Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings

Anonymous ARR submission

Abstract

Large language models (LLMs) are highly adept at question answering and reasoning tasks, but when reasoning in situational context, human expectations vary depending on the relevant cultural common ground. As languages are associated with diverse cultures, LLMs should also be culturally-diverse reasoners. In this paper, we study the ability of a wide range of state-of-the-art multilingual LLMs (mLLMs) to reason with proverbs and sayings in a conversational context. Our experiments reveal that: (1) mLLMs "know" limited proverbs and memorizing proverbs does not mean understanding them within a conversational context; (2) mLLMs struggle to reason with figurative proverbs and sayings, and when asked to select the wrong answer (instead of asking it to select the correct answer); and (3) there is a "culture gap" in mLLMs when reasoning about proverbs and sayings translated from other languages. We construct and release our evaluation dataset **MAPS** (**M**ulticultur**A**l **P**roverbs and **S**ayings) for proverb understanding with conversational context for six different languages.

1 Introduction

Large language models (LLMs) have achieved impressive results on question answering and reasoning tasks (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022a, inter alia). However, when reasoning in situational context, human expectations may vary cross-culturally (Thomas, 1983, i.e., the pragmatic failure, the inability to understand 'what is meant by what is said') and depend on the knowledge of the relevant cultural common ground (i.e., the shared knowledge based on which people within a culture reason and communicate, including concepts, common sense, etc. Hershcovich et al., 2022). Understanding of such common ground in a cross-lingual setting is specifically understudied in NLP (Hershcovich et al., 2022) and neglected in existing LLM literature. As languages and cultures are intertwined (Kramsch, 2014; Hovy and Yang, 2021), it is crucial for models that serve all communities to be able to reason and communicate in a relevant way.

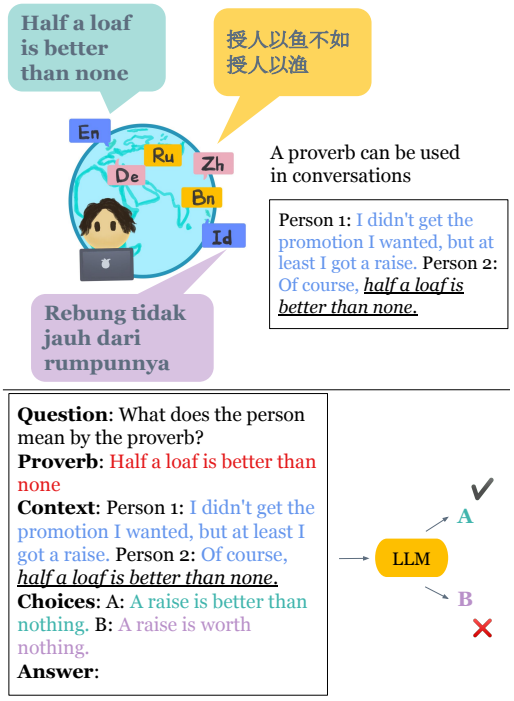


Figure 1: Proverbs are fixed expressions used by different cultures. We collect proverbs from six languages (top) and their usage within conversational contexts. We evaluate mLLMs with a binary-choice inference task in the conversational context that contains proverbs (bottom).

For these reasons, we focus on studying cultural common ground understanding of multilingual LLMs. Several questions arise: (1) Do mLLMs embed knowledge of cultural common ground, and does this knowledge affect their reasoning performance? (2) Can mLLMs reason in contexts that require an understanding of cultural common ground? and (3) Can mLLMs reason cross-culturally (i.e., about another culture's cul-

tural common ground, after translating into the same language) and are there gaps in the cultural knowledge (a "cultural gap")?[1]

In order to answer the above questions, we need to assess mLLMs using fixed, culturally-diverse expressions in multiple languages, that are also used flexibly in situational contexts. Fixed expressions are particularly important for evaluating the memorization of cultural common ground knowledge of LLMs. However, prior work focusing on multicultural concepts such as MaRVL (Liu et al., 2021, which is in multimodal) or MABL (Kabra et al., 2023) do not contain fixed expressions.

Proverbs and sayings (such as the ones illustrated in Figure 1) are fixed expressions that convey traditional wisdom, sometimes viewed as a form of folk literature, and grounded in living experience and social-cultural context (White, 1987; Mieder, 2004; Honeck, 2013). While different proverbs may emerge for different cultures, the underlying meaning of proverbs usually expresses universal human experiences. Yet, their literal expression and interpretation can vary from culture to culture (Honeck, 2013).

For example, the English proverb *The apple doesn't fall far from the tree* — means a child grows up to resemble his/her parents. While a plain version *like father like son* exists in many cultures, this proverb has a similar variant *Rebung tidak jauh dari rumpunnya* "Bamboo shoots are not far from the clump" in Indonesian, and 龙生龙，凤生凤，老鼠的儿子会打洞 "the dragon begets the dragon, the phoenix begets the phoenix, the son of a rat can make a hole" in Chinese. Of course, not all proverbs have parallels in different languages, as they are often culturally dependent.

Furthermore, proverbs are used in writing or conversational settings to offer advice, make arguments, or console others. A proverb's interpretation depends on the context (Mieder, 2004) it is used in and is often figurative, where the interpreted meaning does not entail the literal meaning. This makes them the ideal devices for studying the ability of LLMs to reason in situational contexts.

Hence, in this paper, we propose to use proverbs and sayings as a proxy for studying culturally diverse reasoning. In particular, we study (1) Do mLLMs know the proverbs and how well do mLLMs memorize them? (2) Can mLLMs choose the correct interpretation of a proverb given a situational context? and (3) Can mLLMs reason cross-culturally and are there cultural gaps in the interpretation of proverbs cross cultures?

We first present a dataset, **MAPS** (**M**ulticultur**A**l **P**roverbs and **S**ayings). The dataset consists of a collection of proverbs and sayings, an inference task for interpreting the meaning of proverbs in situational contexts (i.e., conversations), and binary labels indicating if the proverb is figurative. The dataset covers six languages with geographical diversity: English, German, Russian, Bengali, Chinese, and Indonesian.

We design a suite of experiments with **MAPS** for a wide range of *open source* state-of-the-art mLLMs. We find that mLLMs do possess knowledge of proverbs and sayings to varying degrees (significantly biased toward English and Chinese), and the amount of knowledge scales with model size. Through our inference task, we also find that the memorization of proverbs does not indicate better reasoning ability with proverbs, and figurative proverbs are more difficult for mLLMs to reason about in many languages. On the ability of mLLMs to reason cross-culturally with cultural common ground, we find that significant cultural gaps exist when reasoning with translations. Our results indicate that despite the apparent multilingual reasoning abilities of mLLMs, further research to improve the cultural-diversity (in terms of cultural common ground) of mLLMs is needed.

To summarize, our contributions are: **1)** we provide an analysis of the ability of a wide range of state-of-the-art open-source mLLMs to reason with cultural common ground, through the lens of proverbs and sayings; **2)** We disentangle the effects of memorization versus reasoning with proverbs and sayings, and reveal culture gaps in mLLMs; and **3)** We construct a multicultural dataset of proverbs and sayings for six different languages with multiple levels of annotations.

2 Related Work

Prior work has evaluated LLMs' ability for abstract reasoning (Ghosh and Srivastava, 2022, recognize proverbs from short stories) in English and assessed the models' ability for matching proverbs across three languages (BIG-bench authors, 2023,

[1]Reasoning with cultural common ground may be independent of language. For example, communications among different cultural groups within a multi-cultural country, or communication between L1/L2 speakers of a language where the L2 speaker has acquired the grammatical competence but not the cultural or pragmatic competence.

with a small evaluation set). To the best of our knowledge, **MAPS** is the largest multilingual dataset that focuses on proverbs and sayings, with conversational contexts and an inference task.

MABL (Kabra et al., 2023) is a task similar to ours, but focuses on multicultural novel metaphors understanding and cross-lingual transfer. It is less suitable to studying memorization vs. reasoning and does not study reasoning within a conversational context. Ruis et al. (2022) and Hu et al. (2023) use conversational context to study pragmatic reasoning in English LLMs and the identification of parallels between human and models, respectively. However, they provide limited insights beyond English. While we also use conversational context in our work, we focus on cultural common ground and multilingual aspects of mLLMs (with a larger dataset).

Finally, Haviv et al. (2023) aims to understand the memory-retrieval mechanism in LLMs with English idioms, which is a different in goals from this work.

3 MAPS - MulticulturAl Proverbs and Sayings

To help investigate our proposed research questions, we first present **MAPS** — a dataset of proverbs across six geographically and topologically-diverse languages. **MAPS** consists of: (1) proverbs and sayings, (2) conversational usages as context, (3) interpretations of proverbs (one correct, one wrong), and (4) labelling of whether the usage of the proverb is figurative or not (data examples in Table 2, Figure 6 in Appendix A.6 illustrates the annotation process).

3.1 Dataset Creation

Language Choices. We chose six languages for this dataset: English, German, Russian, Bengali, Chinese, and Indonesian. Several factors where considered when choosing the languages, including geographical diversity such as Eastern vs. Western (to increase the potential concept diversity), topological diversity, and resource availability (high-resource vs. lower-resource).

Proverbs and Sayings. We collect all proverbs and sayings (along with explanations) from Wikiquote[2] and Wiktionary.[3] Bengali has a significantly higher quantity of proverbs compared to other languages, thus, we perform a random subsampling of the proverbs for annotation to keep the final data roughly balanced.

Conversational Context. While proverbs and sayings are self-contained, they are typically used in conversations and writing to offer advice or console others. In order to investigate the ability of mLLMs to reason with proverbs, next we created short conversations that use proverbs (i.e., conversational context for the inference task).

To aid the data creation process, we use a human-model hybrid approach (i.e., model-in-the-loop), inspired by the recent work (Chakrabarty et al., 2022; Liu et al., 2023). We first use GPT3.5 (gpt-3.5-turbo-0301; a sibling model of Ouyang et al., 2022b) by prompting it with fixed templates to generate the seed conversational context (see Appendix B for the model templates).[4] Next, we ask two or more native speakers (experts or crowd, with least one expert per language) to either accept the model-created conversation, or write a new conversation if the human thinks the usage of the proverb is flawed.

In the final dataset, the conversational contexts for English, Chinese, Russian, and Bengali were completely re-written,[5] whereas for Indonesian and German, 22% and 20.5% of the original model-generated contexts were retained (the difference is probably due to variations in individual annotator preferences).

Interpretation of Proverbs in Context. We formulate this part as an inference task (following Liu et al., 2022). We ask annotators to create one correct answer and one wrong answer to the following question based on the conversational context:

What does the person mean by {proverb}?

Additionally, we also label the proverb if the interpretation is figurative (i.e., the interpreted meaning of the proverb is different from the expressed literal meaning).[6]

[2] https://en.wikiquote.org/

[3] https://www.wiktionary.org/

[4] The conversational contexts are in each perspective language, except for Russian and Bengali where the contexts are in English due to quality issues. For Russian and Bengali, the contexts are written in English first, then machine translated and fixed by native speakers for two rounds.

[5] The model has significant trouble in creating relevant context when the proverb is figurative. Anecdotally, human annotators found that the machine-generated context is helpful as a 'prompt', which helped to speed up the re-writes.

[6] "An apple a day keeps the doctor away" is a literal proverb that is advocating for apple consumption. "The apple doesn't fall far from the tree" is a figurative proverb where the literal meaning is about apples and a natural phenomenon,

Lang	Code	#Data (Test Size)	Class†
English	En	424 (394)	5
Chinese	Zh	364 (334)	5
German	De	364 (334)	5
Russian	Ru	420 (390)	4
Bengali	Bn	370 (340)	3
Indonesian	Id	371 (341)	3

Table 1: Dataset statistics. †: language class identified in Joshi et al. (2020), where 5 means the language is resource-rich.

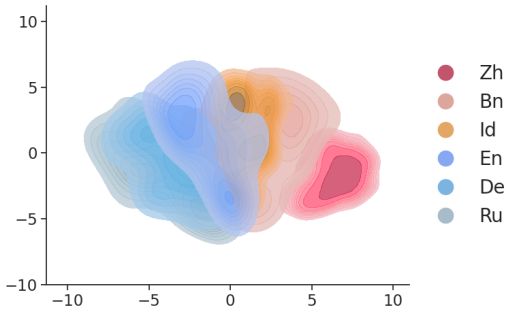


Figure 2: Visualizing proverbs embeddings using kernel density estimation (KDE).

Quality Control. Finally, we sampled 100 conversational contexts with their answers from each language. Then, we asked a separate set of native speakers to ensure the data quality for (1) correct usage of the proverb (i.e., the context is correct), and (2) correct answers for interpreting the meaning. Sometimes, it is possible to have more than one interpretation of a proverb given the context. We asked the native speakers to score the answers as correct as long as the answers aligned with one possible interpretation.

The final dataset consists of 2313 proverbs with conversational context. The statistics for each language are in Table 1 (with additional data statistics, in Table 7 in Appendix A). We further split the data for each language into a test set and a few-shot train-dev set (30 randomly-selected examples each). Table 2 shows examples from our dataset.

3.2 Analysis of MAPS

Proverbs and sayings are cultural artifacts and reflect embodied experiences, which contain diverse concepts often grounded with respect to real-world objects and experiences.

For instance, dairy product concepts (milk, cheese, yogurt etc.) exist in different languages but not in Chinese, whereas concepts that are symbolically meaningful in Chinese culture like dragons or phoenixes exist in the dataset. To illustrate this, we select interesting food items and animals from the final dataset (details in Table 5, Appendix A.2). From the data, we see for example that the tiger is a relatively important concept for Eastern cultures, whereas the lion is more important for Western cultures. Furthermore, we categorized the concepts in 100 sampled figurative proverbs in English, Chinese and Indonesian (see details in Appendix A.3, Figure 7). We observe that Indonesian has a lot more proverbs and sayings that use animals and elements in the nature than English.

We further encode the proverbs (without contexts) using multilingual sentence embeddings (Feng et al., 2022, LaBSE) and plot the embeddings with Kernel Density Estimate (KDE) (after dimensionality reduction to two components using tSNE van der Maaten and Hinton, 2012) to show the distinctiveness and connections between proverbs across different languages and cultures in Figure 2, which further illustrates that proverbs and sayings are culturally-diverse.

From Figure 2, we see that the embedding distributions are interestingly ordered from the West to the East. Indonesian proverbs have some overlap with English, probably due to the use of the Latin script and influences of foreign languages due to historical context. Chinese and Bengali proverbs are relatively distinct from the Western languages. Additional details for the KDE plot interpretation are in Appendix A.5.

4 Experimental Setup

We perform zero-shot evaluations and keep all prompt templates in English (on the test set), as previous studies show better performance with English prompts on mLLMs (Lin et al., 2022; Scao et al., 2022; Muennighoff et al., 2023).[7]

Models. We experiment with the following open source state-of-the-art multilingual models: (1) masked LMs: XLM-R (355m, 3.5B, Conneau et al., 2020); (2) encoder–decoder LMs: mT0 (580m, 3.7B, 13B, multitask and instruction tuned, Muennighoff et al., 2023); and (3) Causal

whereas the actual meaning of the proverb is about a child growing up to resemble his/her parents.

[7]For completeness, we also provide additional baselines using MAPS for cross-lingual transfer and few-shot evaluation in Appendix C and Appendix D.5.

Lang	Proverb	Context	Choices & Answer
Zh	授人以鱼不如授人以渔 (figurative)	A: 你可以帮我做这个项目吗? B: 当然可以，但是我觉得"授人以鱼不如授人以渔"。 (A: Can you help me with this project? B: Of course, but I think "it is better to teach a man fishing than to give him fish".)	A: B想帮A做项目而不是教A做项目。 (B wants to help A with the project instead of teaching A to do the project.) B: B想教A做项目而不是帮A做项目。 (B wants to teach A to do the project instead of helping A to do the project.) Answer: B
Id	Nasi sudah menjadi bubur (figurative)	Orang 1: Bagaimana reaksi bos-mu setelah kamu mengakui kesalahanmu? Orang 2: Kurang baik. Saya sudah mencoba menjelaskan alasan saya berbuat begitu, tetapi saya tetap diberi sangsi. Nasi sudah menjadi bubur. (Person 1: How did your boss react after you admitted your mistake? Person 2: Not good. I've tried to explain why I did this, but I'm still being penalized. The rice has become porridge.)	A: Orang 2 tidak dapat melakukan apapun untuk mengubah reaksi bos. (Person 2 can do nothing to change the boss's reaction.) B: Orang 2 masih bisa mengubah reaksi atasan. (Person 2 can still change the boss's reaction.) Answer: A

Table 2: Examples from selected languages (examples for all languages in Table 8, Appendix A.6).

LMs: BLOOMZ (560m, 3B, 7.1B, Muennighoff et al., 2023), and XGLM (564m, 2.9B, 7.5B, (Lin et al., 2022)). Most of the models cover all 6 languages in **MAPS** except BLOOMZ, which is derived from BLOOM (Scao et al., 2022) and does not cover Russian or German. In addition, despite being primarily an English model, Llama-2 (Touvron et al., 2023, Causal LM) has some multilingual capabilities. As a result, we decided to incorporate two Llama-2 models (7B, 13B) in our studies.[8]

Memorization Evaluation. Following previous work in assessing data memorization (Magar and Schwartz, 2022; Haviv et al., 2023; Carlini et al., 2023, 2021), we mask out the last word of each proverb and prompt the mLLMs to complete the proverb with templates in Table 9, Appendix B.

For the memorization task, let $t_i \in T$ be a prompt template, and let q_j be a proverb with n words where $q_j \triangleq \{w_1, w_2 \cdots w_n\}$. We remove the last word w_n for non-MLM models, if the LM generates (greedily) a string that starts with the missing token, or the entire proverb is a sub-string of generated string, then we count the model as having memorized the proverb. For the MLM model, we mask out the last word with '<mask>' and do predictions (i.e., $w = \arg\max_{w_n \in V} P(w_n | T_i; \hat{q}_j)$, where $\hat{q}_j$ is a proverb with mask token, and V is the vocabulary).

As the zero-shot prompting results are highly sensitive to the input patterns, we create 5 different prompt patterns (Table 9, Appendix B), and take the union of memorized examples among 5 patterns as the memorization accuracy.

[8]While larger models exist, we chose these models due to computational constraints. We can already see differences in performance at these model sizes.

Question: What does the person mean by the proverb?
Proverb: <proverb>
Context: <context>
Choices: A: <answer 1> B: <answer 2>
Answer:

Table 3: Zero-shot testing template, where the coloured part is the template.

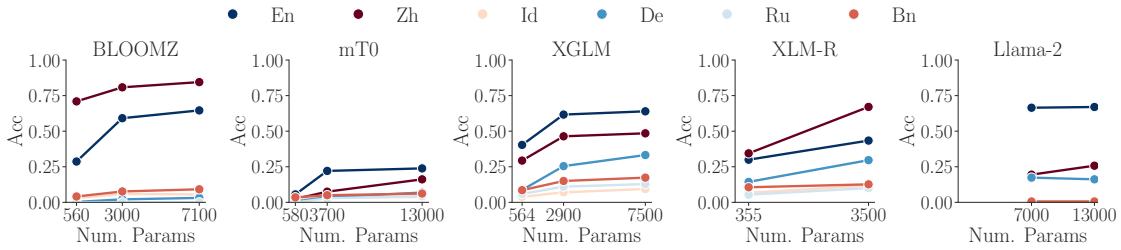
Reasoning Evaluation. For the inference task, we compute the correct answer by comparing logits of the two answer candidates ('A' or 'B') as in Lin et al. (2022). In particular, we use the prompt template t^r for this task (as in Table 3) and compute $P(t^r; q_i; \text{'}A\text{'})$ and $P(t^r; q_i; \text{'}B\text{'})$ and pick the larger one as the correct answer. For the MLM model, we compare the prediction logits of the answer candidates.

5 Results and Discussion

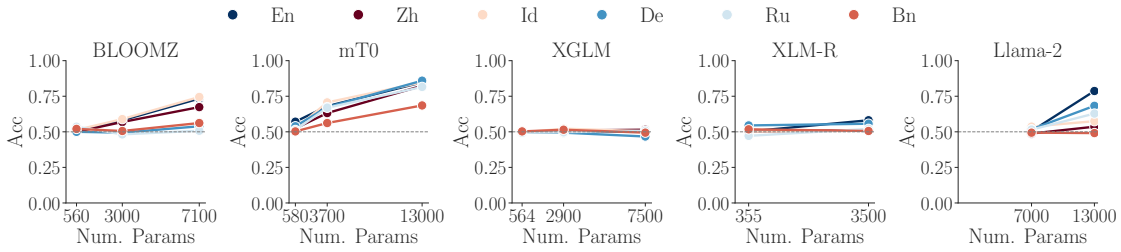
5.1 Knowledge of Proverbs

— A little knowledge is a dangerous thing.

Since proverbs are *fixed* expressions, successfully completing a proverb with greedy decoding likely means that the model has seen the proverb during pre-training. While it is possible that the proverbs in training data appears alone without any contextual usage or explanation, we consider



(a) Memorization of proverbs in different languages.



(b) Zero-shot results of proverbs understanding with context.

Figure 3: Performance of mLLMs on the proposed **MAPS** dataset. The number of parameters is in millions.

such an occurrence to be unlikely.[9] Hence, we make the assumption that memorization of the fixed expression also correlates with LLMs having embedded knowledge of the usage or meaning.

Figure 3a shows the results of proverb memorization, which (unsurprisingly) improves with model size. While XLM-R, XLGM, and mT0 cover all of the languages in our dataset, they don't score particularly well in memorization of proverbs in a single language. All models exhibit disparities in memorization across all languages, and these disparities are particularly pronounced in the case of Indonesian, Bengali, and Russian, which are lower-resource languages. These disparities are potentially due to data exposures, as we don't find any significant attribution, such as well-known versus less well-known, long versus short, or figurative versus non-figurative proverbs, by analyzing the memorized proverbs.

5.2 Reasoning of Proverbs with Conversational Context

— All that glitters is not gold.

While many models embed knowledge about proverbs, it is unclear if memorization translates to better reasoning with proverbs given the context. Next, we assess the models using our inference task.

[9]Webpages such as this https://en.wiktionary.org/wiki/no_pain,_no_gain exist in the training data for LLMs.

Memorization does not indicate ability to reason with proverbs. We prompt models with the pattern in Table 3 and plot the accuracy across languages in Figure 3b. In general, the bigger the model is, the better it performs on the inference task (i.e., the ability emerges with scale).

Overall, comparing the memorization curve and reasoning curve of mT0, XGLM and XLM-R, we observe that memorization does not indicate the ability to reason with proverbs in our experiments. In fact, model architecture has little effect (as BLOOMZ and Llama-2 are Causal LMs, and mT0 is an encoder-decoder model).

Since we know which proverbs are memorized from the previous experiments, we further break down the results into memorized vs. not memorized proverbs for the 3 best-performing models in English and Chinese (in Table 11, Appendix D.1). The benefit of memorization is evident in English, and shows inconsistency for Chinese (which aligned with observations for other languages in Figure 3b).

One possible explanation for the task not being heavily dependent on memorization is that contextual information aids inference, and the model may implicitly learn other culturally-relevant information from the training data during pretraining. Consequently, this suggests that LLMs may prioritize contextual information over memory retrieval when both are available. However, such a hypothesis requires rigorous study, which we will leave as future work.

Figurative proverbs are difficult to understand in general. Many proverbs are figurative, hence, we further divide the results of the model based on this property (described in §3). Looking at Table 4, we can see that, across English, German, and Russian, all models perform worse on the inference task when the interpretation is figurative. Interestingly, the opposite pattern is consistently observed for Chinese. Larger models appear to understand Indonesian and Bengali figurative proverbs better. One conjecture is that while abstract reasoning (the kind required for understanding figurative proverbs) can rely on memorization, but less memorization may lead to better abstract reasoning in LLMs.

Bias towards the correct answer amplifies performance gaps across languages. If the model genuinely understands a proverb's meaning in a situational context, it should be able to select the correct answer as well as the wrong answer when requested, especially for a task with only two choices. Several prior work has shown that negation in the natural language inference task weakens model performance (Hartmann et al., 2021; Truong et al., 2023; She et al., 2023). While not the primary focus of our work, this is a fundamental aspect of reasoning (Blanco and Moldovan, 2011) and we conducted experiments to verify. Here, we aim to ask a 'negative' question rather than provide negative answers. Hence, we change the question in the prompt template to *What does the person not mean by the proverb?*, while keeping everything else the same.

The results are in Figure 4. By simply asking the model to pick the wrong answer, all previous well-performing models now performing badly, except mT0 (which maybe due to the model being instruction-tuned). The 'negative' question enlarged performance gaps across languages as the model size increased. Additional results on asking the model to pick the wrong answer *without* using the word ***not*** are in Appendix D.2, where we observe consistent trends of model failures and inverse scaling in many cases. While we focus on the culture aspect of mLLMs, these results show fundamental work is needed to improve the ability for current mLLMs to handle 'negative' questions.

5.3 Culture Gaps in mLLMs - A Case Study

— When in Rome, do as the Romans do.

An ideal mLLM should perform on texts from all languages and translations in any directions equally well. However, in our experiments, the performance on English data is still stronger than other languages for most of the models we studied. Recently, several works have shown that good performance can be achieved by translating non-English text data in languages into English (Conneau and Lample, 2019; Yang et al., 2019, inter alia). Here, we demonstrate that when a task relies on cultural context, there are two distinct performance gaps to achieve true multilingual ability: one is the language gap (due to mistakes by the translation system, which may be fixed by a perfect translation system), and the other is the culture gap.[10] To demonstrate this, we use English and Chinese as the focus of a case study.

Machine Translation (MT). We translate every Chinese proverb, context and answers into English using Google Translate (Zh-En). By closely examining the translated data, it is evident that current machine translation (MT) systems do not handle cultural context well, producing incomplete or incorrect translations of proverbs. For example, a polysemous phrase 大三was translated to "junior" (third year university student), but in a specific proverbial context, it means someone is "three years older".

Human-Adapted Translation (HT). Next, we perform several adaptations to the machine-translated context: 1) manually correct any mistakes in the literal translation of proverbs, fix the grammatical errors in the contexts and answers; 2) conduct a light adaptation of the translated data inspired by Majewska et al. (2023), by replacing names and locations in the dataset to align with the culture (e.g., Xiao Ming to Michael etc.) in case models are confused about whether an entity is a person or a place. This represents our best-effort adaptation to reduce the language gap.

Next, we perform zero-shot evaluation with the best-performing multilingual models (mT0-XXL, 13B) and English model (Llama-2 13B) for Zh-En (in Figure 5). In fact, both models show a performance gap in the translated data compared to the target language. Interestingly, mT0 also shows a performance degradation comparing to the inference results in the original language (Llama-2 is near chance level for Zh, the improvement is not surprising). In all cases, HT improves over MT,

[10] This relates to cross-cultural pragmatic failure.

Model	Non-Figurative / Figurative					
	En	Zh	Id	De	Ru	Bn
BLOOMZ 3B	58.76/57.60	53.12/61.97	53.33/60.52	51.66/47.54	52.43/45.13	55.88/49.26
BLOOMZ 7.1B	79.66/68.20	66.66/68.30	72.00/75.18	54.30/53.55	52.43/49.55	67.64/53.30
mT0-XL (3.7B)	75.14/62.21	62.50/64.08	74.67/69.54	74.17/61.74	73.78/61.94	69.12/52.94
mT0-XXL (13B)	87.01/82.95	81.77/83.09	84.00/84.96	88.74/83.61	87.80/76.99	63.23/69.85
Llama-2 13B	81.36/76.50	53.12/54.23	54.66/58.27	72.19/65.03	67.07/59.73	47.05/49.63

Table 4: Zero-shot accuracy of non-figurative and figurative proverbs (Non-Fig./Fig.). The gray colour results indicate that the language is not officially supported by the model.

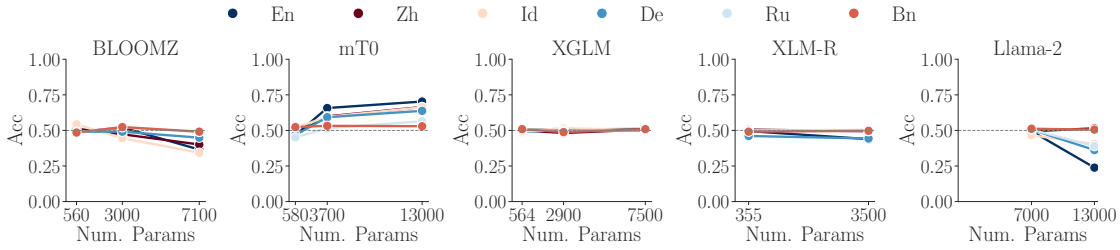


Figure 4: Performance of mLLMs on the proposed **MAPS** - Inference task when asking the 'negative' question.

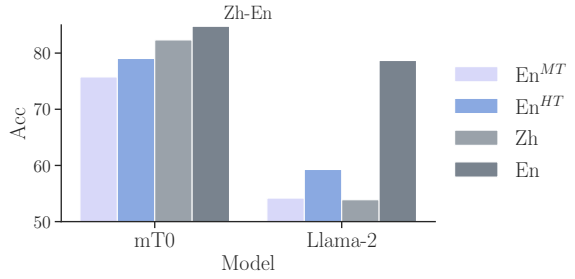


Figure 5: Performance gap between machine translated, human translated data and results in the original source language (Zh), and target language (En).

where the gain can be considered as the language gap. More interestingly, we define the gap between HT and the max of source and target language is the *culture gap* in mLLMs, i.e., *culture gap* = $|\text{Acc}^{HT} - \max(\text{Acc}^{Src}, \text{Acc}^{Tgt})|$. The culture gap for Zh-En is 5.73 for mT0 and 19.40 for Llama-2.[11] In an ideal situation, these gaps should be 0, indicating that the model is culturally aware and capable of understanding a language when speakers come from diverse cultural backgrounds. These results suggest that additional research is needed to improve cultural awareness and the inclusion of cultural priors in MT models and mLLMs (Yao et al., 2023; Shaikh et al., 2023).

[11] We also perform the same experiment in the reverse direction En-Zh with mT0 (Appendix D.3), similar results were observed. Other evaluation results on machine translated data for other languages with Llama-2 are in Appendix D.3.

6 Conclusion

In this work, we use proverbs and sayings from different languages as an investigative tool to assess the ability of mLLMs to reason with cultural common ground. Specifically, we study various mLLMs to evaluate their ability to memorize proverbs, reason with proverbs and sayings in different situational contexts, and understand cross-cultural communications using proverbs.

To aid the investigation, we present a multicultural proverbs and sayings dataset **MAPS**. Our analysis shows that many models possess knowledge of proverbs and sayings, however, knowing proverbs does not mean the model is able to reason with proverbs in contextual settings. Indeed, we found that mT0 shows some culturally-diverse reasoning ability, but only to a very limited extent. We also found that the ability to reason in a zero-shot manner emerges with model scale, but the ability to understand a 'negative' question inversely correlates with the model scale. The disparities in culturally-diverse reasoning ability between languages grow with the model size, which raises concerns in terms of multilingual availability and points to the need for more robust mLLMs. Finally, we defined and observed several culture gaps in cross-lingual communications. We hope to explore different aspects of cultural common ground in the future and to inspire novel work around mLLMs to facilitate inclusive cross-cultural understanding and communication.

7 Limitations

Our work uses proverbs and sayings as a proxy for cultural common ground, and we explore mLLMs' ability in understanding cultural common grounds in a limited setting. One potential limitation is we only collect one conversation per proverb or saying. Another limitation is the evaluation data is relatively small compared to many automatically generated benchmarks and may introduce lexical biases. However, these are not major concerns as 1) we want to focus on cultural common ground, which automatically limit us to a subset of lexical items (lexical biases is an intended feature); 2) to our best knowledge, this is the largest proverbs dataset for reasoning in context, and there is enough signal to distinguish between the tested models and uncover insights on current mLLMs ability and limitations in understanding proverbs and sayings. We hope to explore aspect of culture beyond proverbs and sayings, and with a more diverse set of languages (such as African languages or American indigenous languages) in the future.

In this work, we evaluate models of size up to 13B parameters (the biggest available size of mT0) due to computational constraints. However, full evaluation of larger models or task-specific models is necessary, especially when asking 'negative' questions and assessing the culture gaps in the future. Moreover, we focus on studying open-source LLMs in this paper for scientific reproducibility, and closed-source LLM evaluations are beyond our scope. As our dataset is publicly available at `anonymous_url`,[12] it can be used to evaluate closed-source LLMs in the future and we encourage others to do so.

References

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

[12] CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0/). We release the data and code. We will update the paper with the proper url after peer views.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Richard P Honeck. 2013. *A proverb in mind: The cognitive science of proverbial wit and wisdom*. Psychology Press.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore. Association for Computational Linguistics.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *CoRR*, abs/2304.14399.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Wolfgang Mieder. 2004. *Proverbs: A handbook*. Greenwood Publishing Group.

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

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

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *CoRR*, abs/2210.14986.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. 2023. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen

Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2012. Visualizing non-metric similarities in multiple maps. *Mach. Learn.*, 87(1):33–55.

Geoffrey M White. 1987. *Proverbs and cultural models: An American psychology of problem solving*. Cambridge University Press.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *CoRR*, abs/2305.14328.

A Dataset

A.1 Annotations

We recruit crowd annotators through Prolific[13] with the requirement of corresponding language as their first language, and fluent in English. Expert annotators are Master's, PhD and Post-doc researchers, including the authors of this paper. The annotation process is illustrated in Figure 6.

Instructions to create the conversational context:

> **Step 1: Check if the proverb is used correctly in the conversation.**
>
> Note: Sometimes, the proverb is figurative, meaning that the underlying meaning and the literal meaning of the proverb are different! The conversation should fits the figurative usage/meaning of the proverb.
>
> Example:
>
> *Person 1: "I'm scared of my boss." Person 2: "Well, barking dogs seldom bite."*
>
> "Barking dogs seldom bite" -It has a literal meaning of dogs that bark rarely taking actions and bite you, so you don't need to be afraid of getting hurt. The proverb metaphorically describes people that threaten you a lot rarely take actions and harm you. Although this conversation maybe missing some contexts, it should be labelled as correct.
>
> Example:
>
> *Person 1: "My dog is barking." Person 2: "Well, barking dogs seldom bite."*
>
> The proverb is used in a literal way, when it has a figurative meaning. This should be labelled as wrong.
>
> **Step 2: Re-write the conversation if the proverb is not used correctly from step 1.**
>
> The conversation should be 1-turn (1 round between 2 people), and maximum 2-turn (2 rounds between 2 people).
>
> Note: Please do not produce a conversation where one person is asking about the meaning of the proverb.

Instructions to create the answers:

> What does the person mean?

[13] http://prolific.com/

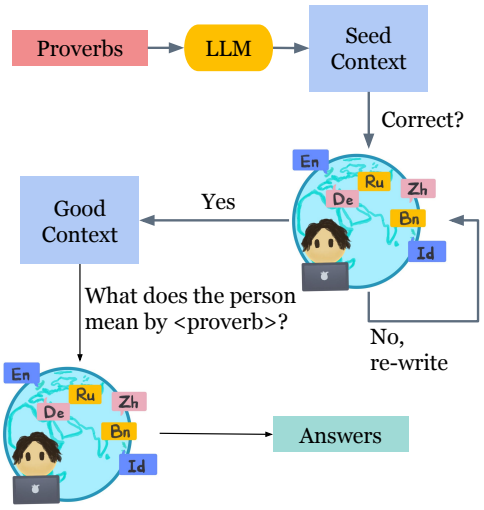


Figure 6: The data annotation process of **MAPS**.

> - Identify the person that used the proverb in the conversation.
> - Write down a short sentence in the OPT1 column, state what the person mean by the proverb in this conversation.
> - Write down a negative of OPT1 in the OPT2 column.

A.2 Animal and Food Terms in the Dataset

Table 5 shows selected animal and food concepts across different languages. From the data, we could see proverbs naturally contain culturally-important concepts. For examples, we could see the tiger is a relatively important concept for Eastern cultures, whereas the lion is more important for Western cultures; while bread is enjoyed by many people around the world, rice is culturally more important in the East, etc.

A.3 Additional Qualitative Analysis of Proverbs

We provide a qualitative analysis of how similar proverbs are expressed differently across languages and cultures. Similar to the ones in our introduction, many proverbs have a similar variant across cultures but are expressed differently. These proverbs differ by either using concepts that are familiar with the culture or using a local place name or person name (but this is very rare). Table 6 shows examples.

Next, when proverbs are figurative, different languages and cultures tend to use different types

Lang	Animals & Food
En	fox, wolf, lion bread, loaf, cookie
De	luchs, wolf, löwen, adler (lynx, wolf, lion, eagle) brot, kuchen, schinken (bread, cake, ham)
Ru	лиса, волк, сорока, соловья (fox, wolf, magpie, nightingale) хлеб, каравай, пирог, квас (bread, loaf, pie, kvass)
Bn	শিয়াল, হাতি, বাঘ, সাপ (fox, elephant, tiger, snake) চাল, ঘি, দই (rice, ghee, yoghurt)
Zh	龙，虎，凤（凰） (dragon, tiger, phoenix) 米 (rice)
Id	buaya, singa, harimau, merak (crocodile, lion, tiger, peacock) beras, sagu (rice, sago)

Table 5: Selected food and animal concepts from the proverbs.

Proverbs
En - When the cat's away the mice will play Id - Kalau di hutan tak ada singa, beruk rabun bisa menjadi raja (If there were no lions in the forest, the short-sighted monkey could become king.) Zh - 山中无老虎猴子称大王 (There are no tigers in the mountains, but the monkey is called the king.)
En - Rome wasn't built in a day Ru - Москва не сразу строилась (Moscow was not built in a day.)
Zh - 一山不容二虎 (One mountain cannot tolerate two tigers.) Bn - জলে তেলে খাপ/মিশ খায় না (Does not mix with water and oil.) Ru - Два медведя в одной берлоге не живут (Two bears don't live in the same den.)
En - Barking dogs seldom bite Id - Harimau mengaum takkan menangkap (The roaring tiger will not catch.)

Table 6: Parallel or closely related proverbs across different languages.

of concepts to draw parallels. We randomly sampled 100 figurative proverbs in English, Indonesian and Chinese, and classified contained concepts into one of the 5 categories, namely: Animals & Insects, Food, Cultural (including religious and spiritual entities, historical figures or names from the local culture), Nature (including metals, plants and other in-animated objects) and Others. Most of the time, a proverb only contains a single type of concept. However, when there are multiple types of concept, we pick the dominant one (such as part of the object of the sentence). The distributions are in Figure 7. Here, we observe noticeable differences in distributions across different cultures. There are more concepts related to Animals & Insects and Nature in Indonesian than the other languages, which is probably due to Indonesia's unique geographical location.

A.4 Additional Data Statistics

We include additional dataset statistics in Table 7. To calculate the average tokens in the context for Chinese, we take each character as a word.

Lang	Avg Tok in Context	Avg Turns
English	28.41	1.18
Chinese	31.30	1.14
German	27.91	1.12
Indonesian	25.35	1.15
Russian	31.25	1.47
Bengali	35.16	1.63

Table 7: Additional dataset statistics: average number of tokens in the context, and average turns in the context.

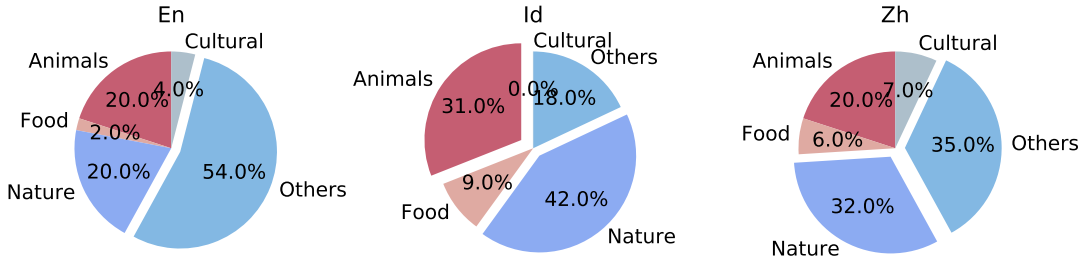


Figure 7: Distributions of concepts categories in figurative proverbs.

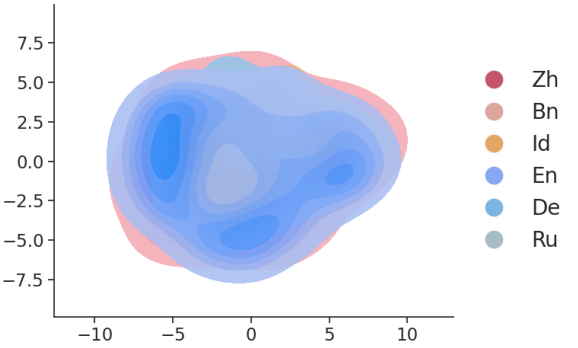


Figure 8: Visualizing embeddings with Kernel Density Estimate (KDE) when the sentences are sampled from a parallel dataset (topic coherent across languages).

A.5 Interpreting the KDE Plot

For better comparison, we produce the Kernel Density Estimate (KDE) plot of 400 randomly sampled sentences each language (2400 sentences in total), from a parallel multilingual dataset (Li et al., 2023a) in Figure 8. As the original data is much larger (67k sentences per language), subsampled sentences are likely not translations of each other, but rather topic coherent.

When sentences are topic coherent, their embeddings overlapping on top of each others and inseparable (Figure 8). In comparison with the KDE plot of proverb embeddings (Figure 2), we can clearly see the difference of proverbs across languages and cultures.

A.6 Data Examples

We balance the labels in **MAPS** and we show example data for all languages in Table 8.

B Templates

We use *Generate a very short 1-turn dialogue ends with "proverb" in language* as the template to query GPT3.5 (gpt-3.5-turbo-0301) for the seed conversational data. The model does not strictly generate seed conversation with 1-turn. We also experimented with translated template and did not observe quality improvements for our task.

Table 9 contains all the templates we used in our memorization experiments. As the prompting results are highly variable based on the input patterns, we created five different prompt patterns. We take the union of memorized examples among 5 patterns as the memorization accuracy.

C Cross-lingual Transfers Baselines

For completeness, we provide cross-lingual transfer baselines on **MAPS**. For cross-lingual transfer baselines, we re-split the English dataset into the train and test set (274/150 data point each), and evaluate on the original test set for other languages (i.e., same as zero-shot). We randomly sampled 20 data points from the training set as validation. We formulate the task as binary classification and experimented with XLM-R-Base (125m)/XLM-R-Large (355m)/XLM-R-XL (3.5B) and mT0-Base (580m)/mT0-Large (1.2B)/mT0-XL (3.7B).

The data input format is: *Context: {context} Choices: A: {answer 1} B: {answer 2}*.

We use AdamW optimizer (Loshchilov and Hutter, 2019) and conduct hyperparameter search of learning rate of [5e-5, 1e-4, 1e-5] and batch size of [8, 10, 16], trained for 30 epochs with bfloat16 precision, on a single A100 GPU.

The zero-shot transfer results are in Table 10 and averaged over 4 random seeds. The final hyperparameters for all models are [lr=1e-4, batch size=10], except for mT0-Large, which is [lr=1e-4, batch size=8]. Following previous work, we also include results for the translate-test baselines (Conneau et al., 2018) in Table 10.

Similar to our findings in the main paper, the model does not perform well on the task with models under billion parameters. The performance gap

Lang	Proverb	Context	Choices & Answer
En	half a loaf is better than none	Person 1: I didn't get the promotion I wanted, but at least I got a raise. Person 2: Of course, half a loaf is better than none.	A: A raise is better than nothing. B: A raise is worth nothing. Answer: A
Zh	授人以鱼不如授人以渔	A: 你可以帮我做这个项目吗? B: 当然可以，但是我觉得“授人以鱼不如授人以渔”。 (A: Can you help me with this project? B: Of course, but I think "it is better to teach a man fishing than to give him fish".)	A: B 想帮 A 做项目而不是教 A 做项目。 (B wants to help A with the project instead of teaching A to do the project.) B: B 想教 A 做项目而不是帮 A 做项目。 (B wants to teach A to do the project instead of helping A to do the project.) Answer: B
De	Es ist noch kein Meister vom Himmel gefallen	Person 1: Ich habe Schwierigkeiten beim Lernen dieser Sprache. Person 2: Mach dir keine Sorgen, es ist noch kein Meister vom Himmel gefallen. (Person 1: I'm having trouble learning this language. Person 2: Don't worry, no master has fallen from the sky yet.)	A: Eine Sprache zu lernen ist schwer und Person 1 sollte vielleicht mehr Zeit in die Praxis investieren. (Learning a language is difficult and Person 1 should perhaps invest more time in practice.) B: Eine Sprache zu lernen ist schwer und Person 1 sollte wahrscheinlich nicht mehr Zeit in das Üben stecken. (Learning a language is difficult and Person 1 probably shouldn't spend more time practicing.) Answer: A
Id	Nasi sudah menjadi bubur	Orang 1: Bagaimana reaksi bos-mu setelah kamu mengakui kesalahanmu? Orang 2: Kurang baik. Saya sudah mencoba menjelaskan alasan saya berbuat begitu, tetapi saya tetap diberi sangsi. Nasi sudah menjadi bubur. (Person 1: How did your boss react after you admitted your mistake? Person 2: Not good. I've tried to explain why I did this, but I'm still being penalized. The rice has become porridge.)	A: Orang 2 tidak dapat melakukan apapun untuk mengubah reaksi bos. (Person 2 can do nothing to change the boss's reaction.) B: Orang 2 masih bisa mengubah reaksi atasan. (Person 2 can still change the boss's reaction.) Answer: A
Ru	До свадьбы заживёт	Человек 1: О нет! Думаю, что я умру! Посмотри, как я порезал себе палец! Человек 2: До свадьбы заживёт. (Person 1: Oh no! I think I'll die! Look how I cut my finger! Person 2: It will heal before the wedding.)	A: Человек 1 не почувствует себя лучше. (Person 1 will not feel better.) B: Человек 1 скоро будет себя лучше чувствовать. (Person 1 will feel better soon.) Answer: B
Bn	বিপদ এড়িয়ে চলা বুদ্ধিমানের কাজ।	ব্যক্তি ১: আমরা কি এখানে সহজ রাস্তা টি নেব? ব্যক্তি ২: ওটা কিন্তু পাহাড়ের ধার ঘেঁষে যায়। ব্যক্তি ১: হ্যাঁ, কিন্তু আমাদের এক ঘন্টার পথ কমবে, আমরা কি সহজ রাস্তা টি নেব? ব্যক্তি ২: অযথা বিপদের মধ্যে যাওয়া যুক্তিযুক্ত নয়। (Person 1: Shall we take the easy way out here? Person 2: But it approaches the edge of the mountain. Person 1: Yes, but our journey will be less than an hour, shall we take the easy way? Person 2: It is not advisable to take unnecessary risks.)	A: তাদের বিপজ্জনক শর্টকাট নেওয়া উচিত নয়। (They should not take the dangerous shortcut.) B: তাদের বিপজ্জনক শর্টকাট নেওয়া উচিত। (They should take the dangerous shortcut.) Answer: A

Table 8: Examples for all six languages from **MAPS**.

between English and other languages remains significant.

D Additional Results

D.1 Memorized versus Not Memorized

We break down the results into memorized group versus not memorized group for the three best per-

Templates
1. **Proverb:** `no pain, no`
2. **Complete this proverb:** `no pain, no`
3. **Finish the proverb:** `no pain, no`
4. **What's the last word of this proverb:** `no pain, no`
5. **What's missing at the end of this proverb:** `no pain, no`

Table 9: Memorization templates, and the coloured part is the template.

forming models. We only show results when there are more than 50 proverbs in a group in Table 11 (which left us with English and Chinese). The benefit of memorization only shows for English, but not for Chinese.

D.2 '*Negative*' Questions.

We experimented with 4 additional versions of '*negative*' questions / instructions (randomly created), without the use of the word **'not'**, they are:

- Which answer is contrary to what the person means by the proverb?
- Which answer is impossible as the interpretation of what the person means by the proverb?
- Pick the opposite answer to what the person means by the proverb.
- Pick the wrong answer to what the person means by the proverb.

We use the same prompt template to evaluate the models. The results are in Figure 9. While our work focus on reasoning with cultural common grounds, this shows the importance and urgent need to improve model's ability in answering 'negative' questions.

We speculate this is due to the biases in training data. Often, users seeking for the correct solution to solve problems online (which we refer to as positive biases) rather than the wrong solution. Hence, when using web corpora as training data for LLMs, such positive biases will propagate to the behaviour of LLMs. To demonstrate this further, we conducted an additional experiment *without* asking a question in the prompt on BLOOMZ, mT0 and Llama-2. In an ideal situation, a good model should score nearly random when no question is asked (analogously to human confusion when data is given, but no question is asked). From Figure 10, all LLMs can score above random for multiple languages, which indicates all models *failed*. This failure mode further hints at the inability for mLLMs to handle negative question maybe due to the nature of the training data.

D.3 Culture Gaps

In addition to the results in §5.3, we follow the same procedure and perform the experiment with mT0 for En-Zh translated data. We observe similar results in Figure 11, and the culture gap for En-Zh is 5.33.

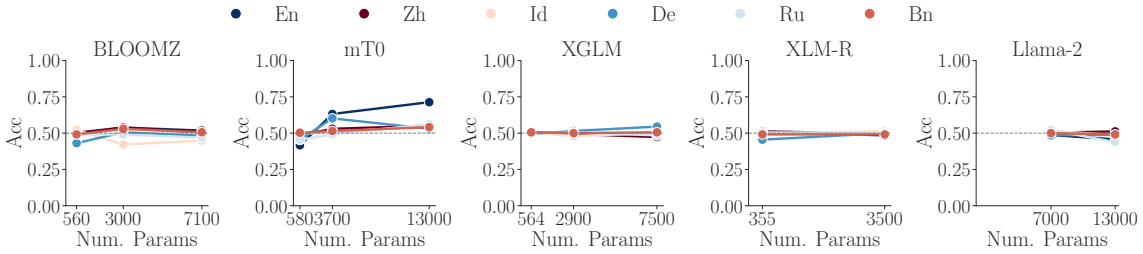
D.4 Additional Results on Llama-2 with Translations

Since Llama-2 13B is one of the recent state-of-the-art (English officially) models, we further conducted an zero-shot experiment by translating all date from other languages into English. We use Google Translate for translation and reported the results in Table 12. From the Table, we can see significant performance gaps (to English). It is also interesting to see the gaps increase as the corresponding geographical location of the language moves further away from English. While we consider this gap to be a combination of language gap and the defined culture gap, a future interesting direction is to closely examine the cultural gap in cross-cultural communications and how this is related to how LLMs internal representations are organized.

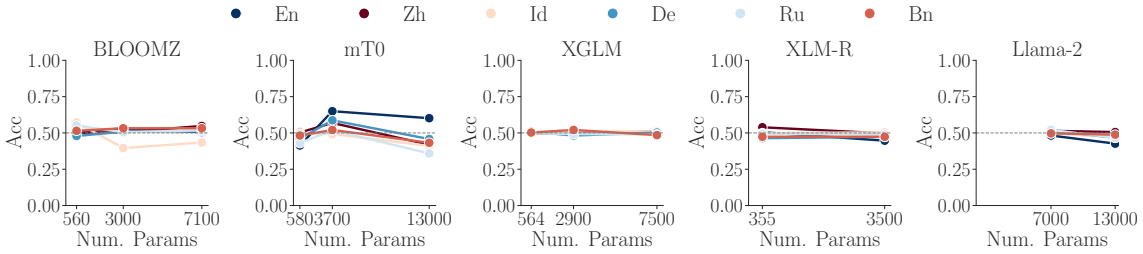
D.5 Few-shot (In-context) Evaluation

For completeness we also provide evaluation results with few-shot demonstrations. We perform 2-shot and 5-shot experiments by randomly sample 5 sets of n-shot demonstrations from the few-shot training set (using the same template as zero-shot evaluation by concatenation). We evaluate on BLOOMZ 7.1B, mT0-XXL 13B and Llama-2 13B models, and Table 13 shows the results.

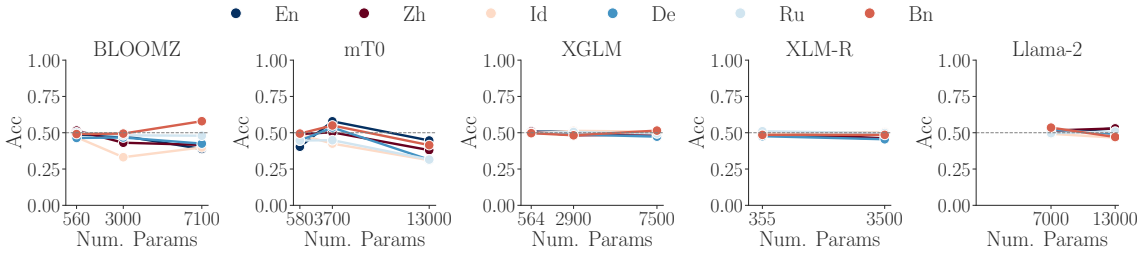
From Table 13, we do not observe any improvements with few-shot demonstrations comparing to zero-shot. In fact, model performances consistently degrade with more demonstrations. Since our task has very long context that may affects the n-shot performance. Nonetheless, this degradation has been observed recently in other work such as in Li et al. (2023b); Koto et al. (2023) with few-shot evaluations.



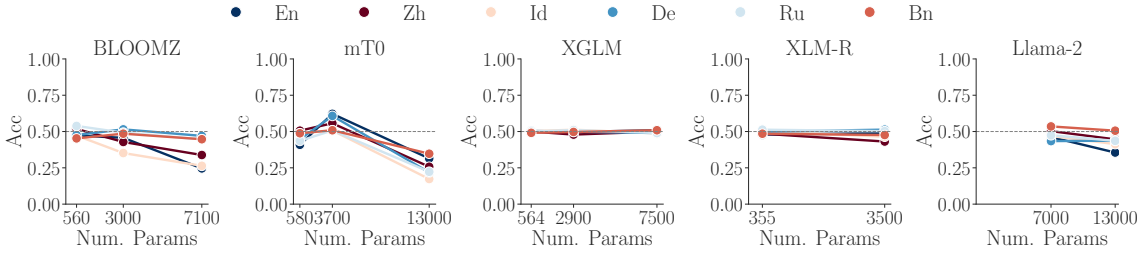
(a) Results using *Which answer is contrary to what the person means by the proverb?.*



(b) Results using *Which answer is impossible as the interpretation of what the person means by the proverb?.*



(c) Results using *Pick the opposite answer to what the person means by the proverb.*



(d) Results using *Pick the wrong answer to what the person means by the proverb.*

Figure 9: Performance of mLLMs on the proposed **MAPS** dataset when asking the model a 'negative' question.

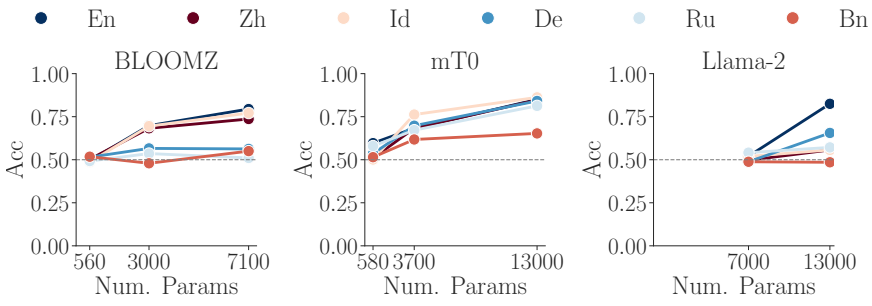


Figure 10: Performance of mLLMs on the proposed **MAPS** dataset when only the proverb, context and choices are provided, but without a question. Ideally, all models should score around random guessing.

Model	En	De	Zh	Ru	Id	Bn	Cross-lingual Avg
XLM-R-Base (125m)	52.06	50.00	50.07	50.19	50.37	50.22	50.17
XLM-R-Large (355m)	49.85	50.00	50.07	50.00	49.93	50.00	50.00
XLM-R-XL (3.5B)	58.38	53.67	52.25	53.65	52.79	53.01	53.07
mT0-Base (580m)	60.74	55.01	52.02	50.77	50.29	53.75	52.37
mT0-Large (1.2B)	65.00	56.89	56.59	53.53	50.44	55.59	54.61
mT0-XL (3.7B)	72.65	67.51	60.63	61.54	60.26	53.82	60.75
Translate-Test							
XLM-R-Base (125m)	-	50.60	50.75	49.23	51.47	49.85	50.38
XLM-R-Large (355m)	-	50.00	50.00	50.00	49.85	50.00	49.97
XLM-R-XL (3.5B)	-	50.90	51.20	52.31	49.85	51.47	51.15
mT0-Base (580m)	-	51.80	51.05	51.15	49.56	54.26	51.56
mT0-Large (1.2B)	-	54.04	55.09	54.62	53.67	57.21	54.93
mT0-XL (3.7B)	-	67.96	62.72	63.46	57.92	58.68	62.15

Table 10: Zero-shot cross-lingual transfers and translate-test baselines. Cross-lingual averages are calculated over all languages except English.

	En		Zh	
Model	∈Mem.	∉Mem.	∈Mem.	∉Mem.
BLOOMZ 7.1B	77.23	65.07	-	-
mT0-XXL (13B)	86.17	84.33	81.48	82.50
Llama-2 13B	80.30	75.38	54.65	53.22

Table 11: Result on memorized versus not memorized proverbs on 3 best performing models for English and Chinese. Results omitted due to less than 50 proverbs in the not memorized group.

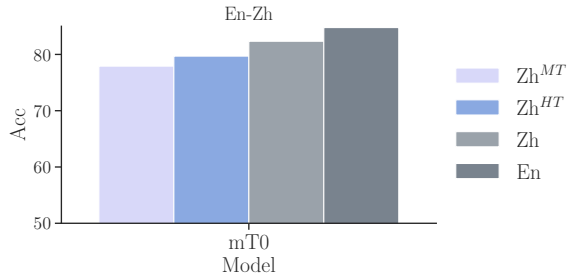


Figure 11: Performance gap between machine translated, human translated English data and results in the original source language (En), and target language (Zh).

Lang	Ori. Lang	MT	Δ_{En}
En	78.68	-	-
De	68.26	73.35	5.33
Ru	62.82	71.02	7.66
Id	57.47	69.79	8.89
Bn	49.11	61.76	16.92
Zh	53.59	54.19	24.49

Table 12: Results of machine translated data with Llama-2 13B. Δ_{En} is the result gap to model's performance on English data.

Model	En	De	Zh	Ru	Id	Bn	Cross-lingual Avg
BLOOMZ 7.1B 2-shot	59.49	61.55	56.59	53.77	51.53	50.00	52.65
BLOOMZ 7.1B 5-shot	51.57	52.39	50.85	50.35	50.25	50.52	50.30
mT0-XXL 13B 2-shot	78.37	72.63	76.95	78.74	74.87	63.82	76.81
mT0-XXL 13B 5-shot	68.48	67.90	70.38	71.50	67.64	60.00	69.57
Llama-2 13B 2-shot	74.87	56.52	55.42	60.77	56.76	51.00	58.77
Llama-2 13B 5-shot	64.16	52.69	54.89	55.56	52.71	50.17	54.14

Table 13: Few-shot evaluation results from **MAPS**. Cross-lingual averages are calculated over all languages except English.