ProverbEval: Exploring LLM Evaluation Challenges for Low-resource Language Understanding

Israel Abebe Azime[1,*,†]**, Atnafu Lambebo Tonja**[2,3,*,†]**, Tadesse Destaw Belay**[4,†]**,**
Yonas Chanie[5,†]**, Bontu Fufa Balcha** [6,†]**, Negasi Haile Abadi** [7,†]**, Henok Biadglign Ademtew**[8,†]**,**
Mulubrhan Abebe Nerea[9,†]**, Debela Desalegn Yadeta** [6]**, Derartu Dagne Geremew**[10,†]**,**
Assefa Atsbiha tesfau[7,†]**, Philipp Slusallek**[1]**, Thamar Solorio**[2]**, Dietrich Klakow**[1]**,**

[†] Ethio NLP, [1] Saarland University, [2] MBZUAI, [3] Lelapa AI, [4] Instituto Politécnico Nacional,
[5]Pindo, [6] AAIT, [7]Lesan AI, [8]EAII, [9]University West, [10]Haramaya University,

Abstract

With the rapid development of evaluation datasets to assess LLMs understanding across a wide range of subjects and domains, identifying a suitable language understanding benchmark has become increasingly challenging. In this work, we explore LLM evaluation challenges for low-resource language understanding and introduce **ProverbEval**, LLM evaluation benchmark for low-resource languages, focusing on low-resource language understanding in culture-specific scenarios. We benchmark various LLMs and explore factors that create variability in the benchmarking process. We observed performance variances of up to 50%, depending on the order in which answer choices were presented in multiple-choice tasks. Native language proverb descriptions significantly improve tasks such as proverb generation, contributing to improved outcomes. Additionally, monolingual evaluations consistently outperformed their cross-lingual counterparts in generation tasks. We argue that special attention must be given to the order of choices, the choice of prompt language, task variability, and generation tasks when creating LLM evaluation benchmarks[1].

1 Introduction

Large language models (LLMs) evaluation is gaining increasing attention as these models are typically trained on general-domain datasets while demonstrating notable performance on tasks out of their training domains (Mosbach et al., 2023). The creation of evaluation datasets helps to identify the capabilities of LLMs, pinpoint shortcomings, and establish a measurable path for improvement. Based on Chang et al. (2024), LLM evaluation addresses questions such as what to evaluate (subjects and topics), where to evaluate (selecting appropriate datasets), and how to evaluate (the evaluation process).

To improve LLMs' capabilities and effectively assess their performance, researchers are creating benchmark datasets using a diverse range of domains and languages. This inclusive methodology allows for a more comprehensive evaluation of LLMs' performance across various domains and languages. Popular benchmark datasets like MMLU (Hendrycks et al., 2020) and MEGA-VERSE (Ahuja et al., 2023) cover a wide range of extensive world knowledge tasks and subjects.

To create evaluation benchmarks that are multilingual, researchers Koto et al. (2024); Li et al. (2023); Son et al. (2024) introduced benchmark datasets for different languages by translating a subset of the MMLU dataset. Beyond research efforts, translating existing benchmarks into different languages is an effective strategy to evaluate the multilingual capabilities of closed-source LLMs. These benchmarks evaluate multilingual understanding of models by presenting a range of extensive world knowledge tasks in the language of interest. While combining different subjects in a benchmark dataset may seem beneficial, it does not always provide a clear picture of the model's shortcomings. For example, using MMLU in different languages tests language and subject understanding simultaneously (Hendrycks et al., 2020). There should be evaluation benchmarks that disentangle language understanding and specific subject knowledge.

Language understanding of LLM can be measured in numerous ways, and it is crucial to introduce benchmarks that evaluate complex text comprehension while considering each language's specific linguistic, cultural, and contextual nuances. Creating benchmarks tailored to individual languages' unique values and customs is essential for ensuring comprehensive and accurate evaluations

---

* Equal Contribution.
[1]Evaluation data available at https://huggingface.co/datasets/israel/ProverbEval evaluation code https://github.com/EthioNLP/EthioProverbEval

of language models (Liu et al., 2023).

> *"If culture was a house, then language was the key to the front door, to all the rooms inside."* — Khaled Hosseini, Afghan-born American novelist and physician

Language plays a vital role in shaping and preserving cultural identity (Wang et al., 2024a). It serves as a medium for not only communication but also for the transmission of traditions, values, and beliefs from one generation to another. Through language, individuals can express their emotions, share their stories, and form deep connections with others. With approximately 7000 spoken languages across the globe, each language reflects the unique history, customs, and perspectives of the community that speaks it (Zheng et al., 2024).

A proverb is a short, well-known pithy saying, stating a general truth or a piece of advice. Proverbs are like windows into a culture, offering brief but powerful insights into how people think and live. They carry lessons, reflect shared values, and communicate wisdom passed down through generations. They are a rich manifestation of a society's values, beliefs, and worldview and serve valuable didactic and communicative purposes (Lomotey and Csajbok-Twerefou, 2021). For example, the English proverb *The apple does not fall far from the tree* — means a child grows up to resemble his/her parents. While a plain version of this proverb exists in many cultures, it is expressed differently in different languages and cultures (Liu et al., 2023). For instance, the above proverb might be equivalent in meaning to an Ethiopian Proverb "ልጅ አባቱን አይብ አንቱን ይመስላል" — literally meaning the son resembles his father, the cheese its milk.

In this paper, we introduce **ProverbEval**: LLM evaluation dataset with three distinct tasks based on cultural proverbs for 4 Ethiopian languages and English. The contributions of this work are as follows:

- Introduce **ProverbEval**, a comprehensive LLM evaluation dataset comprising three distinct tasks, derived from cultural proverbs in four Ethiopian languages and English.

- Explore zero-shot performances of a wide range of LLMs on monolingual and cross-lingual language understanding abilities for low-resource languages.

- Explore LLM evaluation challenges for low-resource language understanding.

2 Related Work

Significant efforts have been made to include diverse languages in the development of multilingual language models (Conneau et al., 2020; Xue et al., 2021). Rust et al. (2021) conducted a comparison between multilingual and monolingual language models, employing metrics such as subword fertility. Subword fertility, defined as the ratio of subtokens to total tokens, has been shown to have a direct correlation with model performance across languages, illustrating the impact of tokenization on multilingual language model efficacy. Apart from architecture-based evaluation, multilingual benchmarks help us to track the progress toward multilingualism.

Current evaluation benchmarks prioritize multiple-choice questions due to the relative ease of automatic scoring, as opposed to open-ended question benchmarks that demand significant human involvement (Son et al., 2024; Wang et al., 2024b). For example, MMLU-Pro (Wang et al., 2024b) places a strong emphasis on prompt variations and their influence on large language model (LLM) performance.

Cultural significance of LLM benchmarks is crucial factor to consider as part of language understanding. To incorporate cultures into benchmarks, Myung et al. (2024) introduced BLEnD, which covers 16 countries and 13 languages to prepare datasets that have tests of significance for users in their region. Additionally Liu et al. (2023) shows proverbs can be used to assess LLMs cultural understanding in several languages and introduces **MAPS** (Multicultural Proverbs and Sayings) dataset based on proverbs and sayings to evaluate LLMs multilingual and cultural understanding ability. Our work adopts the same motivation to use proverbs and expands it to different languages and task types.

3 Methodology

3.1 Languages Covered

We create **ProverbEval** benchmark dataset for four low-resource languages along with English to evaluate the cross-lingual capability of LLMs. From these languages, three languages were written in Ethiopic script: Amharic, Tigrinya, and Ge'ez, and
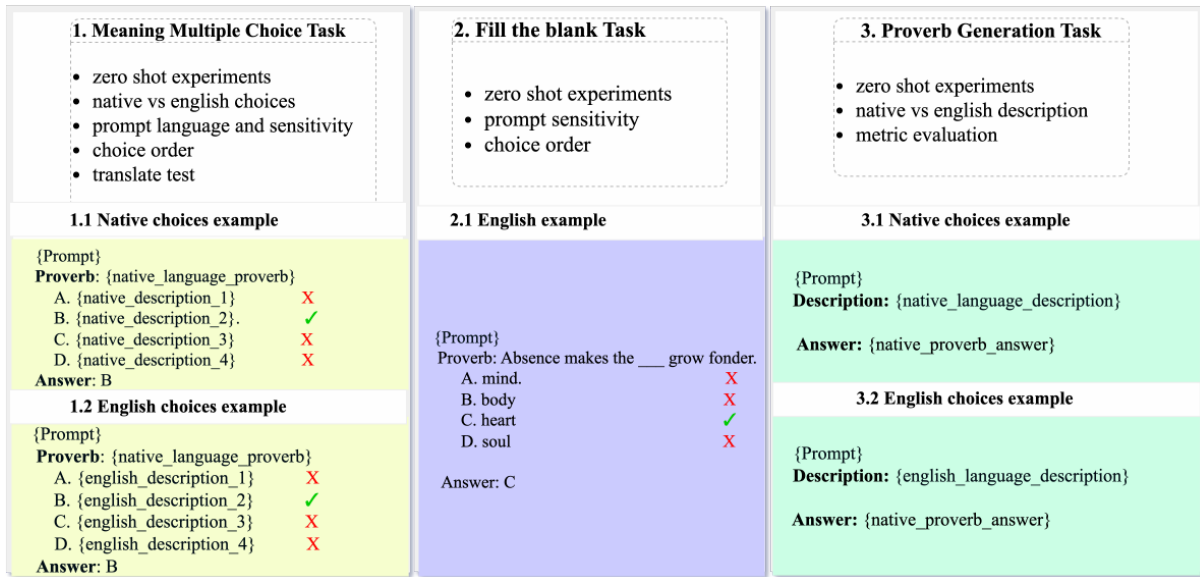
Figure 1: Detailed overview of **ProverbEval**, which consists of three distinct tasks. Native languages include those included in Table 1. Detailed prompt descriptions can be found in Appendix B.

two languages in Latin script: English and Afaan Oromo. We begin with these languages due to the availability of native speaker access to construct the dataset.

Language	# Task 1	# Task 2	# Task 3
Amharic	483	494	484
Afaan Oromo	502	493	502
Tigrinya	380	503	380
Ge'ez	434	429	434
English	437	462	437

Table 1: **ProverbEval** languages and data sizes. All numbers show the test set data size that was prepared.

3.2 Data Collection

Proverbs belong to the public domain and are widely regarded as shared cultural expressions. Their public domain status allows us to freely collect and utilize these resources without licensing restrictions. We collect proverbs from books, online sources, and the common knowledge of volunteer annotators. Our data collection focuses on collecting proverbs, writing detailed explanations in native and English languages, and verifying the correctness of the collected data.

The data collection was carried out by volunteers who are contributing to this research as co-authors. Data collectors utilized existing machine transla-

tion (MT) systems to verify and supplement any vocabulary gaps they encountered while writing proverbs in English after completing explanations in native languages.

As shown in Table 1, we focused on collecting only the test sets for all tasks. Additionally, we included five items that can serve as few-shot examples for *Task 2: Fill in the Blank*.

Biases in Proverbs: The compact and metaphorical language in proverbs is intriguing, but it can also serve as a tool to reinforce gender stereotypes and racial inequalities. In this work, we gave special attention to proverbs that reflect these values and removed all instances of such proverbs.

3.3 Tasks

ProverbEval benchmark contains three main tasks: multiple choice, fill-the-blank, and generation tasks with various evaluation settings.

Task 1: Meaning Multiple Choice In meaning-based multiple-choice tasks, we aim to assess the model's language understanding capabilities by asking the model to select the option with the most similar meaning. For each proverb, four options are provided, each with a detailed explanation of its possible meaning, with only one being correct.

Native vs English choices – One of the factors we are currently exploring in our experiment is the

selection of language used in the multiple-choice options. This exploration will help us access the cross-lingual capability in addition to the monolingual capability of models where the proverb is given, and the model has to choose a sentence that closely resamples it. Figure 1 explains details of task variations. Due to the extremely low resource availability for `Ge'ez, proverb descriptions` are carried out using Amharic, a closely related language.

Task 2: Fill in the Blank   The fill-in-the-blank task is designed to evaluate: the ability of the model to recall proverbs despite containing unconventional word order. For example, the proverb "*Don't let the cat out of the ___*" commonly should be followed by *house* rather than *bag* if we do not have an understanding of that specific proverb, as cats are more commonly associated with houses rather than bags. In this task, we will assess how well the LLMs understand the common proverb.

Task 3: Generation   The ability to determine which proverb best aligns with a particular meaning or situation serves as a way to assess and measure a model's understanding of language. In order to evaluate this, we designed a generation task in which a detailed description of the proverb is provided, and the model is required to select the most appropriate proverb that aligns with the description given. We chose this approach for easier evaluation, though the dataset could also be used for tasks involving generating descriptions based on a given proverb.

Native and English descriptions   - For this task, we utilized both *native* language and *English* descriptions. Descriptions provided in English with the expectation of receiving a native proverb allowed us to evaluate cross-lingual capabilities. Conversely, descriptions given in the native language with the expectation of a corresponding native proverb enabled us to assess monolingual comprehension.

4   Experimental Setup and setting

4.1   Model Selection

Given the wide range of available model options, we established criteria to guide model selection. The models chosen for this experiment were based on the following key factors: (1) different models in terms of the number of parameter size, (2)
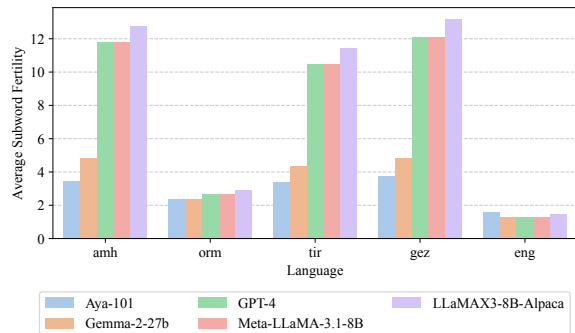

Figure 2: Subword fertility of proverbs for each model's tokenizer in our study. Models that share the same tokenizers are grouped together. Lower values indicate better performance, as they reflect that words are not being excessively split on average.

closed-source versus open-source models, (3) multilingual models versus general-purpose models, and (4) instructed models versus base models.

In this experiment, we did not include open-source instruction-finetuned models due to the difficulty in accessing the specific instruction-finetuning data used for their training. Instead, we utilized LLaMAX3-8B-Alpaca, which is fine-tuned on the Alpaca dataset, and Aya-101, which incorporates a combination of various task-oriented and generative datasets. For large models, we include Meta-LLaMA-3-70B (Dubey et al., 2024)and Gemma-2-27b (Team et al., 2024); for average size models, we included Meta-LLaMA-3-8B (Dubey et al., 2024) and Gemma-2-9b (Team et al., 2024); for multilingual models, we include Aya-101 (Üstün et al., 2024) and LLaMAX3-8B-Alpaca (Lu et al., 2024); finally, we included Gpt-4o (Achiam et al., 2023) from closed source models. From the model list, we select Aya-101 model since it is mT5 based model used to compare with decoder-only models.

4.2   Evaluation

We used ElutherAI's open-source Language Model Evaluation Harness (lm-eval) framework (Gao et al., 2024) to evaluate the models. The library supports evaluation strategies, including log-likelihood, generation, and perplexity, using *YAML* to configure and manage the evaluations. We used log-likelihood and generation for open-source models for multiple-choice and fill-the-blank tasks. In the multiple-choice task and fill-the-blank, each option is appended to the corresponding question and prompt, and the log-likelihood score is subse-

Model Name	Amharic		Afaan Oromo		Tigrinya		Ge'ez		English	average		
prompt language	*native*	*english*	*native*	*english*	*native*	*english*	*native*	*english*		*native*	*english*	*all*
Meta-LLaMA-3-8B												
English prompts	24.72	24.98	32.54	25.37	26.93	29.83	30.11	29.27	49.43	28.58	27.36	30.35
Native Prompts	31.54	26.43	26.23	24.97	27.11	25.09	26.42	24.19		27.83	25.17	26.50
Gemma-2-9b												
English prompts	31.06	30.85	29.22	26.43	29.82	30.88	38.1	45.93	63.31	32.05	33.52	36.18
Native Prompts	29.41	34.77	25.30	26.69	28.07	26.93	26.74	26.04		27.38	29.46	28.27
Gemma-2-27b												
English prompts	35.06	36.3	34.99	27.69	32.39	33.95	41.86	42.71	68.12	36.08	35.16	39.23
Native Prompts	38.36	39.54	25.57	26.89	25.17	25.53	27.65	25.04		29.19	30.65	29.82
Meta-LLaMA-3-70B												
English prompts	41.67	37.61	32.67	27.96	36.49	30.96	55.07	47.62	71.70	41.48	36.04	42.42
Native Prompts	26.24	27.19	27.09	25.76	26.67	27.11	26.27	25.20		26.57	26.32	26.44
LLaMAX3-8B-Alpaca												
English prompts	28.99	25.38	31.94	25.77	29.21	28.25	35.02	30.72	42.71	31.29	27.53	30.89
Native Prompts	30.09	26.15	26.16	26.69	27.17	25.97	26.42	25.12		27.46	25.98	26.72
Aya-101												
English prompts	**48.21**	52.38	**49.40**	32.80	42.19	**55.09**	75.34	**82.49**	77.42	**53.79**	**55.69**	**57.26**
Native Prompts	50.96	**55.21**	41.97	28.82	**49.74**	32.72	45.00	48.69		46.92	41.36	44.14
Gpt-4o												
English prompts	40.19	46.24	49.01	50.80	32.37	35.00	24.20	59.29	**89.97**	36.44	47.83	47.45
Native Prompts	44.42	48.93	27.22	**55.31**	24.82	7.54	0.08	1.15		24.13	28.23	26.18

Table 2: Zero-shot scores of Task 1 (*meaning multiple choice task*) across all models for English and native prompts for choosing from **native** choices and **english** choices. All scores are average of 3 distinct prompts. prompt details in Appendix B and detailed results in E.

quently computed for evaluation. Finally, the accuracy score is reported to be the highest selected option.

For Generation tasks, we heavily rely on ChrF (Popović, 2015) scores but included BLEU and translation edit rate (ter) (Snover et al., 2006) scores in the Appendix G.

For Gpt-4o evaluation, we used the generate_until output type for all tasks since it does not support log-likelihood. We wrote a verbalizer to extract answers from generated answers and calculate accuracy scores for all tasks except for generation.

4.3 Experiments

4.3.1 Zero-shot evaluation of the models

In our first experiment, we performed a comprehensive zero-shot evaluation on all tasks. This involved rigorously testing the LLMs language understanding capabilities by subjecting it to our carefully curated test set.

4.3.2 Key Factors Influencing Zero-Shot Performance

Most LLM evaluation benchmarks rely on multiple-choice tasks due to the ease of evaluation. Compared to generative tasks, multiple choice tasks are simpler to assess using automatic metrics, as they eliminate the possibility that the model provides a correct answer in a different form from the ground truth (Zhang et al., 2024). This approach ensures consistency in the evaluation and avoids ambiguity when assessing the model's performance.

In this work, in addition to introducing proverb-based tasks, we are interested in exploring the reliability of multiple-choice evaluations. To answer this question, we explored the following factors.

Prompting language is one factor that affects the performance of the model. Models can be sensitive to different prompts and prompts given in several languages (Zhang et al., 2023). To evaluate the effect, we tested three English prompts to assess model performance with diverse English inputs and three native prompts for each language to assess performance with instructions in the respective native language.

Order of choices affects the performance of the models in multiple-choice tasks (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023). To evaluate the effect of this problem in a low-resource scenario, we compared the average of three random shuffle performances of the models to correct answers appearing first (all "A") or last choice (all "D").

Few-shot Experiments For *task 2: proverb fill the blank task*, we explored if introducing examples can improve the performance of the models using our validation set.

Effect of Translation Cross-linguistic translation of proverbs is challenging because these ex-

| Model Name | Amharic | | Afaan Oromo | | Tigrinya | | Ge'ez | | English | Average | | |
shuffling strategy	native	english	native	english	native	english	native	english		native	english	all
Meta-LLaMA-3-8B												
3 random shuffle	26.86	26.98	31.54	25.77	29.30	26.05	30.34	27.67	50.73	29.51	26.62	30.58
all option A	58.88	73.91	69.92	80.08	55.79	80.26	69.12	77.63	89.47	63.43	77.97	72.78
all option D	7.23	9.94	16.73	4.98	3.95	2.89	4.38	4.47	23.57	8.07	5.57	8.68
Gemma-2-9b												
3 random shuffle	29.68	33.89	31.54	28.89	29.47	27.46	40.4	27.37	64.23	32.77	29.4	34.77
all option A	63.02	81.16	69.12	88.65	50.53	87.63	73.27	86.05	90.85	63.99	85.87	76.70
all option D	22.11	13.66	24.10	4.38	24.47	5.00	31.34	5.05	50.80	25.51	7.02	20.10
Gemma-2-27b												
3 random shuffle	34.64	34.57	36.25	29.02	28.16	29.91	40.02	29.82	66.13	34.77	30.83	36.50
all option A	65.29	69.57	62.35	74.50	55.00	73.68	72.81	75.00	90.39	63.86	73.19	70.95
all option D	19.21	20.7	25.7	9.76	18.16	7.11	27.19	8.68	54.23	22.57	11.56	21.19
Meta-LLaMA-3-70B												
3 random shuffle	41.94	40.30	35.13	31.51	38.16	30.18	61.44	30.26	74.52	44.17	33.06	42.60
all option A	50.62	53.21	56.77	50.40	41.84	58.68	71.66	58.68	79.86	55.22	55.24	57.97
all option D	28.51	21.33	21.71	14.74	20.53	9.47	42.63	10.26	71.85	28.35	13.95	26.78
LLaMAX3-8B-Alpaca												
3 random shuffle	33.95	25.33	31.48	26.29	30.79	27.63	33.18	27.19	39.51	32.35	26.61	30.59
all option A	36.98	64.39	47.81	72.11	25.79	79.47	39.86	80.53	75.97	37.61	74.13	58.10
all option D	34.5	10.56	21.91	4.38	37.63	4.47	30.65	4.21	24.03	31.17	5.91	19.15
Aya-101												
3 random shuffle	51.24	54.98	51.06	32.8	43.16	55.35	78.88	55.88	80.78	56.09	49.75	56.01
all option A	61.98	67.91	54.58	44.62	51.32	67.63	85.71	70.26	82.38	63.4	62.61	65.15
all option D	57.23	56.73	51.99	31.27	50.53	50.53	81.80	49.21	80.78	60.39	46.94	56.67
Gpt-4o												
3 random shuffle	59.51	52.86	78.75	75.43	43.33	26.05	51.92	22.79	86.96	65.66	66.41	69.75
all option A	52.27	43.48	80.48	76.10	43.16	23.42	91.94	23.68	99.54	68.13	59.71	67.88
all option D	50.00	45.55	72.71	77.89	35.00	11.32	77.42	13.42	99.08	59.35	53.51	61.17

Table 3: Zero-shot accuracy scores of Task 1 (*meaning multiple choice task*) across all models for ***native choices*** and ***English choices***.

pressions often carry culturally specific meanings that may not have direct equivalents in other languages. When proverbs are translated, the nuances and cultural significance can be lost, making it difficult for non-native speakers to fully understand the intended message. Our analysis of closed-source models indicates that LLMs mitigate their lack of language understanding by translating questions from low-resource languages to English and conducting reasoning in English. We translated our proverbs and compared them with the native ones to see if our task is easily solvable by translating, as shown in Table 5.

5 Results and Analysis

5.1 Proverb Multiple Choice

Does model size significantly improve performance for low-resource languages? In Table 2, the result indicates that the size of the open-source base models has a notable impact on the prompt that is being used. Generally, the bigger the models, the better, but Gemma-2-27b takes the lead in *native* prompt, and Meta-LLaMA-3-70B takes the lead in *english* prompt. This directly correlates with Figure 2 that the model with the lowest subword fertility is the better and Gemma models are better multilingual models. This is more reflected in Gemma models having better performance in *na-*

tive choices than *English* choices. We can conclude that a better tokenizer (lower monolingual fertility) is very important in monolingual evaluation compared to cross-lingual evaluation.

Does the choice of language in the prompt affect performance for low-resource languages? For tasks using native prompts, show lower results compared to using English prompts for non-English languages. Using in-language prompt results min 0 and max ±3 differences between *native* and *english* multiple choice in task 1.

As seen in Table 2, only the biggest or multilingual fine-tuned models show promising results in *meaning multiple choice task*. The results in English also show that the task is answerable with a specific focus on languages, and this dataset will be an important resource to identify whether LLMs will achieve meaningful reasoning ability in low-resource languages. Additionally, as we can see from the table, Gpt-4o shows better results when choices are given in English than in native languages.

Are LLMs sensitive to choice order in low-resource languages? As shown in Table 3, smaller models show a difference of accuracy close to 30% and 50% when the answers are provided in the first choice. This number decreases signifi-
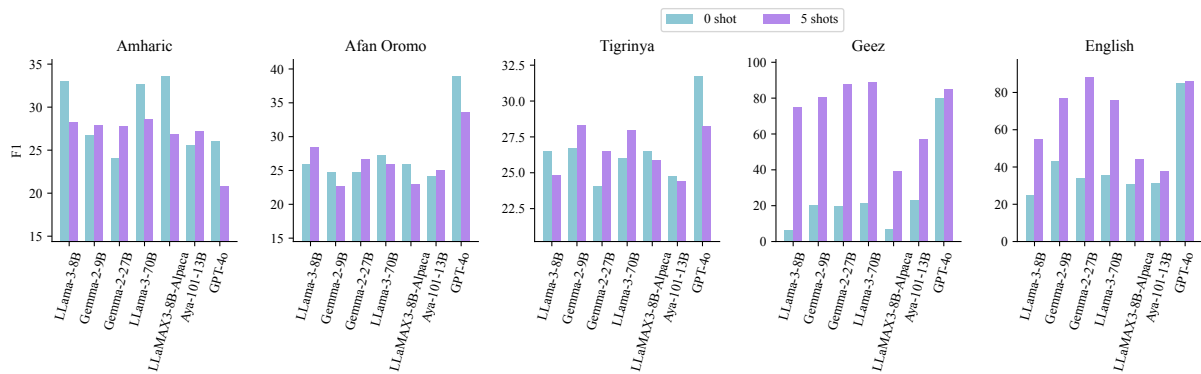
Figure 3: Average accuracy of fill-the-blank results (0 and 5 shots). Zero-shot and five-shot results are an average of three random shuffles using English prompt.

cantly when using larger models or when testing models that pass through supervised fine-tuning. Aya-101 model shows resistance to this disturbance probably because of the training data containing several tasks, whereas the Gpt-4o model shows persistent results regardless of choice order. Looking at *native* and *English* choices for all prompts, we can clearly see that choice order affects cross-lingual tasks more than monolingual tasks.

Monolingual vs Cross-lingual understandings
We evaluate both monolingual and cross-lingual understanding by using native and English choices. The results indicate that in most cases the models demonstrate more robust performance in monolingual tasks than in cross-lingual ones, except for Gpt-4o. Sensitivity to choice order is also less apparent when using monolingual (native) choices, as shown in Table 3.

Does translating proverbs into English improve low-resource language performance? Table 5 shows the effect of translating proverbs written in low-resource languages into English. As we can see from the average results, translating proverbs into English does not significantly help models.

5.2 Task 2: Proverb Fill in the Blank

Zero-shot & Few-shot results of fill the blank
Looking at Figure 3, we observe that all models perform poorly in the fill-in-the-blank task, with the exception of Ge'ez and English for Gpt-4o. The task appears to be easily solvable in English, probably because of the strong focus on English in these models. The examples presented demonstrate a modest performance improvement for open-source models, whereas Gpt-4o shows less benefits from few-shot examples.

5.3 Task 3: Proverb Generation Task

Can LLMs generate coherent proverbs for a given description in low-resource language?
Table 4 shows the ability of the models to generate proverbs for a given description in *native* language and in *English*. LlaMA models show strong generation ability when the description is given in the native language, and Gemma-2-27b becomes competitive when the description is given in the English language, looking at the average scores. There is a huge difference between languages that use Latin script and others that use Ge'ez script.

English vs. Native Descriptions In most cases, models are more likely to generate proverbs in native languages when provided with native descriptions compared to English descriptions. This is because, when given native input, the models tend to anchor their generation around key culturally specific terms or phrases. This context-sensitive approach often results in more accurate and culturally relevant proverb generation, as the models are better able to capture nuances inherent in the native language.

5.4 General Takeaways

Building Models Optimized for Multilingual Functionality Designing an effective tokenizer is crucial, as it serves as a strong foundation for developing more advanced LLMs.

Size Is Not Always the Answer Models with better tokenizers and fine-tuned on carefully curated datasets can be competitive with larger models.

Monolingual vs. Cross-Lingual Evaluations When designing LLM evaluations, it is crucial to consider the differences between monolingual and cross-lingual properties.

6271

| Model Name | Amharic | | Afaan Oromo | | Tigrinya | | Ge'ez | | English | Average | | |
	native	*english*	*native*	*english*	*native*	*english*	*native*	*english*		*native*	*english*	*all*
Meta-LLaMA-3-8B	1.83	1.94	13.79	7.54	1.99	1.81	2.72	1.65	22.41	5.08	3.24	6.19
Gemma-2-9b	1.84	1.20	8.39	4.24	2.61	0.73	2.99	1.21	6.58	3.96	1.85	3.31
Gemma-2-27b	1.34	1.21	8.41	**10.17**	1.72	1.04	2.39	1.28	23.18	3.47	3.43	5.64
Meta-LLaMA-3-70B	2.23	2.74	10.12	5.73	3.72	3.03	2.75	2.87	21.61	4.71	3.59	6.09
LLaMAX3-8B-Alpaca	5.29	4.90	18.11	10.16	3.38	2.54	3.06	0.00	31.25	7.46	4.40	8.74
Aya-101	**6.44**	**5.58**	**19.17**	4.70	4.71	2.80	**7.06**	**6.12**	19.17	**9.35**	**4.80**	8.41
Gpt-4o	5.63	0.03	16.94	3.27	**6.38**	**4.70**	6.00	3.88	**50.39**	8.73	2.97	**10.80**

Table 4: ChrF Generation Scores. For *native*, descriptions were provided in the native language, while for *English*, descriptions were given in English to generate proverbs in each language. Native and English choice averages do not include the English language

| Model Name | Amharic | | Afaan Oromo | | Tigrinya | | Average | |
	native	*english*	*native*	*english*	*native*	*english*	*native*	*english*
Meta-LLaMA-3-8B								
native proverb	24.72	24.98	32.54	25.37	26.93	29.83	28.06	26.73
translated proverb	32.10	27.33	26.49	32.37	23.51	32.37	27.37	30.69
Gemma-2-9b								
native proverb	31.06	30.85	29.22	26.43	29.82	30.88	30.03	29.39
translated proverb	27.13	33.68	27.09	38.98	28.25	34.12	27.49	35.59
Gemma-2-27b								
native proverb	35.06	36.30	34.99	27.69	32.39	33.95	34.15	32.65
translated proverb	31.10	34.99	31.04	33.34	30.32	34.04	30.82	34.12
Meta-LLaMA-3-70B								
translated proverb	41.67	37.61	32.67	27.96	36.49	30.96	36.94	32.18
translated proverb	42.15	38.44	31.41	43.63	32.46	34.65	35.34	38.91
LLaMAX3-8B-Alpaca								
native proverb	28.99	25.38	31.94	25.77	29.21	28.25	30.05	26.47
translated proverb	28.17	27.33	28.75	29.48	26.93	30.44	27.95	29.08
Aya-101								
native proverb	48.21	52.38	49.40	32.8	42.19	55.09	46.6	46.76
translated proverb	40.57	41.27	46.48	41.77	36.76	44.91	41.27	42.65
Gpt-4o								
native proverb	40.19	46.24	49.01	50.80	32.37	35.00	40.52	44.01
translated proverb	59.40	39.06	42.57	44.15	49.34	34.73	50.43	39.31
Average native	34.95	31.02	32.27	26.64	30.97	30.77	32.73	29.48
Average translated	33.54	33.84	31.88	36.6	29.71	35.09	31.71	35.18

Table 5: Accuracy scores of proverb translate-test. Can translating proverbs using NLLB-200 3.3B (NLLB Team et al., 2022) improve the performance of task 1 (*meaning multiple choice task*)? This experiment covers languages supported by NLLB.

Subject vs. Language Understanding   Distinguishing between language and subject understanding is crucial in LLM evaluation.

Translate Test Experiment   Creating a benchmark that captures the cultural and linguistic nuances of a language is crucial for evaluating LLMs. This ensures that language understanding assessments are robust and not artificially inflated by simple translation systems.

Distinct Patterns in Ge'ez Proverbs   We carefully analyzed the linguistic patterns in Ge'ez and observed distinct behaviors in certain tests. Our findings suggest the following characteristics: (1) Ge'ez proverbs are predominantly derived from biblical sources, making them more predictable. (2) Instead of focusing on everyday activities, they emphasize spiritual traditions and customs, providing limited contextual diversity. (3) Ge'ez proverbs are generally shorter and more predictable than those in other languages. (4) The dataset used for Ge'ez proverbs was sourced from a single collection, increasing the likelihood of its inclusion in common LLM training datasets.

6 Conclusion

In this work, we explore the challenges of LLM evaluation for low-resource language understanding. We also introduce a **ProverbEval**, LLM evaluation benchmark for low-resource language based on proverbs to focus on low-resource language understanding in culture-specific scenarios. Our results indicate that LLMs still significantly underperform in non-English languages when it comes to understanding proverbs, as compared to their performance in English. We observed that prompting LLMs in their native languages leads to lower accuracy, and the models have high sensitivity to the order in which choices are presented. In the fill-in-the-blank task, few-shot prompting showed minimal improvement. In generative tasks, LLMs perform better when descriptions are provided in native languages.

In benchmarks focused on cultural understanding, some results may not be transferable to other languages or to broader evaluations. However, this highlights the need for specialized evaluations that capture cultural nuances to ensure that LLMs demonstrate true language understanding.

Acknowledgment

Limitations

Should open-source and closed-source model results be reported together? Open-source models can be easily evaluated using log-likelihood scores. However, this approach is not feasible for closed-source models. As a result, we converted all tasks to generation-based evaluation for closed-source models, a method widely adopted across various evaluation benchmarks. Despite its popularity, these results should not be considered directly comparable. The performance of closed-source models highly depends on the specific verbalizer (tool used to extract answers from long generations) used for each task.

Label-based vs sequence-based evaluations An interesting question explored by Lyu et al. (2024) is whether to evaluate large language models (LLMs) based on the probability assigned to the multiple-choice letter (e.g., "A" for the first option) or the content of the choice itself. In this work, given the extensive number of experiments we conducted, we opted to use label-based evaluation.

Language coverage The scope of this work includes a limited number of languages, primarily constrained by the availability of volunteer native speakers and resource limitations. Expanding the language coverage to include a broader range of cultures and languages would significantly enhance the utility of the benchmark, making it a more comprehensive tool for evaluating model performance across diverse linguistic and cultural contexts.

Limited LLMs evaluation The number of LLMs evaluated in this work is limited. Expanding the study to include a broader range of both open-source and closed-source models could provide deeper insights. Additionally, an important avenue for future research is exploring how this type of language understanding can inform the development of more robust multilingual models. However, this particular question falls outside the scope of the present study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela

Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. http://www.ethnologue.com/. [Accessed 10-10-2024].

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

"Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin". 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Benedicta Adokarley Lomotey and Ildiko Csajbok-Twerefou. 2021. A pragmatic and sociolinguistic analysis of proverbs across languages and cultures. *Journal of Pragmatics*, 182:86–91.

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

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

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

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction fine-tuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024a. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

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

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust llm evaluators. *arXiv preprint arXiv:2405.11966*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11248–11259, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Details of covered languages

There are more than 2000 languages spoken in the African continent, and more than 80 of them are spoken in Ethiopia[2]. Amharic, Afaan Oromo, and Tigrinya are the top languages in Ethiopia by the number of speakers. Ge'ez language is also known as Ethiopic script, the origin of Amharic and Tigrinya languages.

Amharic (amh): is a Semitic language written in Ge'ez script, which consists of 33 primary characters, each with seven vowel sequences. It is the second most widely spoken Semitic language, next to Arabic.

Afaan Oromo (orm): is an Afro-Asiatic language written in Latin script. It is the most widely spoken language in Ethiopia and the third most widely spoken in Africa, next to the Arabic and Hausa languages

Tigrinya (tir): is a Semitic language spoken in the Northern part of Ethiopia and Eritrea. The language uses Ge'ez script with additional Tigrinya alphabets and it is the fourth widely spoken language in Ethiopia next to Somali (Eberhard et al., 2024).

Ge'ez (gez): is a language of Ethiopia that is used only as a second language and does not have an ethnic community. It belongs to the Afro-Asiatic language family.

English (eng): we have created a new proverb dataset for the English language. The proverb descriptions of the other languages also have English descriptions for parallel evaluation.

---

[2]https://www.statista.com/statistics/1280625/number-of-living-languages-in-africa-by-country/

B Prompts for zero-shot and ICL

Table 6 presents all prompts used in the three proposed tasks: Task1: multiple choice, Task2: fill in the blank, and Task: proverb description generation, respectively.

English multiple choice prompts

Prompt 1: You are LLM capable of understanding {language} language. I will give you a prompt and a list of descriptions that have the same meaning. Return a letter for the correct choice among four choices given

Prompt 2: You are LLM capable of understanding language. I will give you a prompt and a list of descriptions that have the same meaning. Return a letter for the correct choice among four choices given

Prompt 3: Which choice are similar?

Afaan Oromo multiple choice prompts

Prompt 1: Ati LLM dandeetti Afan {language} hubachuu qabdudha. Gaaffii fi fillannoowwan hiikaa/eergaa ibsan sif nan laadha. Filannoowwan afur keennaman keessaa quubee deebiin sirri irra jiru naaf deebisii.

Prompt 2: Ati LLM dandeetti Afan hubachuu qabdudha. You are LLM capable of understanding language. Gaaffii fi fillannoowwan hiikaa/eergaa ibsan sif nan laadha. Filannoowwan afur keennaman keessaa quubee deebiin sirri irra jiru naaf deebisii.

Prompt 3: Fiilannoowwan armaan gadii keessaa kamtuu hiika/eergaa walfakkaataa qaba

Amharic multiple choice prompts

Prompt 1: አንተ {language} ቋንቋ መረዳት የምትችል የኮምፒተር ማሽን ነህ፡፡በመቀጠል ምሳሌያዊ አነጋገር እና ትርጉሞች ስጥሃለሁ ከተሰጡት አራት ምርጫዎች መካከል ለተከከለኛው ምርጫ ፊደል ይመልሱ

Prompt 2: አንተ ቋንቋ መረዳት የምትችል የኮምፒተር ማሽን ነህ፡፡በመቀጠል ምሳሌያዊ አነጋገር እና ትርጉሞች ስጥሃለሁ ከተሰጡት አራት ምርጫዎች መካከል ለተከከለኛው ምርጫ ፊደል ይመልሱ።

Prompt 3: የትኛውው ምርጫ ተመሳሳይ ነው?

Tigrinya multiple choice prompts

Prompt 1: ንስኻ {language} ዝተበሃለ ቋንቋ ምርዳእ ኣቕሚ ዘለካ ዓብዪ ናይ ቋንቋ ስልጡን ኢኻ፡፡ዝርዝር ተመሳሳሊ ትርጉም ዘለዎም መግለጺታት ምስ መጠየቅታ ከህበካ እየ፡፡ካብቶም ኣርባዕተ ምርጫታት ቅኑዕ ዝኾነ መልሲ ዝሓዘ መማረጺ ፊደል ምረጽ፡፡

Prompt 2: ንስኻ ቋንቋ ምርዳእ ኣቕሚ ዘለካ ዓብዪ ናይ ቋንቋ ስልጡን ኢኻ፡፡ዝርዝር ተመሳሳሊ ትርጉም ዘለዎም መግለጺታት ምስ መጠየቅታ ከህበካ እየ፡፡ካብቶም ኣርባዕተ ምርጫታት ቅኑዕ ዝኾነ መልሲ ዝሓዘ መማረጺ ፊደል ምረጽ፡፡

Prompt 3: ኣየናይ ምርጫ እዩ ተመሳሳሊ?

Ge'ez multiple choice prompts

Prompt 1: አንተ ውቱ ኪን ወዘተአምር ልሳን {language}:: አምዘታሕቱ ዘተዳለው አማሪያት በገሮ አውስ (ጓይ) ፊደል ዘስንኦ አምዘተደለወ አንጋሪ ምሳሊ ።

Prompt 2: አንተ ውቱ ኪን ወዘተአምር ልሳን ፡፡አምታሕቱ ዘተዳለው አማሪያት በገሮ አውስ (ጓይ) ፊደል ዘስንኦ አምዘተደለወ አንጋሪ ምሳሊ ።

Prompt 3: አይ ውኡቱ አንጋሪ ምሳሊ ዕሩየ ዘኮነ አው ዘይቀርብ ምስለ አለተደለው አንጋሪ ምሳሊያት አምዘታሕቱ?

Fill the blank prompt

English: You are LLM capable of understanding {language} language. Given a proverb, can you fill the blank with an appropriate word from the choices? blank is shown with '___'.
{language} Proverb: {Proverb}
Choices:
 A: {A}
 B: {B}
 C: {C}
 D: {D}

Answer:

Proverb generation prompt

 You are LLM capable of understanding {language} language. Based on the detailed description provided in {source_language}, generate an appropriate proverb in {target_language} that captures the essence and meaning of the context.

{source_language} Description: {Description}
{target_language} Proverb:

Table 6: Prompts (in five languages) used for decoder-only zero-shot and in-context learning experiments

C   Multiple choice detail results

In Table 7, we present an alternative analysis of choice order variance. We averaged the results from the first three randomly shuffled prompts and compared them against instances where the correct choice appeared as either the first or last option. The numbers reflect the deviation from the average random shuffle baseline. For example, a value of +33.92 indicates an increase in accuracy by 33.92 percentage points, while -19.63 signifies a decrease of 19.63 points relative to the baseline.

Model Name	Amharic		Afaan Oromo		Tigrinya		Ge'ez		English	average		
shuffling strategy	*native*	*english*	*native*	*english*	*native*	*english*	*native*	*english*		*native*	*english*	*all*
Meta-LLaMA-3-8B												
3 random shuffle Avg.	26.86	26.98	31.54	25.77	29.3	26.05	30.34	27.67	50.73	29.51	26.61	30.58
All answers A Diff	+32.02	+46.93	+38.38	+54.31	+26.49	+54.21	+38.78	+49.96	+38.74	+33.92	+51.35	+42.20
All answers D Diff	-19.63	-17.04	-14.81	-20.79	-25.35	-23.16	-25.96	-23.2	-27.16	-21.44	-21.05	-21.9
Gemma-2-9b												
3 random shuffle Avg.	29.68	33.89	31.54	28.89	29.47	27.46	40.4	27.37	64.23	32.77	29.4025	34.77
All answers A Diff	+33.34	+47.27	+37.58	+59.76	+21.06	+60.17	+32.87	+58.68	+26.62	+31.21	+51.35	+42.20
All answers D Diff	-7.57	-20.23	-7.44	-24.51	-5.00	-22.46	-9.06	-22.32	-13.43	-7.27	-22.38	-14.67
Gemma-2-27b												
3 random shuffle Avg.	34.64	34.57	36.25	29.02	28.16	29.91	40.02	29.82	66.13	34.7675	30.83	36.50
All answers A Diff	+30.65	+35.00	+26.1	+45.48	+26.84	+43.77	+32.79	+45.18	+24.26	+29.09	+42.35	+34.45
All answers D Diff	-15.43	-13.87	-10.55	-19.26	-10.00	-22.8	-12.83	-21.14	-11.9	-12.20	-19.27	-15.30
Meta-LLaMA-3-70B												
3 random shuffle Avg.	41.94	40.3	35.13	31.51	38.16	30.18	61.44	30.26	74.52	44.17	33.06	42.60
All answers A Diff	+8.68	+12.91	+21.64	+18.89	+3.68	+28.5	+10.22	+28.42	+5.34	+11.06	+22.18	+15.36
All answers D Diff	-13.43	-18.97	-13.42	-16.77	-17.63	-20.71	-18.81	-20.00	-2.67	-15.82	-19.11	-15.82
LLaMAX3-8B-Alpaca												
3 random shuffle Avg.	33.95	25.33	31.48	26.29	30.79	27.63	33.18	27.19	39.51	32.35	26.61	30.55
All answers A Diff	+3.03	+39.06	+16.33	+45.82	-5.00	+51.84	+6.68	+53.34	+36.46	+5.26	+47.51	+27.50
All answers D Diff	+0.55	-14.77	-9.57	-21.91	+6.84	-23.16	-2.53	-22.98	-15.48	-1.17	-20.71	-11.44
Aya-101												
3 random shuffle Avg.	51.24	54.98	51.06	32.8	43.16	55.35	78.88	55.88	80.78	56.09	49.75	56.01
All answers A Diff	+10.74	+12.93	+3.52	+11.82	+8.16	+12.28	+6.83	+14.38	+1.56	+7.31	+12.85	+9.13
All answers D Diff	+5.99	+1.75	+0.93	-1.53	+7.37	-4.82	+2.92	-6.67	+0.00	+4.30	-2.82	+0.66
Gpt-4o												
3 random shuffle Avg.	54.13	66.39	76.73	80.35	44.30	42.28	87.48	76.62	99.46	65.66	69.41	69.75
All answers A Diff	-1.65	-8.21	+3.95	-1.27	+1.75	+6.67	+5.84	-23.99	+0.08	+2.47	-6.70	-1.87
All answers D Diff	-5.99	-8.21	-3.62	-4.25	-7.19	-2.28	-8.45	-36.88	-0.38	-6.31	-12.90	-8.58

Table 7: Zero-shot scores of task 1 (*meaning multiple choice task*) across all models for ***native*** choices and ***english*** choices. The first row shows the average across ***three different random shuffles*** of the choice order. We compare this with providing the ***correct choice at choice "A" (first choice)*** or providing the ***correct choice at choice "D" (last choice)***. For choices "A" and "D," the closer the numbers are to zero, the better since we don't see huge variance from the shuffle.

D Choice sensitivity

In Table 8, we present three different results based on the order of choices. To provide detailed insights, we have listed all outcomes, revealing that random shuffling yields consistent results. However, when the correct answer is consistently positioned as either the first option ('A') or the last option ('D'), we observe significant variations in performance.

	Amharic		Afaan Oromoo		Tigrinya		Ge'ez		English	Model Name	Choice Order
	native	english	native	english	native	english	native	english	native		
shuffle 1	27.48	24.22	32.07	25.90	27.11	28.95	32.26	29.74	51.03		random
shuffle 2	25.21	27.54	31.47	26.10	31.05	22.63	29.26	23.68	49.89		random
shuffle 3	27.89	29.19	31.08	25.3	29.74	26.58	29.49	29.58	51.26	Meta-LLaMA-3-8B	random
shuffle 4	58.88	73.91	69.92	80.08	55.79	80.26	69.12	77.63	89.47		A
shuffle 5	7.23	9.94	16.73	4.98	3.95	2.89	4.38	4.47	23.57		D
shuffle 1	30.79	31.47	30.88	28.09	29.21	31.58	37.1	31.58	64.53		random
shuffle 2	27.27	34.58	33.07	30.88	28.16	24.21	42.17	23.68	64.53		random
shuffle 3	30.99	35.61	30.68	27.69	31.05	26.58	41.94	26.84	63.62	Gemma-2-9b	random
shuffle 4	63.02	81.16	69.12	88.65	50.53	87.63	73.27	86.05	90.85		A
shuffle 5	22.11	13.66	24.10	4.38	24.47	5	31.34	5.05	50.80		D
shuffle 1	34.30	35.40	35.25	29.68	28.42	35.79	41.71	34.74	65.90		random
shuffle 2	31.61	34.99	37.65	28.88	26.58	26.05	36.64	26.84	67.96		random
shuffle 3	38.02	33.33	35.86	28.49	29.47	27.89	41.71	27.89	64.53	Gemma-2-27b	random
shuffle 4	65.29	69.57	62.35	74.5	55.00	73.68	72.81	75.00	90.39		A
shuffle 5	19.21	20.70	25.70	9.76	18.16	7.11	27.19	8.68	54.23		D
shuffle 1	41.53	37.47	33.86	31.27	39.47	30.26	61.06	28.95	75.29		random
shuffle 2	41.94	40.17	36.06	30.68	37.63	30.53	60.14	30.26	75.97		random
shuffle 3	42.36	43.27	35.46	32.58	37.37	29.74	63.13	31.58	72.31	Meta-LLaMA-3-70B	random
shuffle 4	50.62	53.21	56.77	50.4	41.84	58.68	71.66	58.68	79.86		A
shuffle 5	28.51	21.33	21.71	14.74	20.53	9.47	42.63	10.26	71.85		D
shuffle 1	33.88	23.81	30.88	26.69	30.79	31.05	30.18	29.74	38.67		random
shuffle 2	33.68	27.12	30.88	27.09	32.37	23.42	36.18	24.21	43.25		random
shuffle 3	34.30	25.05	32.67	25.10	29.21	28.42	33.18	27.63	36.61	LLaMAX3-8B-Alpaca	random
shuffle 4	36.98	64.39	47.81	72.11	25.79	79.47	39.86	80.53	75.97		A
shuffle 5	34.50	10.56	21.91	4.38	37.63	4.47	30.65	4.21	24.03		D
shuffle 1	50.00	55.62	50.20	33.47	42.89	57.89	78.57	57.89	80.09		random
shuffle 2	51.45	54.66	51.20	33.47	43.42	53.95	78.34	53.95	80.32		random
shuffle 3	52.27	54.66	51.79	31.47	43.16	54.21	79.72	55.79	81.92	Aya-101	random
shuffle 4	61.98	67.91	54.58	44.62	51.32	67.63	85.71	70.26	82.38		A
shuffle 5	57.23	56.73	51.99	31.27	50.53	50.53	81.8	49.21	80.78		D
shuffle 1	53.51	67.29	76.2	79.88	45.53	53.16	87.33	51.05	99.54		random
shuffle 2	53.31	65.01	76.69	80.28	42.89	47.11	87.10	90.09	99.31		random
shuffle 3	55.58	66.87	77.29	80.88	44.47	52.89	88.02	88.71	99.54	Gpt-4o	random
shuffle 4	52.48	58.18	80.68	79.08	46.05	48.95	93.32	52.63	99.54		A
shuffle 5	48.14	58.18	73.11	76.10	37.11	40.00	79.03	39.74	99.08		D

Table 8: Accuracy scores for *Task 1: Meaning Multiple Choice* task given three different randomly shuffled choices and when we make the choices first or last choice for the whole dataset.

E English Prompt Sensitivity for Multiple choice

Table Table 9 provides a detailed analysis of the model's sensitivity and accuracy when responding to three distinct prompts in English. The table is focused on assessing how well the model adapts to varying prompt formulations and maintains accuracy in its responses. Model outputs can be different depending on the prompt used in our task and on Table 9 and Table 10. We explored both native and English prompts.

	Amharic		Afaan Oromoo		Tigrinya		Ge'ez		English	Model Name
	native	*english*	*native*	*english*	*native*	*english*	*native*	*english*	*native*	
Prompt 1	26.65	24.22	31.67	25.9	27.89	28.95	32.26	29.74	51.26	
Prompt 2	23.55	26.29	33.27	24.7	25.26	29.74	29.49	29.95	52.63	Meta-LLaMA-3-8B
Prompt 3	23.97	24.43	32.67	25.5	27.63	30.79	28.57	28.11	44.39	
Prompt 1	30.79	31.47	30.88	28.09	29.21	31.58	37.1	44.01	64.53	
Prompt 2	30.79	28.78	26.89	26.29	29.47	30.26	36.41	44.93	59.04	Gemma-2-9b
Prompt 3	31.61	32.3	29.88	24.9	30.79	30.79	40.78	48.85	66.36	
Prompt 1	34.3	35.4	38.25	29.68	28.42	35.79	41.71	38.49	65.9	
Prompt 2	35.33	35.4	29.47	26.29	35.06	32.89	40.09	44.01	64.99	Gemma-2-27b
Prompt 3	35.54	38.1	37.25	27.09	33.68	33.16	43.78	45.62	73.46	
Prompt 1	41.74	37.47	33.47	31.08	37.11	31.05	61.06	49.77	75.06	
Prompt 2	42.36	40.17	32.87	27.69	36.58	30.79	57.14	48.39	75.06	Meta-LLaMA-3-70B
Prompt 3	40.91	35.2	31.67	25.1	35.79	31.05	47.00	44.70	64.99	
Prompt 1	29.13	24.02	31.27	26.1	28.42	28.95	34.33	31.34	39.36	
Prompt 2	27.27	26.64	32.67	25.7	27.63	28.95	33.87	30.65	44.16	LLaMAX3-8B-Alpaca
Prompt 3	30.58	25.47	31.87	25.5	31.58	26.84	36.87	30.18	44.62	
Prompt 1	50	55.69	50.2	33.47	42.89	57.89	78.57	83.64	80.09	
Prompt 2	48.35	54.24	50.6	32.27	41.84	56.58	74.65	83.87	79.41	Aya-101
Prompt 3	46.28	47.2	47.41	32.67	41.84	50.79	72.81	79.95	72.77	
Prompt 1	62.81	67.08	78.29	79.88	46.58	52.89	31.57	88.25	99.50	
Prompt 2	17.58	70.39	19.72	72.31	18.16	51.84	16.82	88.94	99.50	Gpt-4o
Prompt 3	0.00	1.24	0.00	0.20	0.00	0.26	0.00	0.69	70.71	

Table 9: English Prompt sensitivity Accuracy results for three distinct prompts.

F   Native Prompt Sensitivity for Multiple choice

Table 10 provides a detailed breakdown of results from three native (in-language) prompts used in a multiple-choice task. The prompts are designed in the respective native languages of the evaluation, and the task aims to assess the model's performance in understanding and responding correctly to multiple-choice questions.

	Amharic		Afaan Oromoo		Tigrinya		Ge'ez		Model Name
	native	*english*	*native*	*english*	*native*	*english*	*native*	*english*	
Prompt 1	33.06	26.71	27.09	25.3	27.37	26.84	26.96	24.19	
Prompt 2	34.09	26.29	24.9	24.7	27.11	23.68	26.73	25.12	Meta-LLaMA-3-8B
Prompt 3	27.48	26.29	26.69	24.9	26.84	24.74	25.58	21.66	
Prompt 1	34.09	38.72	25.3	27.09	28.95	27.11	26.5	26.04	
Prompt 2	28.31	36.85	24.9	26.49	28.16	26.84	26.04	26.5	Gemma-2-9b
Prompt 3	25.83	28.74	25.7	26.49	27.11	26.84	27.68	25.58	
Prompt 1	39.26	41.61	25.7	26.69	23.68	26.05	26.04	23.5	
Prompt 2	37.6	39.34	25.9	27.89	27.89	25	26.96	24.65	Gemma-2-27b
Prompt 3	38.22	37.68	25.1	26.1	23.95	25.53	29.95	26.96	
Prompt 1	28.1	29.81	27.09	27.09	27.37	26.58	25.35	25.35	
Prompt 2	26.03	27.95	28.09	24.9	27.11	27.63	25.35	25.12	Meta-LLaMA-3-70B
Prompt 3	24.59	23.81	26.1	25.3	25.53	27.11	28.11	25.12	
Prompt 1	31.4	24.64	26.69	26.49	27.29	25.79	27.19	25.58	
Prompt 2	31.4	25.65	25.5	27.09	27.11	26.32	26.04	24.65	LLaMAX3-8B-Alpaca
Prompt 3	27.48	28.16	26.29	26.49	27.11	25.79	26.04	25.12	
Prompt 1	51.03	53.00	43.03	27.29	48.16	32.11	30.41	30.18	
Prompt 2	53.72	57.35	41.24	29.28	50.00	32.63	30.41	30.65	Aya-101
Prompt 3	48.14	55.28	41.63	29.88	51.05	33.42	74.19	85.25	
Prompt 1	64.26	66.87	18.33	59.56	13.95	13.95	0.23	2.76	
Prompt 2	62.81	65.63	49.00	59.76	41.05	3.68	0.00	0.69	Gpt-4o
Prompt 3	6.2	14.29	14.34	46.61	19.47	5.00	0.00	0.00	

Table 10: Prompt experiments based on three distinct native proverbs.

G Generation Results

In Table 11 This table presents the performance evaluation metrics for Task 3, which involves generating proverbs. The table includes three key evaluation scores: ChrF, BLEU, and Translation Edit Rate (TER). These metrics are used to assess the quality and accuracy of the generated proverbs compared to the reference (ground truth) proverbs. Results show using BLEU score might be challenging for this tasks and ter doesn't tell us a clear picture of improvement.

	Amharic		Afaan Oromoo		Tigrinya		Ge'ez		English
	native	*english*	*native*	*english*	*native*	*english*	*native*	*english*	
Meta-LLaMA-3-8B									
ChrF	1.83	1.94	13.79	7.54	1.99	1.81	9.98	9.37	22.41
ter	1295.16	388.73	578.43	654.78	1041.28	270.42	907.84	224.33	447.44
BLEU	0.02	0.01	0.09	0.02	0.01	0.01	0.05	0.05	3.73
Gemma-2-9b									
ChrF	1.84	1.20	8.39	4.24	2.61	0.73	6.45	3.50	6.58
ter	1371.83	1298.30	950.79	2054.45	629.45	1145.38	1804.11	2416.84	2423.57
BLEU	0.02	0.00	0.04	0.00	0.02	0.00	0.02	0.00	0.77
Gemma-2-27b									
ChrF	1.34	1.21	8.41	10.17	1.72	1.03	8.44	8.97	23.18
ter	1015.48	900.26	459.48	435.41	862.70	753.32	417.14	258.83	528.82
BLEU	0.01	0.00	0.03	0.01	0.01	0.00	0.01	0.04	4.75
Meta-LLaMA-3-70B									
ChrF	2.23	2.74	10.12	5.73	3.72	3.03	11.30	6.80	21.61
ter	628.61	354.65	946.16	1389.60	346.14	375.99	813.35	697.51	370.52
BLEU	0.02	0.02	0.09	0.01	0.02	0.02	0.04	0.02	3.06
LLaMAX3-8B-Alpaca									
ChrF	5.29	4.90	18.11	10.16	3.38	2.54	9.73	9.06	31.25
ter	179.81	157.10	165.60	332.72	310.93	191.25	157.21	146.29	106.14
BLEU	0.06	0.05	0.24	0.05	0.03	0.01	0.04	0.04	13.68
Aya-101									
ChrF	6.44	5.58	19.17	4.70	4.71	2.80	9.51	8.62	19.17
ter	132.58	128.09	165.50	965.69	133.47	115.04	158.54	135.70	112.41
BLEU	0.37	0.14	0.61	0.01	0.14	0.03	0.04	0.05	4.34
Gpt-4o									
ChrF	5.63	0.03	16.94	3.27	6.38	4.70	6.00	3.88	50.00
ter	132.58	128.09	165.50	965.69	133.47	115.04	191.74	291.64	112.41
BLEU	0.37	0.14	0.61	0.01	0.14	0.03	0.05	0.02	4.34

Table 11: ChrF, Bleu, translation edit rate (ter) scores for Task 3: Proverb Generation Task