Is Translation All You Need?
A Study on Solving Multilingual Tasks with Large Language Models

Chaoqun Liu[*12] **Wenxuan Zhang**[†23] **Yiran Zhao**[24] **Anh Tuan Luu**[1] **Lidong Bing**[‡5]
[1]Nanyang Technological University, Singapore;
[2]DAMO Academy, Alibaba Group, Singapore; [3]Hupan Lab, 310023, Hangzhou, China;
[4]National University of Singapore; [5]Shanda AI Research Institute
{chaoqun.liu,saike.zwx}@alibaba-inc.com; lidong.bing@shanda.com

Abstract

Large language models (LLMs) have demonstrated multilingual capabilities, yet they are mostly English-centric due to the imbalanced training corpora. While prior works have leveraged this bias to enhance multilingual performance through translation, they have been largely limited to natural language processing (NLP) tasks. In this work, we extend the evaluation to real-world user queries and non-English-centric LLMs, offering a broader examination of multilingual performance. Our key contribution lies in demonstrating that while translation into English can boost the performance of English-centric LLMs on NLP tasks, it is not universally optimal. For culture-related tasks that need deep language understanding, prompting in the native language proves more effective as it better captures the nuances of culture and language. Our experiments expose varied behaviors across LLMs and tasks in the multilingual context, underscoring the need for a more comprehensive approach to multilingual evaluation. Therefore, we call for greater efforts in developing and evaluating LLMs that go beyond English-centric paradigms.[1]

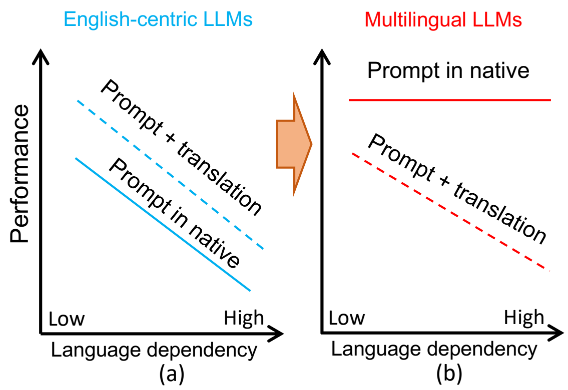


Figure 1: Illustration of two types of LLMs on tasks with varying language dependencies. "English-centric LLMs" refers to LLMs trained mainly in English corpora. "Multilingual LLMs" refers to ideal LLMs equally capable in all languages.

1 Introduction

Large language models (LLMs) frequently demonstrate the capability to understand and generate text across multiple languages, a skill attributed to their training on vast corpora composed of texts from various languages (OpenAI, 2023; Shi et al., 2022; Muennighoff et al., 2023; Jiang et al., 2023; Nguyen et al., 2023). However, these datasets are often disproportionately dominated by English content (Brown et al., 2020; Chowdhery et al., 2022; Workshop et al., 2023; Lin et al., 2022), resulting in an English-centric bias in LLMs. This imbalance can subsequently hinder the models' proficiency in other languages, often leading to suboptimal performance in non-English contexts (Ahuja et al., 2023; Lai et al., 2023; Zhang et al., 2023b).

To enhance performances in multilingual natural language processing (NLP) tasks with English-centric language models, translating training or test data into English has proven an effective strategy (Conneau et al., 2018; Ponti et al., 2020; Artetxe et al., 2023; Moghe et al., 2023; Bareiß et al., 2024). Recent investigations have expanded this idea by incorporating translation, either implicitly or explicitly, into the intermediate stages of prompting LLMs (Huang et al., 2023; Qin et al., 2023b; Etxaniz et al., 2023) for multilingual NLP tasks. For example, Shi et al., 2022 demonstrates that translating test questions into English enhances performance on multilingual reasoning tasks, as illustrated in Figure 2(a). Similarly, Huang et al., 2023 and Etxaniz et al., 2023 have shown that prompting LLMs to first translate or comprehend questions in English, then solve them step by step, improves

*Chaoqun Liu is under the Joint PhD Program between DAMO Academy and Nanyang Technological University.
†Wenxuan Zhang is the corresponding author.
‡Work done while at Alibaba Group.
[1]Our code is publicly available at https://github.com/DAMO-NLP-SG/translation-all-you-need.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9594–9614

performance.

Despite these advancements, methodologies in various studies differ significantly, and the impact of translation on multilingual task performance remains underexplored. Furthermore, these studies focus on specific NLP tasks and English-centric LLMs, but did not study real-world user queries in various languages. This gap highlights a need for more nuanced research into the effectiveness of translation techniques across multilingual contexts. As shown in Figure 1, we hypothesize that English-centric LLMs generally perform better with English translations of prompts, while "Multilingual LLMs" excel with native prompts, particularly for tasks highly dependent on language.

To address the limitations of existing empirical studies, we perform an in-depth analysis of the utility of translation with large language models for various scenarios. Firstly, we compare translating multilingual tasks into English, with an optional step of translating responses back into the original languages (i.e., the "translate-test" method), against several baselines on multilingual NLP tasks. Secondly, we extend the evaluation to real user queries, which are more likely to contain knowledge related to culture and language. Thirdly, we broaden the scope of LLM evaluations to include non-English-centric models to explore how they differ in behavior from English-centric LLMs. To the best of our knowledge, *this is the first work to analyze the impacts of translating real user queries on multilingual LLMs*.

Our results demonstrate that simply translating queries into English can already achieve the best results in multiple NLP task categories. For real user queries, the effect of translation depends on the languages and the LLMs. When working with advanced LLMs and certain languages, employing prompts in native languages appears to be the more effective strategy. In addition, the non-English-centric LLMs also behave differently from English-centric LLMs, where prompts in the native languages yield superior results by capturing the nuances related to culture and language.

The main contributions of this work are:

- We conduct a comprehensive comparison of multilingual prompting strategies in NLP tasks, finding that translation remains a strong baseline even for LLMs, and identifying factors impacting multilingual performance.
- We expand multilingual evaluation to include actual user queries and and non-English-centric LLMs, addressing the limitations of previous studies.
- We expose critical gaps in current multilingual evaluations, underscoring the need for more comprehensive benchmarks and a broader range of LLMs.

2 Translation for NLP Tasks

This section explores various prompting strategies across multiple languages and LLMs, covering a wide range of NLP tasks. This helps us understand how different prompting methods and other factors affect task performance.

2.1 Experiment Setup

2.1.1 Tasks

We conduct assessments on six benchmarks covering reasoning, understanding, and generation tasks that encapsulate various abilities of LLMs: **MGSM** (Shi et al., 2022), **XCOPA** (Ponti et al., 2020), **XNLI** (Conneau et al., 2018), **PAWS-X** (Yang et al., 2019), **MKQA** (Longpre et al., 2021) and **XL-Sum** (Hasan et al., 2021). Following Huang et al., 2023, we choose a subset of 9 languages for MKQA and 5 languages for XL-Sum. For evaluation metrics across our study, we employ the token overlap F1 score specifically for the MKQA dataset, the ROUGE-1 score for assessing XL-Sum, and accuracy as the standard metric for all other benchmarks. More details of the benchmarks can be found in Appendix A.1.

These tasks cover a wide array of 24 diverse languages, including German (de), Russian (ru), French (fr), Chinese Simplified (zh), Spanish (es), Japanese (ja), Italian (it), Vietnamese (vi), Turkish (tr), Indonesian (id), Swahili (sw), Arabic (ar), Korean (ko), Greek (el), Thai (th), Bulgarian (bg), Hindi (hi), Estonian (et), Bengali (bn), Tamil (ta), Urdu (ur), Telugu (te), Haitian Creole (ht), and Southern Quechua (qu). We categorize languages larger than 1% frequency in Common Crawl[2] as high-resource languages (i.e., de, ru, fr, zh, es, ja, it and vi), and the rest as low-resource languages. We exclude English since we want to evaluate the efficient prompting strategy for non-English tasks.

For each task, we sample 500 examples from the test set per language or use the entire test set

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

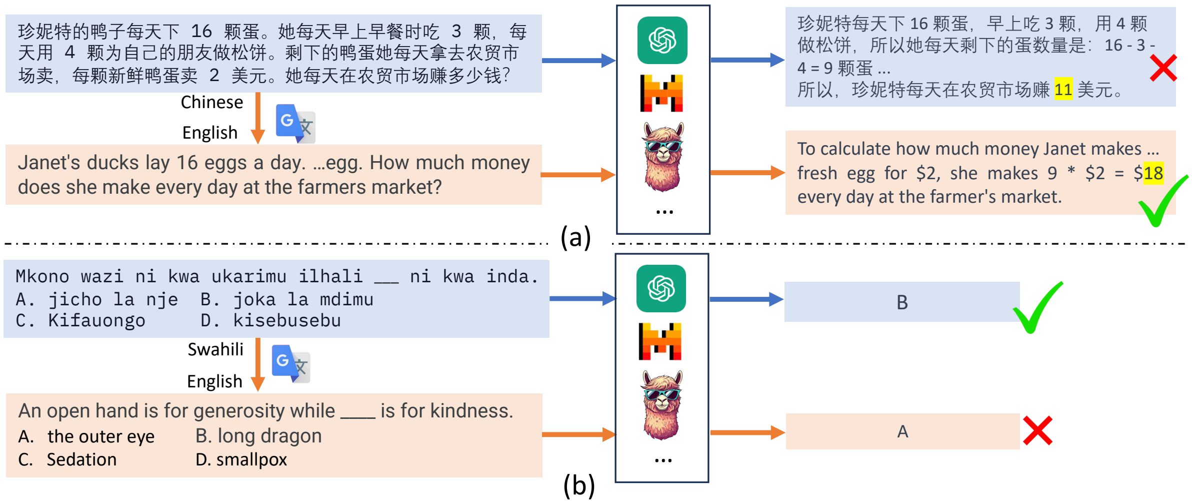


Figure 2: Examples illustrating how translation can both improve (a) and degrade (b) the performance of LLMs. The Chinese example is from MGSM (Shi et al., 2022) and the Swahili example is from M3Exam (Zhang et al., 2023a). Translation is beneficial when the questions are semantically equivalent across languages. However, for questions that demand deep cultural knowledge, translation can hinder the ability to answer accurately.

if there are fewer than 500 examples. For generation tasks like MKQA and XL-Sum, answers will be translated back to the original language if the prompting strategy uses a translator.

2.1.2 Models

We mainly conduct experiments on the following two LLMs, consisting of one closed-source language model and one open-source language model:

ChatGPT This is the most capable and cost-effective model in the GPT-3.5[3] family optimized for chat. We chose the latest version (gpt-3.5-turbo-1106) for the experiment.

Llama-2-70B-Chat This is the largest chat models in Llama-2 family (Touvron et al., 2023). Due to computational resource limitations, we use the AWQ (Lin et al., 2023) version for evaluation.

We also conducted experiments on some other models, including Mistral-7B-Instruct (v0.2) (Jiang et al., 2023), Llama-2-13B-chat (Touvron et al., 2023) and bloomz-7b1 (Muennighoff et al., 2023). More details are shown in Appendix A.1.

2.1.3 Prompting Strategies

We assess experimental strategies based on language of instruction, chain-of-thought reasoning, and translation tools, using a zero-shot approach as the selected models are fine-tuned for instruction-following.

[3] `https://platform.openai.com/docs/models/gpt-3-5`

Basic prompt with native instructions (NATIVE-BASIC) The questions are posed directly without using prompting strategies like chain-of-thought. Both the query and instructions are presented in their original language.

Basic prompt with English instructions (EN-BASIC) Compared with NATIVE-BASIC, EN-BASIC instructs LLMs with English but the query information is in the original language.

Native chain-of-thought (NATIVE-COT) In NATIVE-COT, we ask the question in the native language and ask the model to reason with the native language with the instruction "*Let's think step by step.*" translated into that language.

English chain-of-thought (EN-COT) We pose the question in the native language but instruct the model to reason in English with the instruction "*Let's think step by step in English*".

Cross-lingual-thought (XLT) XLT (Huang et al., 2023) is a state-of-the-art prompting method to handle multilingual NLP tasks. It prompts LLMs to translate the question into English and solve the problem step-by-step in English.

Translate to English with Google Translate (TRANS-GOOGLE) It uses Google Translate API to translate the original questions into English and then solve the problem step by step.

Translate to English with NLLB models (TRANS-NLLB) Instead of using commercial

Model	Prompt type	MGSM		XCOPA		XNLI		PAWS-X		MKQA		XL-Sum		AVG	
		high	low	high	low	high	low	high	low	high	low	high	low	high	low
ChatGPT	NATIVE-BASIC	44.4	19.4	84.6	69.7	56.9	48.6	51.6	40.6	35.1	36.4	32.5	29.9	50.8	40.8
	EN-BASIC	50.3	27.3	88.3	73.3	64.6	61.8	64.3	50.4	37.4	33.3	**33.3**	**30.0**	56.4	46.0
	NATIVE-COT	65.1	27.1	84.1	69.8	54.9	47.4	51.6	43.4	35.5	35.1	31.9	27.9	53.8	41.8
	EN-COT	70.5	47.1	89.9	75.9	60.2	53.6	63.7	51.2	**43.3**	41.2	30.0	28.6	59.6	49.6
	XLT	70.4	50.1	89.3	76.8	60.6	58.1	59.7	58.2	37.7	37.5	22.8	26.1	56.7	51.1
	TRANS-GOOGLE	**74.7**	**72.7**	**90.3**	**83.2**	**62.4**	**59.1**	68.2	62.0	42.5	**48.3**	30.6	28.9	**61.4**	**59.0**
	TRANS-NLLB	65.6	54.1	85.7	78.2	60.5	58.2	**68.4**	**63.4**	35.4	43.6	28.4	27.7	57.3	54.2
Llama-2-70B-Chat	NATIVE-BASIC	35.7	5.6	64.2	48.0	43.0	36.0	53.3	50.4	28.9	10.4	30.1	26.8	42.5	29.5
	EN-BASIC	42.5	7.7	70.7	52.0	52.7	41.9	61.9	52.8	25.7	21.5	30.2	35.3	47.3	35.2
	NATIVE-COT	35.5	5.6	65.3	46.8	41.0	35.6	56.0	49.6	25.3	9.9	26.0	25.2	41.5	28.8
	EN-COT	45.6	7.0	80.7	56.3	52.7	40.9	66.5	57.0	32.7	25.7	29.8	32.0	51.3	36.5
	XLT	49.0	8.4	76.4	54.7	**57.3**	48.4	56.6	51.6	26.5	26.7	19.3	11.5	47.5	33.6
	TRANS-GOOGLE	**55.5**	**50.0**	**86.3**	**79.7**	55.3	**53.0**	69.4	**64.2**	**38.7**	**43.1**	**33.1**	**36.7**	**56.4**	**54.4**
	TRANS-NLLB	46.5	39.7	83.3	75.6	53.7	51.0	**70.5**	62.4	17.8	24.7	32.4	36.2	50.7	48.3

Table 1: Average scores of the high-resource languages and low-resource languages for the six benchmarks in zero-shot setting. The best result for each model is in **bold**.

translators, we use an open-source model, namely NLLB (Team et al., 2022). Specifically, we chose `nllb-200-3.3B` to do the translation.

The examples for each strategy are shown in Table 4 and the templates for EN-BASIC are shown in Table 5 in the Appendix. In addition to the prompting strategies, an output constraint is also included in the template to facilitate answer extraction. When the output format may deviate from the instructions, we utilize "*Therefore, the answer <constraint> is*" in appropriate languages in the second round to retrieve the ultimate answer.

2.2 Main Results

The main results are shown in Table 1. We notice that TRANS-GOOGLE, despite simple, demonstrates the highest overall performance across various models and tasks. While it may not always achieve top performance, it consistently delivers commendable results for both high and low-resource languages. Besides this, we can have the following observations: 1) Utilizing English instructions generally enhances performance across various tasks, regardless of the integration of chain-of-thought. This finding aligns with those reported by Lai et al., 2023. 2) chain-of-thought is quite helpful for strong LLMs like ChatGPT and reasoning tasks like MGSM. For weaker models and tasks that can be answered directly, the basic prompt may be a better option. 3) On average, EN-COT underperforms compared to TRANS-GOOGLE for both high and low-resource languages. While EN-COT surpasses TRANS-NLLB in high-resource languages, it falls short in low-resource ones. We hypothesize that this discrepancy arises because LLMs excel in high-resource languages but need external translation systems to handle low-resource languages effectively.

These findings are also applicable to smaller models, such as Mistral-7B-Instruct, as demonstrated in Table 6 in the Appendix. This suggests that the observations generalize well across different model types and sizes. Further results and discussions are provided in Appendix A.1.4.

2.3 Analysis and Discussions

To investigate the impact of different factors on performance across various languages, we conduct a series of experiments and analyses using the MGSM benchmark.

Is there a relationship between task performance and translation quality? In addition to external translation systems, we can use LLMs to translate the questions. Although XLT includes translation, it is integrated into the solutions. Therefore, we examine the self-translate approach (Etxaniz et al., 2023), translating in a zero-shot manner with the prompt template shown in Appendix A.1.3. Then we prompt LLMs with the translated question the same as TRANS-GOOGLE and TRANS-NLLB. The results are shown in Table 8 in the Appendix.

We use the English subset of MGSM as the reference translation and evaluate translation quality using the SacreBLEU score (Papineni et al., 2002; Post, 2018). The results, shown in Figure 3, indicate that Google Translate achieves the highest quality for all languages except Japanese. Transla-

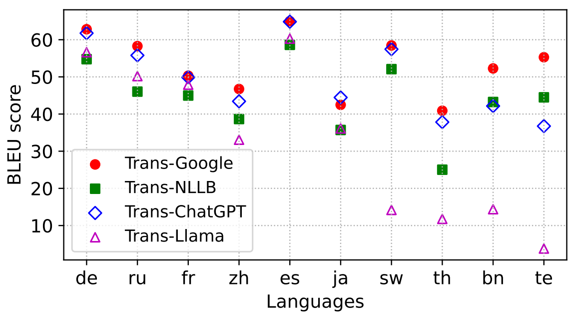


Figure 3: BLEU scores for translating MGSM questions with different translation systems.

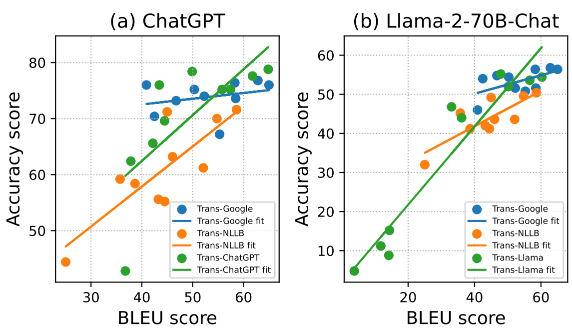


Figure 4: Corrections between BLEU scores of translation and MGSM accuracy for the three prompting techniques: TRANS-GOOGLE, TRANS-NLLB and self-translate. Each dot in the figure represents the performance of one model on one language.

tions by ChatGPT (Trans-ChatGPT) and Llama-2-70B-Chat (Trans-Llama) outperform TRANS-NLLB for high-resource languages but not for some low-resource languages.

To analyze the impact of translation quality on final performance, we plot the correlation between accuracy scores and BLEU scores for each language in Figure 4. The results show that higher translation quality (BLEU scores) generally leads to better task performance, highlighting the importance of an effective translation system.

Does language distance between English and target language affect the performances? Table 1 shows that the LLMs perform better for high-resource languages than low-resource languages on average. We hypothesize that language distance, besides language frequency, is crucial for English-centric LLMs. To verify this, we calculate the correlation between MGSM accuracy and the language distances between the target languages and English. Following Philippy et al., 2023, we examine five types of distances, including the syntactic (SYN), geographic (GEO), inventory (INV), genetic (GEN), and phonological (PHON) distances extracted using `lang2vec` (Littell et al., 2017). As shown in Table 2, MGSM accuracy significantly correlates with syntactic distance but not with other types of distances. The negative values indicate that languages with a larger syntactic distance from English tend to perform worse.

Prompt type	**SYN**	**GEO**	**INV**	**GEN**	**PHON**
`ChatGPT`					
NATIVE-BASIC	-0.786*	-0.336	0.323	-0.403	-0.044
EN-BASIC	-0.820*	-0.160	0.527	-0.299	0.020
NATIVE-COT	-0.795*	-0.184	0.479	-0.313	0.045
EN-COT	-0.841*	-0.286	0.339	-0.436	-0.034
XLT	-0.787*	-0.113	0.445	-0.284	0.117
`Llama-2-70B-Chat`					
NATIVE-BASIC	-0.688*	-0.369	0.250	-0.323	-0.044
EN-BASIC	-0.782*	-0.512	0.134	-0.513	-0.226
NATIVE-COT	-0.706*	-0.403	0.231	-0.475	-0.105
EN-COT	-0.737*	-0.510	0.206	-0.445	-0.219
XLT	-0.697*	-0.432	0.266	-0.423	-0.153

Table 2: Pearson correlation coefficient between MGSM accuracy and five language distances between English and that language. A lower value indicates higher correlation due to the negative coefficients.(*$p < 0.05$, two-tailed)

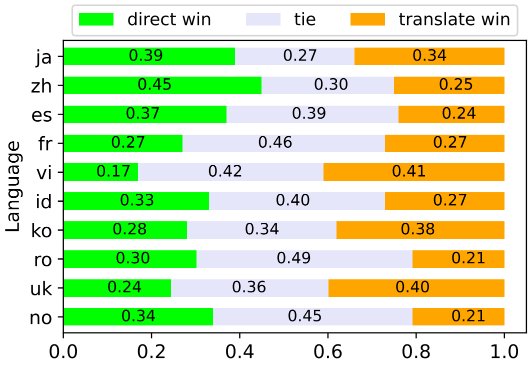
3 Translation for Real User Queries

NLP tasks typically focus on specific linguistic aspects, which may not fully encapsulate the breadth and complexity of real-world user queries which cover diverse topics and require nuanced comprehension. Moreover, these benchmarks are often constructed by translating from the English data (Shi et al., 2022; Ponti et al., 2020; Conneau et al., 2018; Yang et al., 2019; Hasan et al., 2021). This approach leads to datasets that are not truly challenging, as they miss the rich culture-specific elements crucial for truly nuanced language understanding for different languages. To assess the impact of translation on real-world queries, we extract user requests from ShareGPT[4], a website to share real conversations with ChatGPT.

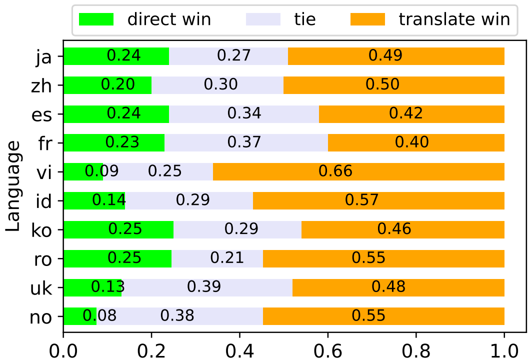
3.1 Experiment Setup

We selected 10 languages, ranging from high to low resource, and randomly sampled 100 requests for each language. However, for Romanian (ro), Ukrainian (uk), and Norwegian (no), we sampled 53, 98, and 53 requests respectively, due to the limited number of samples available from the source

[4] `https://sharegpt.com/`



(a) Win rate with ChatGPT



(b) Win rate with Llama-2-70B-Chat

Figure 5: Win rate comparison for each language using ChatGPT and Llama-2-70B-Chat.

dataset. Since the queries can be in various formats, we only compare two promoting strategies: 1) original queries; and 2) translated queries with Google Translate API. For the second option, we translate the output back to the original language for consistency. To evaluate the quality of the responses, we use GPT-4o[5](`gpt-4o-2024-05-13`) as the judge. The prompt for the judge is shown in Figure 8 in the Appendix, which is adapted from (Zheng et al., 2023). With this prompt, each response will get a score from 1 to 10.

3.2 Main Results

We compared the scores of two response sets from the same model, calculating the win rate for each language. The results are shown in Figure 5, leading to the following observations: 1) ChatGPT's performance varies across languages. For high-resource languages like Japanese, Chinese, and Spanish, original queries have a higher win rate. In contrast, for low-resource languages, the effectiveness of translation can be either better or worse, depending on the specific languages involved. 2) For Llama-2-70B-Chat, translation has a higher win rate for all languages, reflecting its English-centric nature. Despite potential information loss, the improved understanding after translation still enhances performance.

Llama-2-70B-Chat and ChatGPT exhibit distinct behaviors, reflecting their inherent differences. Llama-2-70B-Chat, being English-centric, performs better with translated inputs. Conversely, ChatGPT shows certain characteristics of a "Multilingual LLM", as shown in Figure 1(b), mainly for high-resource languages, indicating the potential for improvement in true multilingual processing.

To determine if answering user queries requires local cultural knowledge, we used GPT-4o with a specially crafted prompt to analyze queries in multiple languages (Figure 9 in the Appendix). Results in Table 14 in the Appendix show that 30% to 74% of queries per language require cultural knowledge, highlighting the rich cultural elements in the data. Further analysis of the ShareGPT subsets requiring local cultural knowledge is in Appendix A.2. We also conduct additional experiments, detailed in Appendix A.2.1, to verify that advanced LLMs can reliably assess the quality of responses.

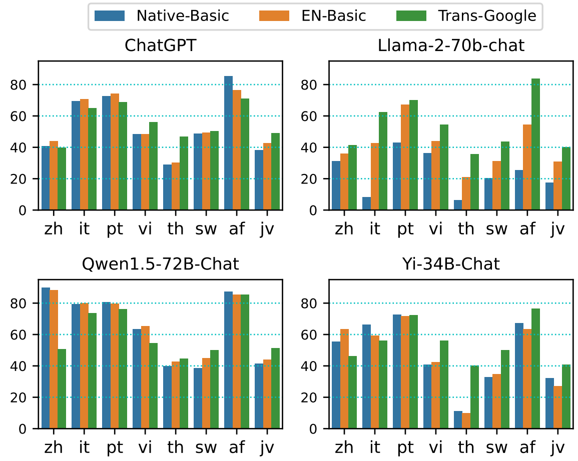
3.3 Analysis and Discussions

Based on the previous results, ChatGPT and Llama-2-70B-chat both tend to be English-centric but ChatGPT demonstrates certain behaviors of a "Multilingual LLM". Consequently, we broaden our analysis to include non-English-centric LLMs and assess their performance across various tasks.

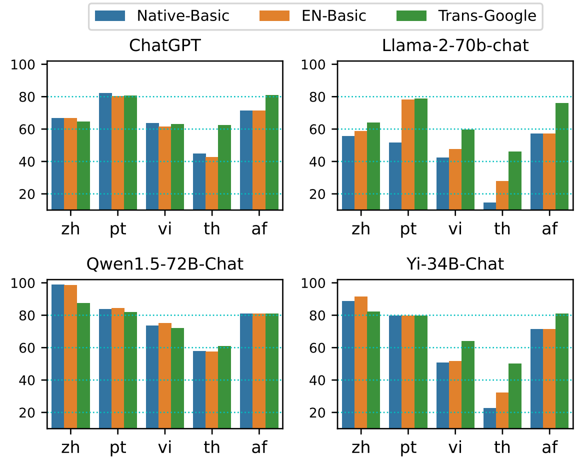
How do non-English-centric LLMs perform on culture-related tasks? To investigate the behaviors of different LLMs on culture-related tasks, we select another two LLMs: Qwen1.5-72B-Chat (Bai et al., 2023) and Yi-34B-Chat (AI et al., 2024), which are not English-centric. These two open-source models demonstrate strong capabilities in both English and Chinese. Therefore, we can check whether they demonstrate multilingual behaviors in Chinese, as illustrated in Figure 1(b).

For the evaluation dataset, we choose M3Exam (Zhang et al., 2023a), as the questions are real-world natural data from different languages instead of translating from English and require strong multilingual proficiency and cultural knowledge to perform well. For example, the question about a Swahili proverb in Figure 2(b) requires local knowledge to answer correctly. We select the `language`

[5] https://openai.com/index/hello-gpt-4o/



(a)



(b)

Figure 6: Accuracies of four LLMs on M3Exam (a) `language` and (b) `social science` subject categories. In M3Exam, not all subjects are available in every language, causing a difference in language coverage between the two subjects.

and `social science` subject categories, which likely contain more native cultural knowledge, and evaluate up to 500 samples per language.

Based on the results shown in Figure 6, we have the following observations: 1) For ChatGPT, translation may not always result in improved performance. This observation aligns with the conclusions in the study by Zhang et al., 2023a. The effectiveness of translation largely depends on whether translation errors outweigh any potential gains in better comprehension. 2) Translation helps Llama-2-70B-chat in all the languages, suggesting that the model's underperformance is due to poor language understanding rather than limitations of cultural knowledge. 3) Qwen1.5-72B-Chat and Yi-34B-Chat excel in Chinese proficiency. The translation hurts Chinese performance, highlighting the significant influence of translationese on comprehension. Despite this, it may boost performance in other languages, notably for Yi-34B-Chat, indicating that they are far from ideal multilingual LLMs.

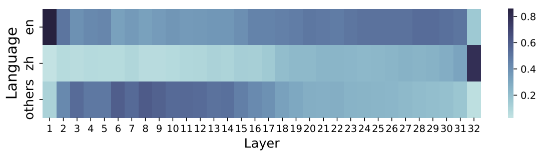
How do non-English-centric LLMs perform on NLP tasks? As shown in Figure 2(b), for an ideal multilingual LLM, prompting in native languages should still have advantages over translation if the tasks are less dependent on languages. To test the hypothesis, we evaluate Qwen1.5-72B-Chat and Yi-34B-Chat on the NLP tasks as discussed in Section 2.1.1. We only evaluate them in Chinese since the two models are optimized for this language.

The results are displayed in Table 3. TRANS-GOOGLE remains competitive among various prompting strategies, achieving the best average scores for Yi-34B-Chat, which surpasses our expectations. The possible reason could be that while both models are optimized for Chinese, their performance in Chinese still lags behind their proficiency in English. Nevertheless, We have the following special observations for the two models. 1) For Qwen1.5-72B-Chat, the best strategy is EN-COT instead of TRANS-GOOGLE. We hypothesize that this prompting strategy utilizes the model's bilingual abilities and simultaneously avoids translationese. 2) Both LLMs perform better with NATIVE-BASIC for the XL-Sum dataset. We hypothesize that the dataset is more language-dependent than other tasks as it is created by considering the local context instead of simply translating from the English version (Hasan et al., 2021). 3) The translation benefits are less pronounced than those of ChatGPT and Llama-2-70B-Chat. For example, the gap between TRANS-GOOGLE and NATIVE-BASIC on MGSM(Chinese) for the two models are 2.8% and 8%. The values for ChatGPT and Llama-2-70b-Chat are 37.2% and 16%, respectively, which are significantly larger.

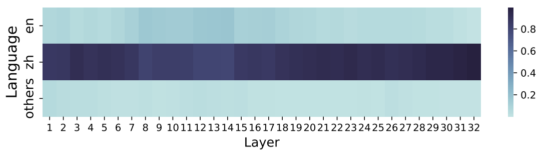
How do different LLMs handle multilingual prompts? To further understand the differences between English-centric LLMs and non-English-centric LLMs, we analyze the layerwise language distribution for Llama-2-7B-Chat and Qwen1.5-7B-Chat, using the method proposed by Zhao et al., 2024. We decode the embedding after each layer and identify each token into different languages

Prompt type	Qwen1.5-72B-Chat							Yi-34B-Chat						
	MGSM	XCOPA	XNLI	PAWS-X	MKQA	XL-Sum	AVG	MGSM	XCOPA	XNLI	PAWS-X	MKQA	XL-Sum	AVG
NATIVE-BASIC	78.8	93.0	55.8	71.8	36.6	**41.3**	62.9	63.2	92.6	46.0	43.6	13.4	**36.9**	49.3
EN-BASIC	77.2	97.0	73.0	**73.0**	32.7	39.7	65.4	66.8	93.6	52.6	**74.6**	15.5	35.1	56.4
NATIVE-COT	**83.2**	95.8	46.4	72.2	35.8	39.5	62.1	65.2	91.8	42.6	43.6	13.0	36.6	48.8
EN-COT	81.6	97.2	71.2	70.6	34.9	38.6	**65.7**	70.0	93.6	48.2	74.8	12.1	33.1	55.3
XLT	78.4	**97.8**	**77.4**	67.6	20.8	35.3	62.9	56.0	93.2	**69.2**	65.6	7.5	31.3	53.8
TRANS-GOOGLE	81.6	94.6	63.8	68.4	**45.7**	31.3	64.2	**71.2**	**94.0**	49.6	70.8	**24.5**	36.3	**57.7**
TRANS-NLLB	58.8	88.2	61.4	70.4	32.0	28.5	56.5	56.0	86.6	48.8	68.2	22.9	28.5	51.8

Table 3: Scores of the two non-English-centric LLMs on NLP tasks for the Chinese language. The best result for each model is in **bold**.



(a) Llama-2-7B-Chat



(b) Qwen1.5-7B-Chat

Figure 7: Layerwise language distribution for (a) Llama-2-7b-chat and (b) Qwen1.5-7B-Chat with Chinese prompts.

with CLD3[6]. As shown in Figure 7, the two LLMs process Chinese prompts differently. While the hidden representations of Qwen1.5-7B-Chat are mainly in Chinese, those of Llama-2-7B-Chat are in various other languages. We hypothesize that processing the information in native without conversion avoids the information loss, making it more suitable for processing multilingual tasks. In addition, we examine the layerwise language distribution in larger models, specifically Llama-2-70B-Chat and Qwen1.5-72B-Chat, as shown in Figure 12 within Appendix A.3.

4 Related Work

Multilingual Evaluation. Since the release of ChatGPT, the evaluation of LLMs has attracted the attention of the research community(Qin et al., 2023a; Bang et al., 2023). Shi et al., 2022 evaluated LLMs on MGSM and found that the models demonstrated strong multilingual reasoning capabilities, even for low-resource languages. Bang et al., 2023 evaluated ChatGPT on 23 datasets covering 8 NLP tasks. They found that ChatGPT failed to generalize its capabilities to non-Latin scripts. To cover tasks, Ahuja et al., 2023 evaluated ChatGPT and GPT-4 on 16 NLP datasets across 70 languages and compared them with state-of-the-art non-autoregressive models. Concurrently, Lai et al., 2023 evaluated ChatGPT on 7 different tasks across 37 diverse languages. However, these evaluations are primarily limited to standard NLP tasks and largely overlook real-world scenarios and cultural knowledge (Fung et al., 2024), which are crucial for understanding the practical applicability of LLMs.

Multilingual Prompting Strategies. The translate-test is a popular technique used to refine the performance of multilingual NLP benchmarks (Conneau et al., 2018; Ponti et al., 2020; Artetxe et al., 2023; Moghe et al., 2023; Qi et al., 2022; Huang et al., 2022). In the era of LLMs, various strategies have been developed to enhance the performance of LLMs using multilingual datasets. Shi et al., 2022 discovered that EN-COT outperforms NATIVE-COT. Huang et al., 2023 introduced cross-lingual-thought prompting to minimize language disparities. In parallel, Qin et al., 2023b introduced cross-lingual prompting, and Etxaniz et al., 2023 suggested self-translate to elevate their performances. Effective in translating prompts into English, these methods excel in NLP tasks but remain uncertain in real-world applications. Their success hinges on the English-centric nature of the LLMs. Our study evaluates translation effectiveness across NLP tasks, real user queries, and non-English-centric LLMs, revealing the limitations of these methods.

5 Conclusion

We have conducted a thorough evaluation of LLMs in various multilingual tasks. These tasks include traditional NLP benchmarks, real user queries, and culture-related tasks. Even though translation-based methods are simple and effective strategies

[6] https://github.com/google/cld3

to overcome the limitations inherent in English-centric LLMs, they are not optimal for all scenarios, highlighting the necessity of more comprehensive multilingual evaluation. The experiment on non-English-centric LLMs and culture-related tasks demonstrates that employing prompts in the native language emerges as a more effective approach. This method is particularly adept at capturing the subtleties and intricacies unique to each language. The challenge of the setting is that it requires LLMs to be proficient in various languages, calling for the prioritization of research and development efforts toward the creation of strong multilingual LLMs.

Limitations

This study aims to systematically assess the effectiveness of various prompting strategies across different tasks and LLMs. Due to limitations in computing resources, it was not possible to evaluate all existing prompting strategies comprehensively. However, we endeavoured to cover the most commonly employed strategies to formulate a broad conclusion. In our evaluation of LLMs on culture-related tasks, we specifically selected two LLMs optimized for Chinese, acknowledging it as one of the most widely spoken languages globally. The dataset used, M3Exam, comprises exclusively multiple-choice questions. It is important to note this specificity as it may influence the applicability of our findings. In our evaluation, we limited our sampling to up to 500 samples for each language within the benchmarks to manage computational constraints and ensure a broad yet feasible analysis scope. Consequently, our results might not be directly comparable with other studies that evaluate performance across the entire benchmark. In future work, we plan to extend our evaluation to LLMs optimized for other languages and to explore benchmarks presented in various formats beyond multiple-choice questions.

Acknowledgements

This research is supported, in part, by DSO Singapore under the research grant DSOCL23216. This research is also supported by DAMO Academy through DAMO Academy Research Intern Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. Chaoqun Liu extends his gratitude to Interdisciplinary Graduate Programme and College of Computing and Data Science of NTU, for their support. We sincerely appreciate the valuable feedback from Hou Pong Chan (DAMO Academy, Alibaba Group).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

01 AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open Foundation Models by 01.AI. ArXiv:2403.04652 [cs] version: 1.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting Machine Translation for Cross-lingual Classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. ArXiv:2302.04023 [cs].

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English Prompts are Better for NLI-based Zero-Shot Emotion Classification than Target-Language Prompts. In *Companion Proceedings of* *the ACM Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

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.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do Multilingual Language Models Think Better in English? ArXiv:2308.01223 [cs].

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. ArXiv:2402.09369 [cs].

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. ArXiv:2305.07004 [cs].

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot Cross-lingual Transfer of Prompt-based Tuning with a Unified Multilingual Prompt. ArXiv:2202.11451 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. ArXiv:2304.05613 [cs].

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, Chuang Gan, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration.

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 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*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406. Place: Cambridge, MA Publisher: MIT Press.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic Evaluation of Machine Translation Metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

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.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. SeaLLMs – Large Language Models for Southeast Asia. ArXiv:2312.00738 [cs].

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the Correlation Between Language Distance and Cross-Lingual Transfer in a Multilingual Representation Space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29. ArXiv:2305.02151 [cs].

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

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing Cross-lingual Natural Language Inference by Prompt-learning from Cross-lingual Templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023a. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? ArXiv:2302.06476 [cs].

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023b. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. ArXiv:2210.03057 [cs].

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. ArXiv:2207.04672 [cs].

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. ArXiv:2211.05100 [cs].

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.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. ArXiv:2306.05179 [cs].

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do Large Language Models Handle Multilingualism? ArXiv:2402.18815 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ArXiv:2306.05685 [cs].

A Appendix

A.1 Translation for NLP Tasks

This section presents more details about the setups and results for the experiments on NLP tasks.

A.1.1 Details about NLP Benchmarks

Here are the detailed descriptions of the NLP benchmarks:

Arithmetic Reasoning The MGSM (Shi et al., 2022) benchmark includes mathematical problems from grade school and requires the model to compute the accurate solution. It spans 10 languages, and we use the accuracy score for assessment.

Commonsense Reasoning The XCOPA benchmark (Ponti et al., 2020) consists of a single premise and two choices. The goal is to identify which choice is the cause or effect of the premise. It covers 11 languages from various families, with an accuracy score used for evaluation.

Natural Language Inference The XNLI (Conneau et al., 2018) benchmark includes one premise and one hypothesis. The model's job is to determine if the hypothesis is entailed, contradicted, or neutral based on the premise. It covers 15 languages, and we evaluate it using the accuracy score.

Paraphrase Identification The PAWS-X (Yang et al., 2019) benchmark consists of two sentences and requires the model to judge whether they are paraphrases. It covers 7 languages, and we assess based on accuracy score.

Question Answering The MKQA dataset (Longpre et al., 2021) contains open-domain questions that require predicting short answers. Questions that are unanswerable or excessively long to have a specific answer are not considered during evaluation. This dataset covers 25 languages, with our focus on 9 languages: de, es, fr, ja, ru, th, tr, vi, and zh. We assess the model's performance using the token overlap F1 score.

Summarization The XL-Sum (Hasan et al., 2021) benchmark requires the model to condense a lengthy news article into a brief summary. It covers 44 languages, and we select a subset of 5 languages: es, fr, tr, vi, and zh. We use the ROUGE-1 score for evaluation.

A.1.2 More LLMs for Experiment

Besides ChatGPT and Llama-2-70B-Chat, we have also evaluated the NLP tasks with the following models:

- Mistral-7B-Instruct (v0.2). This model is the instructed version of Mistral-7B (Jiang et al., 2023).
- Llama-2-13B-chat, which is a chat model in Llama-2 family (Touvron et al., 2023).
- bloomz-7b1, which is a model fine-tuned with multiple tasks, including some multilingual tasks (Muennighoff et al., 2023).

A.1.3 More Details about Prompt Strategies

An example of various prompting strategies is shown in Table 4. The prompts of EN-BASIC for each task are shown in Table 5, which are adapted from Huang et al., 2023. The translation template for self-translate with LLMs is:
Translate the following question from {language} to English:
{question}
Don't answer the question, just translate it!
The prompt templates for other prompting strategies and the instructions for output formats are designed according to the descriptions in Section 2.1.3.

A.1.4 Additional Results

The average performances for high-resource and low-resource languages are shown in Table 6. Table 7, Table 9, Table 10, Table 11, Table 12 and Table 13 shows the detailed results for MGSM, XCOPA, XNLI, PAWS-X, MKQA and XL-Sum, respectively. In addition to the finding in Section 2.2, We find XLT exhibits competitive performance in reasoning tasks; however, its performance in generation tasks is less impressive. Our findings indicate that when employing the XLT prompting strategy, ChatGPT declined to answer 26.4% of the questions in the XL-Sum tasks, responding with "*I'm sorry, I cannot ...*" This refusal pattern was not observed when utilizing other prompting strategies. For open-source models, while we did not observe a refusal pattern, they do not follow the instructions properly, which also degrades their performance with XLT.

A.2 Translation for Real User Queries

The prompt used to assess the response quality is shown in Figure 8. When GPT-4o is prompted with this, it assigns a score ranging from 1 to 10 to each response. Figure 9 illustrates the prompt used to determine if responding to a request requires local cultural knowledge. The Chinese case shows that GPT-4o can identify if queries require knowledge of local culture with explanations and the final answer.

We also analyzed the performance of shareGPT subsets with cultural knowledge only. As shown in Figure 10, the behaviors across languages and models are inconsistent. ChatGPT shows different behaviors for high-resource and low-resource languages. For high-resource languages like Japanese, Chinese, and Spanish, prompting with original

[System]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, expected language and level of detail of the response. Begin your evaluation by providing a short explanation (up to 100 words). Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "Rating: <rating>", for example: "Rating: 5".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]

Figure 8: The LLM-as-a-judge prompt for GPT-4o.

queries has a higher win rate. For low-resource languages, translation is often a better option. In contrast, Llama-2-70B-Chat shows a higher win rate for all languages.

A.2.1 Additional Results

In Section 3.1, we randomly select 100 requests for each language and evaluate the quality of the responses generated by GPT-4o. To ensure a more rigorous and comprehensive analysis, we conduct additional experiments under the following conditions: we heuristically filter queries using GPT-4o to ensure their validity, select 200 queries per language from the filtered set, and employ multiple judge models. Due to an insufficient number of available queries in other languages, we limit our evaluation to Japanese (ja), Chinese (zh), Spanish (es), French (fr), and Korean (ko). For the judging process, we use not only GPT-4o but also Claude-3.5-Sonnet and Gemini-Pro-1.5 to provide a more diverse assessment. The results are presented in Figure 11. ChatGPT performs better when given direct prompts in languages such as Japanese and Chinese, whereas Llama-2-70B-Chat consistently achieves higher performance with translated prompts. These findings align with those discussed in Section 3.2.

A.3 Layerwise Language Distribution in Larger Model

Figure 12 illustrates the layerwise language distribution in larger models, including Llama-2-70B-

Original Question	制作一件袍子需要2 匹蓝色纤维布料和这个数量一半的白色纤维布料。它一共需要用掉多少匹布料
NATIVE-BASIC	{**Original Question** } 您的最终答案的格式应为："答案: <阿拉伯数字>".
EN-BASIC	{**Original Question** } You should format your final answer as "Answer: <Arabic numeral>".
NATIVE-COT	{**Original Question** } 让我们一步步思考。 您的最终答案的格式应为："答案: <阿拉伯数字>".
EN-COT	{**Original Question** } Let's think step by step in English. You should format your final answer as "Answer: <Arabic numeral>".
XLT	I want you to act as an arithmetic reasoning expert for Chinese. Request: {**Original Question**} You should retell the request in English. You should do step-by-step answer to obtain a number answer. You should step-by-step answer the request. You should tell me the answer in this format 'Answer :'.
TRANS-GOOGLE	Crafting a robe requires 2 bolts of blue fiber cloth and half that amount of white fiber cloth. How many pieces of fabric will it use in total? Let's think step by step. You should format your final answer as "Answer: <Arabic numeral>".
TRANS-NLLB	To make a robe, two pieces of blue fiber and half of that amount of white fiber are needed. How many pieces of fabric does it take to make? Let's think step by step. You should format your final answer as "Answer: <Arabic numeral>".

Table 4: An example of zero-shot prompts for a Chinese problem. For NATIVE-BASIC, EN-BASIC, NATIVE-COT, EN-COT and XLT, we provide the original Chinese question as input and expect an answer in the corresponding format; for TRANS-GOOGLE and TRANS-NLLB, we input the translated question in English, and expect a step-by-step solution in English. To obtain the desirable output format, we instruct the models to output in specific format.

Chat and Qwen1.5-72B-Chat. Llama-2-70B-Chat exhibits the same phenomenon as its smaller counterpart, Llama-2-7B-chat, with diverse languages represented in its hidden states. In contrast to Qwen1.5-7B-Chat, the hidden representations of Qwen1.5-72B-Chat incorporate both Chinese and English until the last few layers, possibly reflecting the challenges of constructing such a large model using Chinese exclusively for hidden representations. Nevertheless, it still represents its hidden states more in Chinese than Llama-2-70B-Chat.

Benchmark	#Test	Basic Prompt
MGSM	250	`{problem}`
XCOPA	500	Here is a premise: `{premise}`. What is the `{question}`? Help me pick the more plausible option: -choice1: `{choice1}`, -choice2: `{choice2}`
XNLI	500	`{premise}` Based on previous passage, is it true that `{hypothesis}`? 1: Yes, 2: No, or 3: Maybe?
PAWS-X	500	Sentence 1: `{sentence1}` Sentence 2: `{sentence2}` Question: Does Sentence 1 paraphrase Sentence 2? 1: Yes, 2: No?
MKQA	500	Answer the question in one or a few words in `{target_language}`: `{question}`?
XL-Sum	500	`{article}` Summarize the article.

Table 5: Template of EN-Basic for each benchmark. #Test denotes the number of samples in the test set.

Model	Prompt type	MGSM		XCOPA		XNLI		PAWS-X		MKQA		XL-Sum		AVG	
		high	low	high	low	high	low	high	low	high	low	high	low	high	low
ChatGPT	Native-Basic	44.4	19.4	84.6	69.7	56.9	48.6	51.6	40.6	35.1	36.4	32.5	29.9	50.8	40.8
	EN-Basic	50.3	27.3	88.3	73.3	64.6	61.8	64.3	50.4	37.4	33.3	**33.3**	**30.0**	56.4	46.0
	Native-CoT	65.1	27.1	84.1	69.8	54.9	47.4	51.6	43.4	35.5	35.1	31.9	27.9	53.8	41.8
	EN-CoT	70.5	47.1	89.9	75.9	60.2	53.6	63.7	51.2	**43.3**	41.2	30.0	28.6	59.6	49.6
	XLT	70.4	50.1	89.3	76.8	60.6	58.1	59.7	58.2	37.7	37.5	22.8	26.1	56.7	51.1
	Trans-Google	**74.7**	**72.7**	**90.3**	**83.2**	**62.4**	**59.1**	68.2	62.0	42.5	**48.3**	30.6	28.9	**61.4**	**59.0**
	Trans-NLLB	65.6	54.1	85.7	78.2	60.5	58.2	**68.4**	**63.4**	35.4	43.6	28.4	27.7	57.3	54.2
bloomz-7b1	Native-Basic	1.6	0.9	36.5	18.9	3.7	11.8	-	-	7.1	10.5	-	-	12.2	10.5
	EN-Basic	1.9	2.2	67.5	55.2	**48.2**	40.7	-	-	11.8	6.5	-	-	32.4	26.2
	Native-CoT	1.0	1.4	37.9	17.3	1.2	13.5	-	-	5.2	11.1	-	-	11.3	10.8
	EN-CoT	1.7	1.6	61.3	52.8	37.6	34.7	-	-	10.0	6.9	-	-	27.7	24.0
	XLT	1.9	1.5	58.6	49.2	35.4	35.3	-	-	8.6	5.9	-	-	26.1	23.0
	Trans-Google	**2.5**	**3.0**	**67.5**	**62.8**	44.4	**44.2**	-	-	**15.6**	**23.0**	-	-	**32.5**	**33.2**
	Trans-NLLB	2.0	2.9	64.3	61.2	44.1	43.6	-	-	12.8	21.3	-	-	30.8	32.2
Mistral-7B-Instruct	Native-Basic	15.5	4.9	69.7	50.0	50.6	37.0	44.6	44.8	7.8	8.1	26.3	24.4	35.7	28.2
	EN-Basic	33.7	8.8	42.5	33.8	55.5	46.2	47.0	46.6	6.8	8.0	21.7	21.1	34.5	27.4
	Native-CoT	23.1	8.0	67.7	49.9	50.2	38.3	44.3	44.2	7.7	8.2	25.5	21.1	36.4	28.3
	EN-CoT	37.3	13.1	50.9	38.9	54.2	46.8	46.6	46.4	11.3	12.0	18.7	18.8	36.5	29.3
	XLT	**43.0**	15.0	**78.3**	**57.9**	48.4	44.3	47.9	47.2	9.4	10.4	17.1	19.6	40.7	32.4
	Trans-Google	42.6	**39.4**	67.0	57.5	**56.4**	**53.9**	51.4	52.0	**16.3**	**19.7**	**31.9**	36.5	**44.3**	**43.2**
	Trans-NLLB	32.3	30.8	62.1	52.3	54.4	51.9	**52.2**	**53.6**	14.5	19.3	31.0	**37.3**	41.1	40.9
Llama-2-13b-Chat	Native-Basic	22.7	4.9	59.5	48.4	39.9	33.7	55.2	48.2	20.7	9.6	28.4	23.8	37.7	28.1
	EN-Basic	28.7	4.4	63.9	51.6	48.2	39.8	**59.6**	**56.8**	20.9	17.8	31.3	30.2	42.1	33.4
	Native-CoT	26.9	4.9	59.0	49.3	38.6	33.5	56.2	47.8	17.9	7.8	28.4	22.7	37.8	27.7
	EN-CoT	29.5	5.5	68.2	51.0	46.2	41.8	57.8	56.6	20.5	17.3	30.7	28.0	42.1	33.4
	XLT	32.8	6.5	68.1	52.7	**56.9**	**47.3**	56.0	54.2	19.6	16.8	22.0	18.1	42.6	32.6
	Trans-Google	**38.4**	**40.1**	**77.8**	**70.4**	46.1	46.1	59.2	54.6	**32.6**	**37.8**	**35.1**	**38.0**	**48.2**	**47.8**
	Trans-NLLB	32.8	30.4	72.7	67.1	45.6	45.2	58.1	56.2	26.7	34.7	33.4	37.3	44.9	45.1
Llama-2-70B-Chat	Native-Basic	35.7	5.6	64.2	48.0	43.0	36.0	53.3	50.4	28.9	10.4	30.1	26.8	42.5	29.5
	EN-Basic	42.5	7.7	70.7	52.0	52.7	41.9	61.9	52.8	25.7	21.5	30.2	35.3	47.3	35.2
	Native-CoT	35.5	5.6	65.3	46.8	41.0	35.6	56.0	49.6	25.3	9.9	26.0	25.2	41.5	28.8
	EN-CoT	45.6	7.0	80.7	56.3	52.7	40.9	66.5	57.0	32.7	25.7	29.8	32.0	51.3	36.5
	XLT	49.0	8.4	76.4	54.7	**57.3**	48.4	56.6	51.6	26.5	26.7	19.3	11.5	47.5	33.6
	Trans-Google	**55.5**	**50.0**	**86.3**	**79.7**	55.3	**53.0**	69.4	**64.2**	**38.7**	**43.1**	**33.1**	**36.7**	**56.4**	**54.4**
	Trans-NLLB	46.5	39.7	83.3	75.6	53.7	51.0	**70.5**	62.4	17.8	24.7	32.4	36.2	50.7	48.3

Table 6: Average scores of the high-resource languages and low-resource languages for the six benchmarks in zero-shot setting. The results of PAWS-X and XL-Sum for bloomz-7b1 are not considered since it was already pre-trained on these tasks. The best result for each model is in **bold**.

Model	**Prompt type**	**de**	**ru**	**fr**	**zh**	**es**	**ja**	**sw**	**th**	**bn**	**te**	**avg**
ChatGPT	NATIVE-BASIC	48.8	42.8	42.8	36.0	50.0	46.0	30.8	21.6	15.6	9.6	34.4
	EN-BASIC	49.2	56.0	48.4	52.4	57.2	38.4	42.0	27.2	28.8	11.2	41.1
	NATIVE-COT	66.0	69.6	62.4	64.4	70.0	58.0	49.2	28.4	20.8	10.0	49.9
	EN-COT	74.8	72.4	71.2	67.2	75.2	62.0	58.0	51.6	52.8	26.0	61.1
	XLT	70.8	73.6	69.6	68.8	72.8	66.8	65.6	56.8	50.8	27.2	62.3
	TRANS-GOOGLE	76.8	76.4	75.2	73.2	76.0	70.4	73.6	76.0	74.0	67.2	**73.9**
	TRANS-NLLB	70.0	63.2	71.2	58.4	71.6	59.2	61.2	44.4	55.6	55.2	61.0
bloomz-7b1	NATIVE-BASIC	1.2	1.2	2.0	2.8	1.6	0.8	0.8	0.0	1.6	1.2	1.3
	EN-BASIC	2.0	1.6	2.4	2.8	1.6	1.2	2.0	1.2	3.6	2.0	2.0
	NATIVE-COT	0.0	0.4	1.2	1.6	1.6	1.2	2.4	0.4	1.2	1.6	1.2
	EN-COT	2.0	1.2	2.4	2.0	0.8	2.0	1.6	1.2	2.0	1.6	1.7
	XLT	0.8	1.2	2.0	3.2	1.6	2.4	2.0	0.8	0.8	2.4	1.7
	TRANS-GOOGLE	3.2	2.0	2.4	2.4	2.4	2.4	2.0	3.2	2.0	4.8	**2.7**
	TRANS-NLLB	2.4	1.6	3.2	0.8	2.0	2.0	3.6	2.4	2.4	3.2	2.4
Mistral-7B-Instruct	NATIVE-BASIC	7.6	14.4	12.0	19.2	30.8	8.8	4.0	4.4	6.8	4.4	11.2
	EN-BASIC	38.4	36.4	31.6	28.0	42.4	25.6	7.6	9.6	16.0	2.0	23.8
	NATIVE-COT	9.6	24.0	16.8	26.8	38.4	22.8	6.0	7.6	17.2	1.2	17.0
	EN-COT	39.2	42.0	36.0	33.6	42.0	30.8	8.0	21.6	18.4	4.4	27.6
	XLT	43.6	51.6	45.2	38.4	45.2	34.0	10.4	23.6	19.6	6.4	31.8
	TRANS-GOOGLE	42.0	46.8	41.2	44.0	42.0	39.6	38.8	35.6	42.0	41.2	**41.3**
	TRANS-NLLB	37.6	30.0	34.0	24.8	38.0	29.6	31.6	26.4	31.2	34.0	31.7
Llama-2-13b-Chat	NATIVE-BASIC	25.2	20.0	25.6	24.4	22.0	18.8	3.6	7.2	5.2	3.6	15.6
	EN-BASIC	32.4	26.4	32.0	26.0	34.8	20.8	3.2	5.6	5.6	3.2	19.0
	NATIVE-COT	29.2	23.6	29.2	27.6	28.4	23.2	2.8	7.2	6.4	3.2	18.1
	EN-COT	34.0	32.4	32.0	24.4	35.6	18.4	5.6	6.8	6.0	3.6	19.9
	XLT	34.4	34.4	33.6	29.6	37.2	27.6	4.8	8.4	9.2	3.6	22.3
	TRANS-GOOGLE	38.0	40.4	36.8	35.6	44.8	34.8	38.4	39.2	42.8	40.0	**39.1**
	TRANS-NLLB	29.6	33.2	38.8	31.2	28.0	36.0	32.0	24.8	35.6	29.2	31.8
Llama-2-70B-Chat	NATIVE-BASIC	34.8	28.4	38.8	38.8	41.2	32.0	4.4	8.4	7.6	2.0	23.6
	EN-BASIC	50.4	39.2	48.0	40.0	48.0	29.6	6.0	8.8	11.6	4.4	28.6
	NATIVE-COT	41.2	31.6	36.4	35.6	36.8	31.2	6.4	5.2	9.2	1.6	23.5
	EN-COT	49.6	48.0	50.0	38.0	48.4	39.6	7.6	7.2	10.4	2.8	30.2
	XLT	52.0	49.6	49.6	47.2	52.0	43.6	8.0	8.0	15.6	2.0	32.8
	TRANS-GOOGLE	56.8	56.4	54.4	54.8	56.4	54.0	51.6	46.0	51.6	50.8	**53.3**
	TRANS-NLLB	49.6	43.6	49.2	41.2	50.4	45.2	43.6	32.0	42.0	41.2	43.8

Table 7: Accuracy scores across various languages on the MGSM benchmark.

Model	**Prompt type**	**de**	**ru**	**fr**	**zh**	**es**	**ja**	**sw**	**th**	**bn**	**te**	**avg**
ChatGPT	Trans-ChatGPT	77.6	75.2	78.4	76.0	78.8	69.6	75.2	62.4	65.6	42.8	70.2
Llama-2-70B-Chat	Trans-Llama	53.6	52.0	55.2	46.8	54.4	44.0	8.8	11.2	15.2	4.8	34.6

Table 8: Accuracy scores across various languages on the MGSM benchmark with self-translate approach.

Model	Prompt type	zh	it	vi	tr	id	sw	th	et	ta	ht	qu	avg
ChatGPT	NATIVE-BASIC	88.0	91.8	74.0	81.4	85.4	77.2	65.2	85.4	49.6	63.4	50.0	72.3
	EN-BASIC	90.0	89.8	85.0	86.0	87.2	78.2	75.0	81.4	58.2	65.8	54.8	76.1
	NATIVE-COT	87.0	92.6	72.8	80.8	83.8	75.4	66.8	84.8	48.6	63.2	55.2	72.4
	EN-COT	90.4	92.2	87.0	89.6	90.2	85.6	74.8	85.8	61.4	69.2	50.2	78.6
	XLT	89.4	91.2	87.4	88.0	88.8	82.4	76.4	91.0	60.6	76.8	50.4	79.3
	TRANS-GOOGLE	90.8	91.6	88.4	85.8	88.8	79.4	82.6	88.2	85.6	81.6	73.2	**84.5**
	TRANS-NLLB	85.6	89.2	82.4	85.8	87.0	81.4	73.8	85.4	80.6	76.2	55.6	79.7
bloomz-7b1	NATIVE-BASIC	46.6	48.6	14.4	1.6	48.4	39.0	20.0	0.0	19.0	2.8	20.6	21.4
	EN-BASIC	78.2	55.6	68.6	50.2	62.8	56.8	49.6	50.0	71.4	50.0	50.4	56.5
	NATIVE-COT	43.4	50.0	20.2	0.6	48.6	23.0	39.2	0.0	17.6	0.0	9.4	20.9
	EN-COT	67.4	53.4	63.0	50.4	57.4	51.4	49.6	49.4	64.0	49.6	50.6	53.9
	XLT	63.8	49.6	62.4	45.6	64.0	49.0	51.2	46.0	52.8	48.0	36.6	50.5
	TRANS-GOOGLE	68.0	68.6	66.0	65.2	68.8	60.4	59.4	67.2	61.8	61.6	57.6	**63.7**
	TRANS-NLLB	64.0	67.2	61.6	63.6	64.6	62.2	57.4	62.8	62.8	61.6	54.2	61.8
Mistral-7B-Instruct	NATIVE-BASIC	67.2	82.2	59.8	55.0	65.0	47.6	51.8	36.6	49.2	51.2	43.6	54.2
	EN-BASIC	48.6	43.6	35.4	30.6	43.6	37.8	39.8	28.6	35.2	29.4	25.0	34.9
	NATIVE-COT	64.0	80.4	58.6	54.6	65.4	45.4	50.0	40.0	44.2	51.2	48.2	53.8
	EN-COT	55.8	52.2	44.6	43.8	52.2	39.8	46.0	32.6	29.2	39.4	28.2	40.8
	XLT	82.6	81.4	70.8	66.8	77.8	47.8	64.2	53.6	52.0	56.6	44.0	**61.5**
	TRANS-GOOGLE	69.4	64.8	66.8	61.0	68.8	52.2	62.0	60.8	59.8	52.0	43.6	59.2
	TRANS-NLLB	60.8	63.4	62.2	59.2	63.0	50.8	51.4	60.6	55.0	51.0	27.4	54.4
Llama-2-13b-Chat	NATIVE-BASIC	65.0	62.2	51.4	50.4	57.6	46.2	48.4	50.0	40.2	47.2	47.0	50.1
	EN-BASIC	61.2	74.2	56.2	52.8	62.0	52.0	50.6	50.6	50.2	46.4	48.4	54.3
	NATIVE-COT	62.8	64.6	49.6	53.8	64.8	49.8	51.8	45.4	32.6	49.8	46.6	50.9
	EN-COT	67.4	71.8	65.4	51.4	68.2	48.2	49.0	46.8	48.6	50.4	45.6	54.5
	XLT	65.4	72.6	66.2	57.2	70.0	47.0	49.2	50.8	50.2	50.6	46.6	56.0
	TRANS-GOOGLE	77.8	80.4	75.2	75.0	76.4	66.6	67.6	74.0	71.8	68.8	63.2	**71.9**
	TRANS-NLLB	73.0	75.6	69.6	74.4	73.2	67.4	62.4	73.8	66.2	68.0	51.2	68.2
Llama-2-70B-Chat	NATIVE-BASIC	61.6	81.6	49.4	49.4	55.4	50.6	46.8	49.8	41.0	46.4	44.6	51.5
	EN-BASIC	74.6	79.4	58.0	53.6	63.2	48.8	50.2	49.4	50.4	49.0	51.0	55.3
	NATIVE-COT	65.8	78.0	52.2	51.8	54.8	49.2	49.2	50.2	40.0	43.2	36.2	50.5
	EN-COT	80.4	88.0	73.6	65.4	77.8	53.0	50.0	56.0	48.0	49.8	50.6	61.2
	XLT	79.8	82.0	67.4	64.6	74.4	49.8	51.8	55.0	47.8	46.2	48.2	58.7
	TRANS-GOOGLE	87.2	88.0	83.6	82.2	89.4	76.6	77.6	83.4	83.6	76.4	68.4	**80.9**
	TRANS-NLLB	83.2	86.6	80.2	79.8	85.8	74.4	71.4	79.2	79.6	76.2	58.2	77.1

Table 9: Accuracy scores across various languages on the XCOPA benchmark.

Given the following user request, evaluate whether answering this request necessitates local cultural knowledge in that language. Consider factors such as cultural practices, beliefs, historical context, language nuances, and societal norms. Begin your evaluation by providing a short explanation (up to 50 words). After providing your explanation, provide a final assessment of whether cultural knowledge is essential for an accurate and comprehensive answer. Give you final answer in the format "Answer: [Yes/No]".

User Request:
{question}

Question:
1 / 1 将这段富有戏剧性的对话润色成一个剧本中的对话，要求总句数不少于 8 句，总字数不少于 200 字，内容有戏剧冲突和看点：\nA：你在这里做什么？我叫你来帮我搬东西，不是陪我聊天的。\nB：抱歉，我正在回复一个重要的邮件，马上就好。\nA：我又不是没看到你一直盯着手机，你有多少个重要的邮件？\nB：真的很重要，稍等一下，我处理完就马上来帮忙。

GPT-4o response:
Explanation: The request involves transforming a dramatic dialogue into a script, which requires understanding the nuances of the Chinese language, including tone, context, and cultural expressions. Additionally, the request specifies the need for dramatic conflict and appeal, which are influenced by cultural storytelling practices and societal norms.\n\nAnswer: Yes

Figure 9: Prompt template to check whether answering a request needs local cultural knowledge (upper) and one Chinese example (lower).

Model	Prompt type	de	ru	fr	zh	es	vi	tr	sw	ar	el	th	bg	hi	ur	avg
ChatGPT	NATIVE-BASIC	59.0	58.8	60.2	54.0	60.2	49.2	51.6	51.0	50.6	58.0	39.6	54.8	42.8	40.4	52.2
	EN-BASIC	68.6	58.2	67.4	62.2	68.4	63.0	65.6	65.2	62.4	64.6	56.4	65.4	55.8	59.0	**63.0**
	NATIVE-COT	59.4	54.2	58.0	51.8	58.6	47.6	53.0	50.8	51.2	54.6	37.2	54.4	40.2	37.4	50.6
	EN-COT	62.6	56.4	61.4	57.4	65.8	57.6	58.0	54.0	53.4	59.0	51.0	59.8	48.6	45.0	56.4
	XLT	63.0	57.8	61.4	58.4	63.4	59.8	61.4	58.0	57.8	60.4	55.0	59.6	53.2	59.2	59.2
	TRANS-GOOGLE	65.6	59.6	65.2	62.6	62.6	58.6	60.4	57.6	63.2	62.2	56.4	60.0	57.0	55.8	60.5
	TRANS-NLLB	63.4	62.2	61.6	57.4	62.8	55.6	59.4	58.8	62.4	63.4	54.2	61.6	52.8	53.0	59.2
bloomz-7b1	NATIVE-BASIC	0.4	13.4	0.2	6.6	1.4	0.0	6.8	18.2	1.6	5.2	26.6	15.4	17.8	2.8	8.3
	EN-BASIC	39.8	42.8	50.8	52.4	52.2	51.4	34.2	42.4	45.6	37.2	33.8	40.4	49.2	43.0	43.9
	NATIVE-COT	0.4	3.0	1.2	1.2	1.2	0.2	9.0	27.2	1.6	0.8	33.4	12.4	20.0	3.8	8.2
	EN-COT	36.2	35.2	37.4	42.2	37.4	37.2	33.2	34.8	36.2	33.6	33.2	34.2	37.6	34.4	35.9
	XLT	38.2	34.4	35.0	34.0	35.0	36.0	37.4	35.4	34.6	35.6	35.0	36.6	33.8	34.0	35.4
	TRANS-GOOGLE	45.0	43.4	44.2	44.0	45.2	44.8	43.8	44.0	44.0	44.6	44.4	44.8	43.4	44.4	**44.3**
	TRANS-NLLB	45.6	43.0	44.0	44.0	45.4	42.4	43.6	43.4	44.6	44.6	43.2	44.8	42.8	42.0	43.8
Mistral-7B-Instruct	NATIVE-BASIC	50.4	55.6	59.2	46.0	59.0	33.4	38.8	33.0	34.2	34.2	39.2	46.6	37.0	33.2	42.8
	EN-BASIC	56.4	54.6	59.8	54.0	56.8	51.4	46.8	37.6	45.8	49.4	47.0	54.4	46.4	41.8	50.2
	NATIVE-COT	50.0	55.0	58.4	47.6	54.6	35.8	38.2	32.2	37.6	35.4	40.0	52.0	36.8	33.8	43.4
	EN-COT	55.0	52.2	58.0	52.4	57.0	50.4	48.0	38.0	48.6	51.2	45.8	54.2	46.8	42.0	50.0
	XLT	48.2	44.6	49.6	49.4	52.4	46.0	48.0	39.0	42.2	46.4	45.4	46.6	44.0	42.6	46.0
	TRANS-GOOGLE	58.6	54.2	59.2	52.6	59.0	55.0	54.6	53.0	56.4	58.2	48.8	56.8	52.4	50.6	**55.0**
	TRANS-NLLB	57.0	52.4	55.8	50.2	58.2	53.0	54.2	49.4	53.0	56.4	47.4	55.2	50.0	49.6	53.0
Llama-2-13b-Chat	NATIVE-BASIC	41.4	40.2	44.0	38.6	42.8	32.4	34.6	31.6	32.8	34.2	34.0	37.4	31.4	33.6	36.4
	EN-BASIC	50.2	47.4	51.6	45.0	51.8	43.0	41.8	37.8	38.8	42.6	36.4	45.0	38.4	37.8	43.4
	NATIVE-COT	39.4	42.0	43.4	32.6	42.6	31.8	31.4	33.4	31.2	35.2	32.8	38.2	32.6	33.2	35.7
	EN-COT	45.6	46.8	48.8	44.4	46.6	44.8	41.8	38.6	43.2	43.4	38.6	46.2	42.0	40.8	43.7
	XLT	59.6	55.8	56.4	54.0	59.8	55.6	48.2	37.8	49.4	49.0	44.4	52.0	48.4	49.2	**51.4**
	TRANS-GOOGLE	50.4	44.2	45.4	44.6	46.0	46.0	47.6	42.8	48.4	48.2	43.4	45.4	45.4	47.4	46.1
	TRANS-NLLB	48.6	46.6	47.0	43.2	44.6	43.6	49.0	43.2	44.0	46.0	41.6	48.6	45.6	43.4	45.4
Llama-2-70B-Chat	NATIVE-BASIC	44.0	42.0	45.4	42.6	45.6	38.4	38.4	32.6	35.0	37.6	33.0	41.8	34.8	34.8	39.0
	EN-BASIC	53.6	54.6	57.0	49.6	55.6	46.0	42.8	32.4	50.2	46.2	38.6	52.4	37.6	34.8	46.5
	NATIVE-COT	40.4	42.2	45.4	38.4	41.4	38.4	36.6	32.8	35.2	37.4	32.6	41.0	33.2	36.2	37.9
	EN-COT	53.6	52.8	56.4	50.4	56.8	46.0	40.6	33.4	44.6	47.8	38.2	48.2	37.6	36.6	45.9
	XLT	56.0	59.4	59.6	55.2	61.2	52.6	51.4	36.4	44.4	55.4	44.6	57.8	51.2	45.8	52.2
	TRANS-GOOGLE	58.8	53.4	56.8	56.4	54.8	51.8	55.4	49.6	57.2	56.4	50.2	57.4	50.8	46.6	**54.0**
	TRANS-NLLB	56.4	52.8	54.6	49.8	58.6	50.2	53.4	51.0	52.0	56.0	48.8	52.4	49.0	45.6	52.2

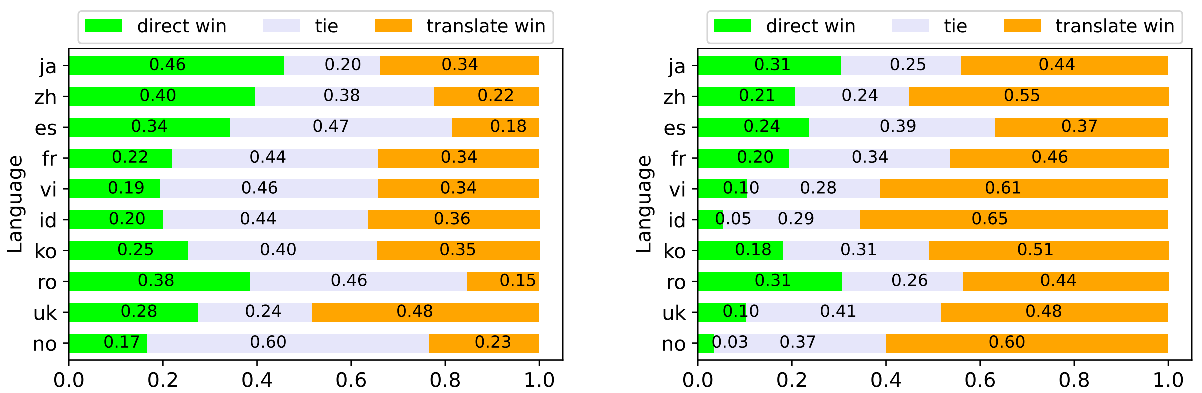
Table 10: Accuracy scores across various languages on the XNLI benchmark.

Model	Prompt type	de	fr	zh	es	ja	ko	avg
ChatGPT	NATIVE-BASIC	62.0	53.6	46.6	46.6	49.0	40.6	49.7
	EN-BASIC	67.6	68.0	58.8	71.4	55.8	50.4	62.0
	NATIVE-COT	61.8	55.0	48.6	48.8	44.0	43.4	50.3
	EN-COT	67.6	64.0	61.2	70.0	55.8	51.2	61.6
	XLT	57.4	63.8	59.8	59.2	58.2	58.2	59.4
	TRANS-GOOGLE	69.0	69.6	66.0	71.4	65.0	62.0	67.2
	TRANS-NLLB	67.0	70.6	68.6	70.2	65.4	63.4	**67.5**
Mistral-7B-Instruct	NATIVE-BASIC	40.6	47.0	49.2	44.2	41.8	44.8	44.6
	EN-BASIC	46.8	47.8	47.8	46.8	45.8	46.6	46.9
	NATIVE-COT	43.8	50.2	38.8	43.6	45.0	44.2	44.3
	EN-COT	46.2	47.4	47.8	47.0	44.8	46.4	46.6
	XLT	47.4	49.6	47.6	46.6	48.2	47.2	47.8
	TRANS-GOOGLE	51.2	49.8	54.0	49.6	52.4	52.0	51.5
	TRANS-NLLB	50.6	52.8	52.4	50.8	54.2	53.6	**52.4**
Llama-2-13b-Chat	NATIVE-BASIC	50.8	57.2	54.0	58.0	55.8	48.2	54.0
	EN-BASIC	60.2	61.0	58.6	59.8	58.2	56.8	**59.1**
	NATIVE-COT	50.4	58.8	59.0	55.8	56.8	47.8	54.8
	EN-COT	59.2	55.8	58.6	59.2	56.4	56.6	57.6
	XLT	54.8	58.0	53.6	56.6	56.8	54.2	55.7
	TRANS-GOOGLE	56.6	62.0	59.6	61.6	56.2	54.6	58.4
	TRANS-NLLB	56.2	60.0	57.4	59.4	57.6	56.2	57.8
Llama-2-70B-Chat	NATIVE-BASIC	53.4	49.8	55.6	61.0	46.8	50.4	52.8
	EN-BASIC	62.8	66.2	58.4	67.0	55.2	52.8	60.4
	NATIVE-COT	53.0	53.4	53.6	65.4	54.6	49.6	54.9
	EN-COT	65.0	70.8	65.0	70.2	61.6	57.0	64.9
	XLT	57.0	61.6	57.6	57.2	49.4	51.6	55.7
	TRANS-GOOGLE	70.6	70.6	68.0	72.2	65.6	64.2	68.5
	TRANS-NLLB	69.8	73.4	69.4	71.2	68.8	62.4	**69.2**

Table 11: Accuracy scores across various languages on the PAWS-X benchmark.

Model	Prompt type	de	ru	fr	zh	es	ja	vi	tr	th	avg
ChatGPT	NATIVE-BASIC	44.1	30.5	46.4	31.4	40.2	20.2	33.0	39.2	33.6	35.4
	EN-BASIC	36.9	30.5	43.3	28.9	44.1	43.5	34.7	32.7	34.0	36.5
	NATIVE-COT	43.6	22.2	46.1	30.0	38.3	33.9	34.1	38.3	31.9	35.4
	EN-COT	44.6	37.4	49.7	38.5	48.0	52.4	32.6	42.0	40.5	42.9
	XLT	36.6	31.0	39.3	31.8	44.0	43.6	37.3	37.9	37.2	37.6
	ransg	42.0	39.2	42.7	48.6	40.8	46.4	37.8	44.2	52.3	**43.8**
	TRANS-NLLB	39.2	34.6	26.7	31.6	29.1	45.3	41.2	41.2	45.9	37.2
bloomz-7b1	NATIVE-BASIC	0.6	3.0	7.6	12.1	11.2	7.5	7.6	0.0	20.9	7.8
	EN-BASIC	7.5	3.7	12.3	21.4	12.2	12.3	13.3	2.1	11.0	10.6
	NATIVE-COT	0.2	0.9	5.9	8.6	8.3	6.0	6.7	0.0	22.2	6.5
	EN-COT	4.0	3.0	11.4	17.9	13.9	8.7	11.1	1.7	12.2	9.3
	XLT	5.7	2.8	10.2	14.8	10.1	7.1	9.6	1.4	10.4	8.0
	TRANS-GOOGLE	13.5	11.5	10.7	25.7	12.5	22.5	12.8	11.7	34.2	**17.2**
	TRANS-NLLB	11.7	8.7	7.2	15.2	9.3	24.5	13.1	11.2	31.3	14.7
Mistral-7B-Instruct	NATIVE-BASIC	8.5	5.2	8.7	7.2	9.5	7.4	8.0	2.6	13.5	7.8
	EN-BASIC	7.9	5.0	7.5	5.1	8.7	6.7	6.3	5.3	10.6	7.0
	NATIVE-COT	9.1	5.4	7.7	8.1	8.2	7.9	7.3	2.8	13.6	7.8
	EN-COT	11.2	7.8	16.0	8.4	14.9	13.1	7.9	7.6	16.4	11.5
	XLT	9.7	7.2	10.4	8.4	10.4	10.5	9.2	6.6	14.2	9.6
	TRANS-GOOGLE	14.6	13.8	14.9	17.7	17.0	22.5	13.4	15.1	24.4	**17.0**
	TRANS-NLLB	13.3	12.7	10.5	14.9	11.8	24.1	13.8	13.5	25.2	15.5
Llama-2-13b-Chat	NATIVE-BASIC	15.0	13.6	31.3	20.6	29.7	13.8	21.2	5.8	13.4	18.3
	EN-BASIC	28.5	11.6	28.7	13.9	27.2	21.0	15.3	15.6	20.0	20.2
	NATIVE-COT	14.6	10.4	29.1	13.3	23.6	10.5	23.8	5.6	10.1	15.7
	EN-COT	28.2	12.6	31.1	11.9	28.9	15.3	15.4	18.3	16.3	19.8
	XLT	23.6	17.0	27.5	10.3	26.2	18.2	14.7	16.4	17.2	19.0
	TRANS-GOOGLE	31.1	29.9	34.6	35.1	31.7	35.4	30.8	31.7	43.9	**33.8**
	TRANS-NLLB	26.1	26.6	19.8	27.4	18.5	36.2	32.1	29.2	40.2	28.4
Llama-2-70B-Chat	NATIVE-BASIC	36.7	23.8	35.2	15.9	39.3	24.7	26.7	8.6	12.1	24.8
	EN-BASIC	33.2	18.1	32.9	18.8	33.7	26.6	16.3	20.7	22.3	24.7
	NATIVE-COT	34.8	19.5	33.9	13.1	38.5	13.1	24.1	9.2	10.6	21.9
	EN-COT	39.5	24.6	39.0	24.2	41.0	35.2	25.3	26.4	25.0	31.1
	XLT	29.8	22.4	29.6	18.0	31.3	29.5	25.0	27.3	26.1	26.6
	TRANS-GOOGLE	37.3	34.0	37.1	43.5	35.4	48.0	35.8	38.3	47.9	**39.7**
	TRANS-NLLB	16.7	16.4	11.9	18.5	14.9	26.8	19.7	21.8	27.6	19.4

Table 12: F1 scores across various languages on the MKQA benchmark.



(a) Win rate with ChatGPT (w/ cultural knowledge)

(b) Win rate with Llama-2-70B-Chat (w/ cultural knowledge)

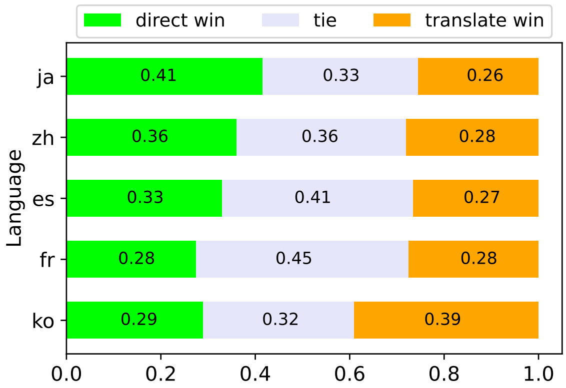
Figure 10: Win rate comparison for each language using ChatGPT and Llama-2-70B-Chat for the subsets of shareGPT with cultural knowledge.

Model	Prompt type	fr	zh	es	vi	tr	avg
ChatGPT	NATIVE-BASIC	29.2	39.3	26.9	34.4	29.9	31.9
	EN-BASIC	28.9	38.8	27.8	37.9	30.0	**32.7**
	NATIVE-COT	28.8	38.5	26.1	34.0	27.9	31.1
	EN-COT	25.4	35.1	26.0	33.5	28.6	29.7
	XLT	24.2	25.5	18.1	23.4	26.1	23.4
	TRANS-GOOGLE	27.2	36.2	26.3	32.6	28.9	30.3
	TRANS-NLLB	26.4	29.7	26.1	31.5	27.7	28.3
bloomz-7b1	NATIVE-BASIC	14.6	24.3	20.0	7.7	8.2	14.9
	EN-BASIC	20.1	23.9	20.9	20.6	14.2	**19.9**
	NATIVE-COT	18.2	25.4	24.1	1.7	8.0	15.5
	EN-COT	18.0	26.1	21.6	19.3	11.3	19.3
	XLT	12.2	19.9	19.3	14.5	5.3	14.2
	TRANS-GOOGLE	10.0	14.2	12.1	9.0	10.7	11.2
	TRANS-NLLB	10.5	8.6	12.5	9.7	11.5	10.6
Mistral-7B-Instruct	NATIVE-BASIC	23.0	34.0	22.3	25.8	24.4	25.9
	EN-BASIC	20.9	16.5	21.5	28.0	21.1	21.6
	NATIVE-COT	19.7	33.6	22.1	26.4	21.1	24.6
	EN-COT	20.6	12.1	19.9	22.2	18.8	18.7
	XLT	15.4	16.5	14.7	21.7	19.6	17.6
	TRANS-GOOGLE	26.8	34.9	26.4	39.5	36.5	**32.8**
	TRANS-NLLB	26.8	30.0	26.6	40.6	37.3	32.2
Llama-2-13b-Chat	NATIVE-BASIC	27.7	21.9	25.3	38.8	23.8	27.5
	EN-BASIC	25.7	38.2	23.6	37.7	30.2	31.1
	NATIVE-COT	27.9	29.0	24.8	31.8	22.7	27.2
	EN-COT	24.0	39.4	23.1	36.4	28.0	30.2
	XLT	24.2	17.7	22.4	23.6	18.1	21.2
	TRANS-GOOGLE	28.0	42.9	27.9	41.6	38.0	**35.7**
	TRANS-NLLB	27.5	37.5	26.9	41.6	37.3	34.2
Llama-2-70B-Chat	NATIVE-BASIC	28.8	34.5	27.3	29.7	26.8	29.4
	EN-BASIC	29.0	31.8	24.3	35.7	35.3	31.2
	NATIVE-COT	25.3	29.5	26.7	22.4	25.2	25.8
	EN-COT	27.0	35.2	22.1	34.8	32.0	30.2
	XLT	18.1	29.7	15.2	14.2	11.5	17.7
	TRANS-GOOGLE	26.8	39.7	27.1	38.7	36.7	**33.8**
	TRANS-NLLB	26.6	37.5	26.3	39.0	36.2	33.1

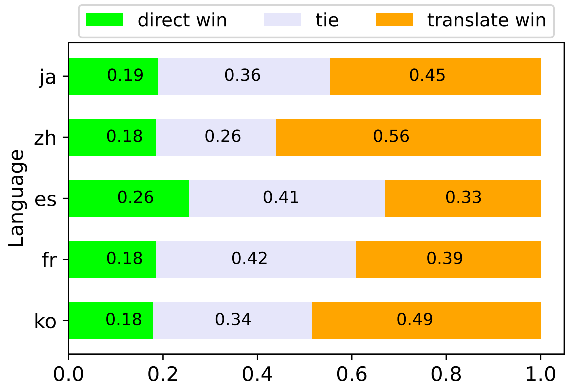
Table 13: ROUGE-1 scores across various languages on the XL-sum benchmark.

Language	**ja**	**zh**	**es**	**fr**	**vi**	**id**	**ko**	**ro**	**uk**	**no**
Ratio (%)	59	58	38	41	67	55	55	74	30	57

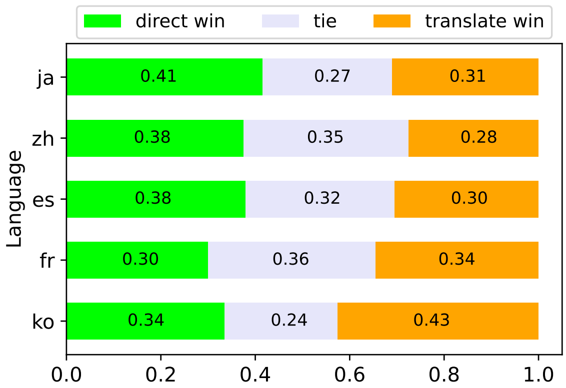
Table 14: The percentage of the questions that necessitate local cultural knowledge.



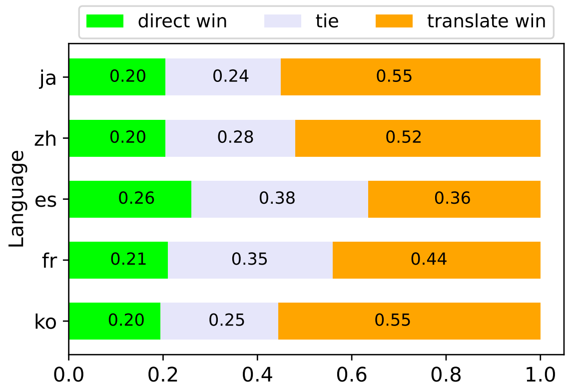
(a) ChatGPT judged by GPT-4o



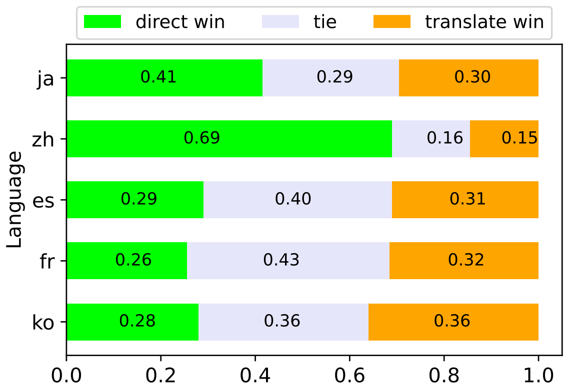
(b) Llama-2-70B-Chat judged by GPT-4o



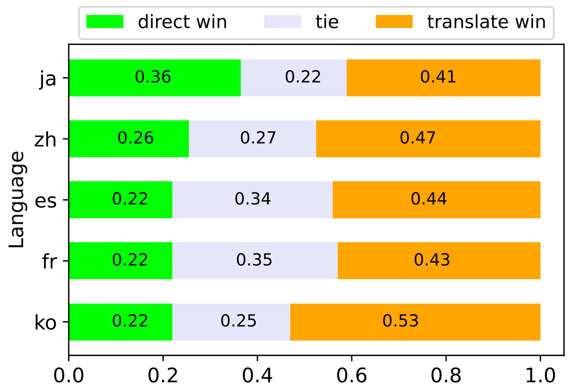
(c) ChatGPT judged by Claude-3.5-Sonnet



(d) Llama-2-70B-Chat judged by Claude-3.5-Sonnet

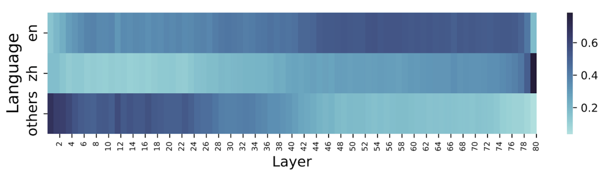


(e) ChatGPT judged by Gemini-1.5-Pro

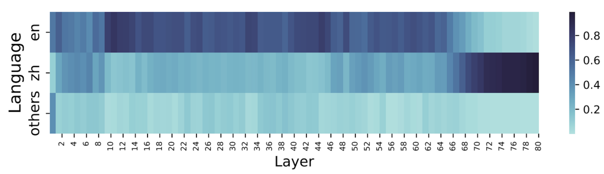


(f) Llama-2-70B-Chat judged by Gemini-1.5-Pro

Figure 11: Win rate comparison for five languages using ChatGPT and Llama-2-70B-Chat judged with three advanced LLMs.



(a) Llama-2-70B-Chat



(b) Qwen1.5-72B-Chat

Figure 12: Layerwise language distribution for (a) Llama-2-70b-Chat and (b) Qwen1.5-72B-Chat with Chinese prompts.