language lift in target language minus the lift in English (E - F).

Cultural appropriateness affects LLM performance across languages

We observe a lift in GPT-family performance across languages on appropriate data as determined by human annotators.

A. Performance on "appropriate" subset

Sample Size	1379	1355	1086	809	1548	1422	1461	1710	1433	1019	1370
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	65.5%	69.1%	81.7%	60.7%	71.7%	63.5%	68.2%	64.8%	64.8%	51.4%	60.0%
GPT 4	61.4%	65.3%	80.0%	60.7%	67.4%	58.7%	65.2%	58.9%	60.1%	52.2%	51.7%
GPT 3.5	51.6%	49.7%	56.7%	51.9%	51.1%	50.4%	49.9%	50.8%	50.7%	49.6%	50.4%

B. Performance on "**NOT** appropriate" subset

Sample Size	385	409	676	956	207	339	306	56	331	737	397
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	65.7%	64.1%	75.2%	61.7%	55.4%	60.6%	64.5%	56.0%	62.8%	49.4%	58.0%
GPT 4	60.9%	59.4%	72.5%	57.4%	54.4%	53.6%	58.7%	59.5%	62.1%	50.0%	49.3%
GPT 3.5	53.9%	50.7%	52.0%	50.0%	53.9%	51.0%	52.7%	56.5%	53.7%	51.5%	51.4%

C. English Performance on "appropriate" subset

Sample Size	1379	1355	1086	809	1548	1422	1461	1710	1433	1019	1370
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	83.9%	85.1%	85.3%	84.9%	84.3%	84.2%	84.0%	83.7%	84.6%	83.4%	83.2%
GPT 4	83.1%	83.1%	84.8%	83.0%	83.3%	83.8%	83.5%	83.0%	83.2%	82.1%	82.5%
GPT 3.5	58.7%	58.5%	59.9%	61.0%	59.0%	59.0%	59.4%	58.8%	58.6%	57.0%	58.7%

D. English performance on "**NOT** appropriate" subset

Sample Size	385	409	676	956	207	339	306	56	331	737	397
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	84.1%	79.9%	81.8%	83.2%	81.2%	82.5%	83.9%	94.0%	81.2%	84.5%	86.6%
GPT 4	83.0%	82.7%	80.0%	83.1%	81.6%	79.7%	80.7%	86.3%	82.3%	84.3%	85.1%
GPT 3.5	59.2%	59.7%	57.1%	56.9%	57.3%	57.7%	55.9%	59.5%	59.4%	61.0%	59.2%

E. Increased Target Language performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.2%	4.9%	6.5%	-1.0%	16.3%	2.9%	3.7%	8.8%	2.0%	2.0%	2.0%
GPT 4	0.6%	5.9%	7.6%	3.3%	12.9%	5.2%	6.5%	-0.6%	-2.1%	2.2%	2.4%
GPT 3.5	-2.2%	-1.0%	4.6%	1.8%	-2.8%	-0.7%	-2.8%	-5.8%	-3.0%	-1.9%	-1.0%

F. Increased English performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.2%	5.3%	3.5%	1.6%	3.2%	1.7%	0.1%	-10.4%	3.4%	-1.1%	-3.5%
GPT 4	0.0%	0.4%	4.8%	-0.1%	1.6%	4.0%	2.8%	-3.3%	0.9%	-2.2%	-2.6%
GPT 3.5	-0.5%	-1.2%	2.7%	4.1%	1.7%	1.3%	3.5%	-0.7%	-0.8%	-4.0%	-0.5%

G. Increased target languagse performance on "appropriate" subset over English increase

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	0.0%	-0.3%	3.0%	-2.6%	13.2%	1.2%	3.6%	19.2%	-1.4%	3.1%	5.4%
GPT 4	0.5%	5.5%	2.8%	3.4%	11.3%	1.1%	3.7%	2.7%	-3.0%	4.4%	5.0%
GPT 3.5	-1.7%	0.2%	1.9%	-2.3%	-4.5%	-1.9%	-6.3%	-5.1%	-2.1%	2.0%	-0.5%

Only Winogrande items that **EITHER** annotator marked as "**good translation**" or "**understandable**" included. Dataset items are considered "inappropriate" if either annotator marked it as "strange, incoherent, or disrespectful".

Table A: Per-language GPT-family performance on "appropriate" dataset items evaluated in the target language.

Table B: Per-language GPT-family performance on "inappropriate" dataset items evaluated in the target language.

Table C: Per-language GPT-family performance on "appropriate" dataset items evaluated in English.

Table D: Per-language GPT-family performance on "inappropriate" dataset items evaluated in English.

Table E: Per-language difference in performance between "appropriate" and "inappropriate" dataset items evaluated in the target language (A - B).

Table F: Per-language difference in performance (lift) between "appropriate" and "inappropriate" dataset items evaluated in English (C - D).

Table G: Per-language lift in target language minus the lift in English (E - F).

Cultural appropriateness affects LLM performance across languages

We observe a lift in GPT-family performance across languages on appropriate data as determined by human annotators.

A. Performance on "appropriate" subset

Sample Size	1371	1339	1086	809	1546	1379	1452	1701	1433	1016	1369
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	65.6%	69.3%	81.7%	60.7%	71.8%	64.1%	68.2%	64.8%	64.8%	51.4%	60.0%
GPT 4	61.6%	65.5%	80.0%	60.7%	67.4%	59.3%	65.4%	58.8%	60.1%	52.2%	51.6%
GPT 3.5	51.6%	49.8%	56.7%	51.9%	51.1%	50.3%	49.8%	50.6%	50.7%	49.7%	50.4%

B. Performance on "**NOT** appropriate" subset

Sample Size	292	324	609	758	146	255	259	30	303	569	390
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	66.2%	65.0%	76.1%	61.3%	59.1%	63.4%	64.2%	58.9%	62.8%	50.7%	58.4%
GPT 4	61.5%	59.3%	72.7%	57.8%	59.6%	55.0%	58.4%	58.9%	62.2%	50.3%	49.8%
GPT 3.5	54.2%	50.0%	51.6%	49.1%	51.1%	52.0%	53.5%	54.4%	53.8%	54.2%	51.8%

C. English Performance on "appropriate" subset

Sample Size	1371	1339	1086	809	1546	1379	1452	1701	1433	1016	1369
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	83.8%	85.3%	85.3%	84.9%	84.3%	84.3%	84.0%	83.6%	84.6%	83.4%	83.2%
GPT 4	83.1%	83.1%	84.8%	83.0%	83.3%	83.9%	83.7%	83.0%	83.2%	82.1%	82.4%
GPT 3.5	58.8%	58.7%	59.9%	61.0%	59.0%	58.9%	59.4%	58.7%	58.6%	57.1%	58.7%

D. English performance on "**NOT** appropriate" subset

Sample Size	292	324	609	758	146	255	259	30	303	569	390
	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	86.2%	80.8%	81.8%	83.2%	76.9%	85.0%	82.9%	93.3%	81.7%	84.2%	86.9%
GPT 4	83.2%	83.1%	79.6%	82.6%	79.7%	79.7%	80.2%	82.2%	83.3%	83.9%	85.0%
GPT 3.5	58.3%	60.7%	57.3%	56.6%	57.8%	58.7%	56.0%	55.6%	59.3%	62.8%	59.7%

E. Increased Target Language performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-0.6%	4.3%	5.6%	-0.6%	12.6%	0.7%	4.0%	5.9%	2.0%	0.7%	1.6%
GPT 4	0.0%	6.3%	7.3%	2.9%	7.8%	4.2%	7.0%	-0.1%	-2.1%	2.0%	1.8%
GPT 3.5	-2.7%	-0.2%	5.1%	2.8%	0.0%	-1.7%	-3.7%	-3.8%	-3.1%	-4.5%	-1.4%

F. Increased English performance on "appropriate" subset

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	-2.4%	4.5%	3.5%	1.7%	7.4%	-0.7%	1.1%	-9.8%	2.8%	-0.9%	-3.8%
GPT 4	-0.2%	-0.1%	5.3%	0.4%	3.6%	4.2%	3.5%	0.7%	-0.1%	-1.9%	-2.6%
GPT 3.5	0.4%	-2.0%	2.6%	4.4%	1.3%	0.2%	3.5%	3.1%	-0.7%	-5.7%	-1.0%

G. Increased target languagse performance on "appropriate" subset over English increase

	Xhosa	Zulu	Afrikaans	Igbo	Shona	Tsonga	Sesotho	Sepedi	Setswana	Bambara	Amharic
GPT 4o	1.8%	-0.2%	2.2%	-2.3%	5.3%	1.3%	2.9%	15.7%	-0.8%	1.6%	5.4%
GPT 4	0.2%	6.3%	2.1%	2.5%	4.2%	0.1%	3.4%	-0.8%	-2.0%	3.9%	4.4%
GPT 3.5	-3.1%	1.8%	2.5%	-1.6%	-1.3%	-1.9%	-7.2%	-6.9%	-2.4%	1.2%	-0.3%

Only Winogrande items that **BOTH** annotators marked as "**good translation**" or "**understandable**" included. Dataset items are considered "inappropriate" if either annotator marked it as "strange, incoherent, or disrespectful".

Table A: Per-language GPT-family performance on "appropriate" dataset items evaluated in the target language.

Table B: Per-language GPT-family performance on "inappropriate" dataset items evaluated in the target language.

Table C: Per-language GPT-family performance on "appropriate" dataset items evaluated in English.

Table D: Per-language GPT-family performance on "inappropriate" dataset items evaluated in English.

Table E: Per-language difference in performance between "appropriate" and "inappropriate" dataset items evaluated in the target language (A - B).

Table F: Per-language difference in performance (lift) between "appropriate" and "inappropriate" dataset items evaluated in English (C - D).

Table G: Per-language lift in target language minus the lift in English (E - F).