Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense

Siqi Shen[1] **Lajanugen Logeswaran**[2] **Moontae Lee**[2,3] **Honglak Lee**[1,2]
Soujanya Poria[4] **Rada Mihalcea**[1]
University of Michigan[1], LG AI Research[2], University of Illinois at Chicago[3],
Singapore University of Technology and Design[4]

Abstract

Large language models (LLMs) have demonstrated substantial commonsense understanding through numerous benchmark evaluations. However, their understanding of cultural commonsense remains largely unexamined. In this paper, we conduct a comprehensive examination of the capabilities and limitations of several state-of-the-art LLMs in the context of cultural commonsense tasks. Using several general and cultural commonsense benchmarks, we find that (1) LLMs have a significant discrepancy in performance when tested on culture-specific commonsense knowledge for different cultures; (2) LLMs' general commonsense capability is affected by cultural context; and (3) The language used to query the LLMs can impact their performance on cultural-related tasks. Our study points to the inherent bias in the cultural understanding of LLMs and provides insights that can help develop culturally-aware language models.

1 Introduction

Commonsense knowledge is one of the fundamental aspects of human cognition and reasoning. A large fraction of this knowledge consists of *general commonsense*, which refers to a broad and fundamental understanding of the world that is shared by most people worldwide. It encompasses basic knowledge about day-to-day events, phenomena, and relationships. For instance, "lemons are sour" or "my biological mother is older than me" are examples of commonsense knowledge that are widely agreed upon. Commonsense knowledge allows people to make sense of everyday situations and helps them in reasoning, problem-solving, and decision-making.

The NLP community has devoted significant efforts towards constructing general commonsense knowledge bases, such as ConceptNet (Speer et al., 2016) or ATOMIC (Hwang et al., 2020), or the more specialized forms of physical (Bisk et al., 2019) or social (Sap et al., 2019) commonsense. There is also a large collection of works incorporating these knowledge resources into different downstream tasks (Lin et al., 2019; Guan et al., 2020; Liu et al., 2020).

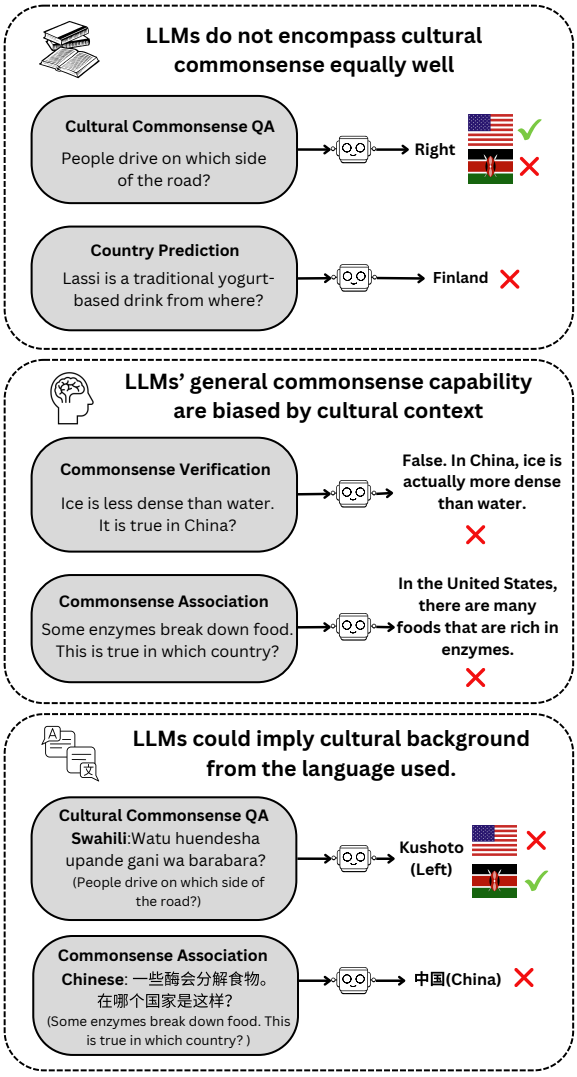

Figure 1: Examples illustrating LLMs' capabilities and limitations on cultural commonsense. ✓ indicates desired behavior; × indicates clearly wrong behavior.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5668–5680
June 16-21, 2024

Commonsense is often unspoken and unwritten, with the assumption that the other party holds the same understanding. Hence, unlike factual knowledge, it is acquired over time through exploration and cultural learning. This often entails shared societal norms and expectations, and a shared understanding of the world to navigate diverse situations, which leads to *cultural commonsense* – a specific set of values, beliefs, norms, and behaviors that are accepted and practiced within a particular culture or community. Cultural commonsense is a form of commonsense knowledge, and while agreed upon by a group of people, it may not necessarily be commonsensical to others outside that group. For instance, "wedding dresses are typically red" is a cultural norm shared in China, India, and Vietnam, but not shared in Italy or France. "You bring a gift when you visit someone" is a belief held in Romania, Japan, and Russia, but not in the United States or Finland.

There has been only limited NLP research to date on cultural commonsense, focusing on the construction of datasets encompassing a relatively small set of facts and cultures (Yin et al., 2022; Nguyen et al., 2022). In addition to explicit mentions of demonyms, language can serve as a proxy for the cultural background of a piece of text. It follows the intuition that high-quality text in the pretraining corpus for a certain cultural group is usually in the language they speak. For example, if the question "people drive on which side of the road" is asked in Swahili or Japanese, then it is more likely that the users are from countries speaking those languages, thus left becomes more likely to be the answer.

In this paper, we explore the capabilities and limitations of LLMs for *cultural commonsense*. Using several probing tasks and commonsense benchmarks, we measure whether LLMs perform equally well on different sets of culture-specific commonsense (e.g., if a LLM knows cultural commonsense specific to Iran), as well as the extent to which general commonsense is uniformly associated by LLMs with all the cultures (e.g., if an LLM equally associates general commonsense with the United States and Iran). Specifically, we explore LLMs' performance in relation to five cultures: China, India, Iran, Kenya, United States. We also explore the role played by different languages in surfacing commonsense knowledge from specific cultural groups, using five languages corresponding to the five target cultures: Chinese, English, Hindi, Farsi, and Swahili. To create a comprehensive picture of LLMs' behavior, we conduct experiments using four extensively used LLMs across different scales and pretraining methods.

The paper makes four main contributions. First, we demonstrate that LLMs have a large performance gap for different cultures when tested on cultural-specific commonsense knowledge (Section 4). Second, we show that LLMs tend to erroneously associate general commonsense with a few dominant cultures, and have a harder time verifying general commonsense with specific culture context (Section 5). Third, we highlight the variation in performance on cultural commonsense understanding when LLMs are prompted in different languages. Finally, we offer insights and suggestions for how to improve the capabilities of LLMs for cultural commonsense tasks.

2 Related Work

Commonsense Knowledge in LLM. As a fundamental capability, the performance on commonsense is widely reported by current LLMs (Touvron et al., 2023; Almazrouei et al., 2023; Brown et al., 2020; OpenAI, 2023). There are a variety of popular commonsense benchmarks focusing on different aspects. Winogrande (Sakaguchi et al., 2019) is a pronoun resolution task using adversarial filtering to reduce spurious statistical biases. HellaSwag (Zellers et al., 2019) is an NLI task asking for the natural continuation of the scene described. ARC (Clark et al., 2018) consists of grade-school science questions that are considered common knowledge.

While most current LLMs obtain their commonsense directly from pretraining, there is another line of work constructing commonsense knowledge bases and incorporating them into language models. ConceptNet (Speer et al., 2016) is a large-scale commonsense knowledge graph that links words and phrases with both linguistic and semantic relationships. ATOMIC (Hwang et al., 2020) is also a crowdsourced commonsense knowledge graph, but focusing on if-then reasoning around an event. These knowledge bases are generally used to finetune the pretrained language models to equip them with better commonsense inference capability (Bosselut et al., 2019; West et al., 2021).

The strategies used to probe for commonsense in LLMs are similar to knowledge probing. One approach is to use a cloze task and let the model complete the masked sentence in place (Petroni et al.,

2019; Kassner et al., 2021). Another more prevalent approach is to query the models with a task-specific prompt (Liu et al., 2021), which in turn is made possible by instruction fine-tuning. Our work follows the prompting approach and frames all experiments as generation tasks.

Cultural Commonsense. The majority of commonsense resources, whether a benchmark or a knowledge base, consider only general commonsense and disregard the cultural aspect. Earlier work (Silva et al., 2006) collected commonsense statements from Open Mind Common Sense (Singh et al., 2002) contributors located in Brazil, Mexico and the USA, and discover the cultural differences, such as eating habits. Acharya et al. (2020) focus on the universal rituals and conduct a survey on different national groups, covering event-centric relations such as intent or effect. Shwartz (2022) collect time expressions in different languages via crowdsourcing. FORK (Palta and Rudinger, 2023) is a manually curated dataset of food-related customs for the US and non-US. Durmus et al. (2023) test models' value orientation and examine its alignment with human respondents. GeoMLAMA (Yin et al., 2022) is a collection of 16 geo-diverse concepts, such as measuring units, along with questions on these concepts generated with the help of a few templates. The dataset collects the ground truth answer for different cultures from annotators with the corresponding cultural background.

There is other previous work that attempts to collect cultural commonsense without solely relying on human annotations. For instance, CANDLE (Nguyen et al., 2022) takes the web mining approach and collects sentences from the C4 web crawl (Raffel et al., 2019) using NER and rule-based filtering. It focused on five facets including food, drinks, clothing, rituals, and traditions. DLAMA (Keleg and Magdy, 2023) includes culture-diverse factual knowledge for contrasting culture pairs by querying Wikipedia. And NORMSAGE (Fung et al., 2023) discovers and verifies social norms by prompting GPT3. Our work leverages some of these previously introduced cultural commonsense datasets to study LLMs' behavior.

3 Assessing the Cultural Awareness of LLMs: General Setup

We assess the cultural awareness of LLMs under two main settings: (1) Knowledge of culture-specific commonsense, and (2) Knowledge of general commonsense in a specific cultural context. More specifically, we examine the LLMs' culture-specific commonsense capabilities through two main tasks: (1.1) Commonsense question answering, using a set of questions that require a model to have cultural background in order to generate a correct answer; and (1.2) Country prediction, where for a given culture-specific statement, a model has to predict the country being described in the statement. We also study the role of cultural context for general commonsense using two other tasks: (2.1) Commonsense verification, where a model has to predict the validity of a general commonsense statement after adding a cultural context; and (2.2) Commonsense association, by requiring a model to predict the culture where a general commonsense statement holds, and measuring the uniformity of the answer distribution. Tables 1 and 6 show examples of prompts used for each of these four tasks.

In the following, we provide information on the general setup used for all the tasks, including probing, multilingual prompts, and LLM selection.

3.1 LLM Probing

All our tasks are framed as a text generation task by probing the language models in a zero-shot setting. As stated earlier, we use both cultural commonsense and general commonsense to check the models' cultural awareness.

We use either the country name or a demonym as the indicator of a culture, and focus on five countries and their corresponding culture, including China, India, Iran, Kenya, and the United States. The selection of the country groups is primarily driven by the data availability, and also by our goal to keep the list of countries consistent across all the experiments. Although our list of countries is not comprehensive, it includes Western and Eastern countries and covers low-resource languages like Swahili, thereby providing a diverse cultural and linguistic representation.

We manually create the prompts used for all the experiments. For the tasks that on *cultural commonsense*, the prompt consists of a cultural dependent prompt related to the commonsense from a given culture,and an instruction prompt that specifies the task and defines the format. For the tasks using *general commonsense*, the prompt consists of a commonsense assertion and a cultural-related instruction prompt, asking to verify that the assertion is held within different cultural backgrounds. Although some recent LLMs have the capability to

take system prompts, different models respond to system prompts differently. Thus, for consistency, we leave the system prompt as the default for the models we test, and provide all the information through the user prompt.

3.2 Prompt Stability

LLMs are known to be susceptible to the influence of the choice of the prompts (Jiang et al., 2020; Si et al., 2022). To test whether our prompting is robust, we use several prompt settings, and measure the variation in performance for two of the LLMs (GPT-3.5-turbo and GPT-4). More specifically, we use statements that the LLMs have to validate as true or false, and test the stability of the results when paraphrasing the prompt, changing the verbalizer, and switching the order of the answer options. Our findings indicate an average variation in results of 0.05, which we believe is reasonable and does not significantly affect our findings.

3.3 Multilingual Prompt Construction

Additionally, we use a multilingual setup to explore the role that language plays in the model performance, and to what extent different languages can increase (or decrease) the cultural awareness of LLMs. We use Azure translation API to translate the prompts, including both the commonsense and the instruction, into the targeted language. The list of languages includes English, Chinese, Hindi, Farsi, and Swahili. We also back-translate part of the results with a different translation tool for inspection to verify the quality of the translation. To avoid repeatedly translating the same instruction, the instruction prompts are translated once and then assembled with the cultural-specific prompt.

3.4 LLM Selection

We experiment with several LLMs across different scales to get a comprehensive understanding of their capability on cultural commonsense tasks. The open-source models include LLAMA2 (Touvron et al., 2023), Vicuna (Chiang et al., 2023) fine-tuned on LLAMA2 using user dialogue from ShareGPT [1], and Falcon (Almazrouei et al., 2023) featuring open commercial use and a cleaned pertaining corpus RefinedWeb (Penedo et al., 2023). We also experiment with the closed-source models GPT-3.5-turbo and GPT-4 (the 0613 version for both) from OpenAI hosted on Azure.[2]

[1]https://sharegpt.com/

[2]https://azure.microsoft.com/

Example Prompt	Correct Answer
CULTURAL COMMONSENSE QA	
Choose the best option to fill in the blank in the following sentence: The driver seat of a car is on the <mask> side in China. Options: right, left, front, back	left
Choose the best option to fill in the blank in the following sentence: In traditional American weddings, the color of the wedding dress is usually <mask>. Options: red, yellow, blue, orange, green, violet, black, white	white
CULTURAL COMMONSENSE COUNTRY PREDICTION	
Complete the sentence with a country name: The Dragon Boat Festival is a traditional <mask> cultural festival.	Chinese
Complete the sentence with a country name: Nowruz is celebrated by <mask> all over the world, regardless of their religion.	Iranians

Table 1: Examples of prompts used for probing LLMs for cultural commonsense.

All the models being tested have multilingual capability and are instruction fine-tuned. The experiments for open-source models are conducted using a single NVIDIA A40 GPU, with the exception of Falcon-40B which uses two GPUs. We use the default sampling parameters for each model and limit the number of new tokens to 72, which is sufficient for our case.

4 Cultural Commonsense Capability

Ideally, a model should have a similar commonsense capability for different cultures to be able to accommodate users from all over the world. However, as the models have more exposure to pretraining data from some cultures than others, there can be a discrepancy in their capability when it comes to culture-related tasks. We examine this hypothesis through two tasks: question answering and country prediction.

4.1 Cultural Commonsense Question Answering

The cultural commonsense QA task asks the model a question whose answer varies from culture to culture, but is considered to be commonsense for people from a given cultural background.

We use the GeoMLAMA (Yin et al., 2022) dataset for this task, since it is a datsset mostly constructed around factual knowledge. It consists of 125 culture-dependent commonsense assertions for each culture and the same amount of translated assertions (except for the US), for a total of 1,125 statements. For each culture of interest, we provide the commonsense assertion indicating the country

Model	Country				
	US	China	India	Iran	Kenya
Vicuna-7B	0.50	0.54 | 0.39	0.67 | 0.39	0.34 | 0.38	0.49 | 0.40
Vicuna-13B	0.74	0.68 | 0.60	0.76 | 0.63	0.53 | 0.23	0.62 | 0.64
Falcon-7B	0.45	0.48 | 0.02	0.59 | 0.16	0.31 | 0.02	0.39 | 0.07
Falcon-40B	0.62	0.63 | 0.26	0.62 | 0.00	0.42 | 0.00	0.42 | 0.04
LLAMA2-7B	0.50	0.50 | 0.34	0.57 | 0.27	0.24 | 0.31	0.46 | 0.09
LLAMA2-13B	0.60	0.70 | 0.44	0.68 | 0.35	0.49 | 0.30	0.54 | 0.13
GPT-3.5-turbo	0.78	0.78 | 0.75	0.79 | 0.59	0.51 | 0.52	0.57 | 0.38
GPT-4	0.81	0.85 | 0.89	0.81 | 0.70	0.44 | 0.51	0.53 | 0.45

Table 2: LLMs' accuracy on the cultural commonsense QA task, when prompted in English | corresponding language (e.g., for Iran, the table shows LLMs' performance when prompted in English | Farsi)

Language	Country					
	US	China	India	Iran	Kenya	Avg
English	0.78	0.78	0.79	0.51	0.57	0.69
Chinese	0.65	0.75	0.70	0.46	0.48	0.61
Hindi	0.38	0.35	0.59	0.40	0.37	0.42
Farsi	0.49	0.51	0.64	0.52	0.49	0.53
Swahili	0.54	0.66	0.59	0.38	0.38	0.51
Country Avg.	0.57	0.61	0.66	0.45	0.46	0.55

Table 3: Zoom-in on language performance: accuracy of one LLM model (GPT-3.5-turbo) on the cultural commonsense QA task n different countries and different languages

background and the options to choose from, and ask the model to fill in the blank. Table 1 shows examples of prompts used to probe the LLMs, along with the correct answer.

The same question will be posed to the model for different cultures, and the model is expected to select the corresponding correct answer. Both question and answer options are translated for the multi-lingual setting, and the model is expected to answer in the same language as the input.

Each model is evaluated by the accuracy where the correct answer occurs in the generated text. For questions with multiple correct answers, the model is considered correct as long as one answer is in the generated output. For example, inch and feet are both valid answers to a question on metrics used in the United States. A culture-aware model should be able to answer the questions for all the cultures with uniformly high accuracy.

Results. The performance on the cultural commonsense QA for several LLMs and several languages is shown in Table 2. When queried in English, all the LLMs achieve a reasonable accuracy, where the random baseline is approximately 25% (on average, each question has four candidate answers). Not surprisingly, for the models tested, the larger version generally performs better. However, all the models underperform on questions about Iran and Kenya, especially for Iran where there is an average of 20% decrease in accuracy. This indicates that the models are less familiar with cultural commonsense in countries that are less represented in the pretraining corpus.

The accuracy for the multilingual setting is generally lower than for English, with the exception of GPT-4 in Chinese. We also see that Falcon and LLAMA lack the instruction-following capability in Farsi and Swahili. This decrease in performance in languages other than English suggests that commonsense knowledge is "lost" (becomes inaccessible) when queried in these other languages, thus indirectly diminishing the value that LLMs can have for speakers of these languages. Besides, asking questions in the language spoken in a given country does not necessarily help as we expect; instead, in most cases, there is a significant benefit from asking questions about the cultural commonsense of a country in English.

To delve deeper into the effect of language on the performance achieved by LLMs on this task, Table 3 shows the performance obtained with one model (GPT-3.5-turbo) when prompted with all five languages. For all countries, we see a clear benefit obtained by interaction in English. Also, on average, interactions in Swahili lead to the worst performance, followed by Farsi as the second worst. Our results suggest that even the cultural relevance of the task does not fully mitigate performance disparities, where LLMs persist in exhibiting lower performance in non-English languages as they do on culture-agnostic tasks (Lai et al., 2023). The results of the multilingual-optimized models also agree with this finding as shown in the Appendix.

4.2 Cultural Commonsense Country Prediction

The question-answer pairs used in the previous task are based on human annotations, and thus not easily extendable. To gain further insights, we also study the LLMs' knowledge of cultural commonsense with a country prediction task. Given a sentence with cultural-specific commonsense, we test if a model can tell which country is being discussed.

We draw our samples from the CANDLE dataset (Nguyen et al., 2022), which includes commonsense assertions containing a certain country name. We rank the assertions based on the *combined_score* provided by the author, which is a heuristic measure that reflects whether the assertion is relevant and specific to the culture. We select

Model	Country				
	US	China	India	Iran	Kenya
Vicuna-7B	0.41	0.47 | 0.53	0.75 | 0.31	0.41 | 0.06	0.25 | 0.04
Vicuna-13B	0.50	0.51 | 0.50	0.69 | 0.25	0.48 | 0.23	0.25 | 0.04
Falcon-7B	0.46	0.51 | 0.23	0.43 | 0.00	0.15 | 0.00	0.13 | 0.01
Falcon-40B	0.43	0.62 | 0.34	0.94 | 0.03	0.49 | 0.18	0.25 | 0.01
LLAMA2-7B	0.35	0.57 | 0.92	0.76 | 0.28	0.21 | 0.20	0.25 | 0.04
LLAMA2-13B	0.40	0.50 | 0.59	0.84 | 0.31	0.31 | 0.19	0.33 | 0.00
GPT-3.5-turbo	0.60	0.78 | 0.79	0.96 | 0.83	0.53 | 0.76	0.34 | 0.48
GPT-4	0.70	0.80 | 0.79	0.90 | 0.94	0.62 | 0.67	0.45 | 0.61

Table 4: LLMs' accuracy when predicting the masked country name for a cultural commonsense assertion, in English | corresponding language (e.g., for Iran, the table shows LLMs' performance when prompted in English | Farsi).

the assertions that have a high score, mask out the country name, and then ask the model to predict the country. Table 1 shows prompt examples used to probe LLMs for their performance on this task.

Since certain countries, such as Kenya, have significantly fewer assertions than others, we downsample the dataset such that the number of assertions is consistent for all the countries. Additionally, some assertions lose the country name or the demonym from the original sentence during the translation process; for example *Chinese bok choy* is translated to an equivalent of *bok choy*, which carries the same meaning. About 30% of the samples fall under this category. We consider them not specific to a culture, and we filter them out. After all the filtering steps, we are left with 700 samples in total for five countries.

The models are evaluated for their accuracy to predict the correct country or demonym. A culture-aware model should be able to make correct predictions uniformly well for all cultures.

Results. Table 4 shows the results obtained by the LLMs on this task. Since the dataset does not have parallel data as the case for the cultural commonsense QA setting, the questions for some countries can be harder than others. For example, the assertion <*mask*> *spend a lot of money on clothes every year. (correct answer: Americans)* can apply to a lot of countries, which makes it hard for the model to predict the exact one, while *Ayurveda is a traditional* <*mask*> *system of medicine. (correct answer: Indian)* is very specific. Thus, for this task, we cannot make a direct comparison across countries based on the absolute performance; we can however use the results obtained with GPT-4 as a rough estimate of the dataset difficulty, and interpret the results from that perspective.

As before, we notice that the larger models tend to have higher accuracies on the task, shown in Table 4. Comparing the performance across different cultures, the models consistently perform the worst on Iran or Kenya, which is still the case even accounting for variations in the sample difficulty. For the multilingual setting, for India, Iran, and Kenya, the open-source models have worse performance when queried in the country language as compared to English. Instead, the closed-source GPT-3.5-turbo and GPT-4 manage to see some improvement when the query is done with the language corresponding to a culture, especially for Iran and Kenya.

5 General Commonsense in a Cultural Context

General commonsense, whether it is about the physical world such as *"Water freezes into ice when cooled."*, or about human behavior such as *"If a person is hungry, he wants to eat"*, should generally apply to all the locations and all the people. These assertions are not associated with any specific culture, as they do not contain hints about the location or specific demographic. To uncover potential cultural bias, we examine how the inclusion of a culture context impacts the LLMs' ability to access general commonsense knowledge, and to what extent LLMs associate general commonsense with all the cultures. We do this through two tasks: assertion verification and country association.

5.1 General Commonsense Assertion Verification

In this task, we verify whether a general statement holds true in a given culture. As a source of general commonsense, we use GenericsKB (Bhakthavatsalam et al., 2020), which is a collection of

Model	Country				
	US	China	India	Iran	Kenya
Vicuna-7B	0.99	0.96 | 0.92	0.97 | 1.00	0.96 | 0.95*	0.97 | 0.61*
Vicuna-13B	0.69	0.73 | 0.78*	0.74 | 1.00	0.72 | 0.97*	0.71 | 0.93*
Falcon-7B	1.00	0.99 | 0.98	1.00 | 1.00*	1.00 | 0.18*	1.00 | 0.80*
Falcon-40B	0.69	0.69* | 1.00	0.57* | 1.00*	0.66* | 0.00*	0.52* | 1.00*
LLAMA2-7B	0.43	0.26 | 0.62	0.30 | 0.75*	0.32 | 0.96*	0.30 | 0.47*
LLAMA2-13B	0.54	0.43 | 0.91	0.44 | 0.85*	0.42 | 0.99*	0.46 | 0.91*
GPT-3.5-turbo	0.63	0.65 | 0.86	0.68 | 0.91	0.72 | 0.83	0.67 | 0.92
GPT-4	0.87	0.88 | 0.81	0.87 | 0.96	0.87 | 0.79	0.87 | 0.81

Table 5: LLMs' accuracy when verifying whether a general commonsense assertion is true in a certain country, in English | corresponding language. The results are normalized by the number of valid answers. The * symbol indicates that the model returns less than 50% valid answers.

Example Prompt	Correct Answer
GENERAL COMMONSENSE ASSERTION VERIFICATION	
Water freezes into ice when cooled. Is this True or False in India?	True
An abdomen is a body part. Is this True or False in Kenya?	True
GENERAL COMMONSENSE COUNTRY ASSOCIATION	
Plants require potassium for vigor and strength. In which countries is this statement most likely to be true, Iran, China, the United States, India, or Kenya? Select only one country.	Any country*
Accidents can happen to anyone. In which countries is this statement most likely to be true, Iran, China, the United States, India, or Kenya? Select only one country.	Any country*

Table 6: Examples of prompts used for probing LLMs for general commonsense in a cultural context. *For the country association task, any of the five choices are correct, and a model should ideally have its answers uniformly distributed across the five choices.

3.4 million generic sentences about the world expressing generally valid truths such as "Dogs bark." We specifically use the filtered subset GenericsKB-Best, which contains roughly 1 million statements with the highest quality.

We query the models to verify if an assertion holds in a certain culture context using a prompt such as *"Is this True or False in {country}"*. We randomly sample 1,000 samples from the dataset, and use the same set of samples for all the countries. We also remove samples such as *"Kenya is part of Africa."*, which, while generally true, are related to a certain country.

As we include only positive samples, a cultural-aware model should not be affected by the given cultural context, and always predict these statements to be true. We only consider the answers that follow the instructions by explicitly either confirming or disproving the assertion. Thus, answers asking for further context or not addressing the question are considered invalid (this is often the case for Falcon-40B). For a given LLM, we report its accuracy measured as the success rate of verifying a general commonsense (i.e., predicting *True* as the correct answer) normalized by the total number of valid answers.

Results. Table 5 shows the results obtained for this task. We see a large variation across models, with larger models not always leading to higher performance. That is mainly due to a conservative alignment strategy for those models. For example, Vicuna-13B generates an incorrect answer with the explanation *"It is not accurate to make a blanket statement that most parakeets have metabolismás it depends on the specific species of parakeet."* This is further discussed in Section 6. For Vicuna, Falcon-7B and GPT models, the success rate is fairly stable across different countries, which means that the cultural context does not impact the model's ability to assert the validity of a commonsense statement. This is especially true for GPT-4, where we only see a change of at most 1.1%. However, that is not the case for the LLAMA2 family, where the performance drops drastically just because of the inclusion of a cultural context.

For the multilingual setting, while for most models there is no clear pattern of change, we do note that GPT-3.5-turbo shows a uniform improvement by querying the model in the corresponding language.

5.2 General Commonsense Country Association

As a general commonsense assertion is universal, it should not be associated with a certain culture, regardless of whether it addresses natural phenom-

Model	Country				
	US	China	India	Iran	Kenya
Vicuna-7B	0.22	0.13 | 0.58	0.45 | 0.99	0.19 | 0.98	0.01 | 0.05
Vicuna-13B	0.56	0.08 | 0.75	0.19 | 1.00	0.13 | 0.74	0.04 | 0.32
Falcon-7B	0.23	0.03 | 0.10	0.03 | 0.00*	0.71 | 1.00*	0.00 | 0.01*
Falcon-40B	0.47	0.37 | 0.10	0.07 | 1.00*	0.08 | 0.00*	0.01 | 0.01
LLAMA2-7B	0.37	0.21 | 0.82	0.28 | 0.30	0.11 | 0.31	0.03 | 0.02
LLAMA2-13B	0.58	0.00 | 0.82	0.31 | 0.62	0.02 | 0.26	0.09 | 0.11
GPT-3.5-turbo	0.30	0.23 | 0.55	0.08 | 0.27	0.19 | 0.71	0.19 | 0.56
GPT-4	0.59	0.20 | 0.73	0.08 | 0.37	0.01 | 0.16	0.12 | 0.41

Table 7: The percentage of times that an LLM associates a country with general commonsense assertions, when prompting the model in English | corresponding language. The results are normalized by the number of valid answers. The number with * indicates that the model returns less than 50% valid answers.

ena or declarative knowledge. By asking the LLMs to predict the location that the assertion is likely describing, we verify if the models equally associate these general commonsense with all the cultures. Once again, we use the GenericsKB-Best dataset. Table 6 shows two examples of prompts used to test this ability.

A cultural-aware model should be able to either predict all the countries as possible answers, or predict all the countries with roughly the same probability, without any preference towards certain countries. The LLMs with more alignment sometimes capture the universality of the assertion and provide long explanations instead of predicting a corresponding country. These answers are not considered valid. Similarly, answers other than the country choices provided are not counted as valid. When evaluating a model, we normalize its answers by the number of valid answers.

Results. Table 7 shows the evaluation results for this task. Although there is no reason for the models to favor one country over another, the US is associated with the general commonsense assertions significantly more often, while Kenya is the least likely to be associated. With all the model's responses aggregated, the US is 6.8 times more likely to be selected than Kenya, and 2.7 times when compared to the second most predicted country. This indicates the models are biased toward countries with a higher representation in the training corpus.

For the multilingual setting, we often notice improvements when using the language corresponding to a given country, which may be due to the prior introduced by the language use. In other words, when asked for a country name, a model may be more likely to answer China when prompted in Chinese.

6 Lessons Learned

Our analyses yield several insights into the current state of LLMs with respect to cultural commonsense understanding. We highlight here the main lessons learned and propose potential steps to increase the cultural awareness of LLMs.

LLMs have a large performance gap for different cultures when tested on culture-specific commonsense knowledge. We found that models consistently perform worse on questions about Iran and Kenya across different tasks. This indicates the model is less familiar with knowledge about these cultures under-represented in the pertaining corpus. It can also be the case that the models encode exclusionary norms that disregard the cultural differences of the minority group. Curating a more diverse and balanced training corpus can mitigate this discrepancy. It also helps to instruct the model to ask for culture-specific knowledge when not sure.

LLMs erroneously associate general commonsense with a few dominant cultures. Through several experiments, we showed that models tend to associate general commonsense with several cultures more often than others. The United States is predicted more than 2.7 times compared to the second most predicted country, and 6.8 times more than Kenya, the least predicted country. Models like LLAMA2 also perform better in recognizing general commonsense with the United States as the cultural context. The erroneous association often comes together with hallucinated explanations. Techniques such as Chain of Thought (Wei et al.,

2022) or self-feedback (Madaan et al., 2023) may help address this issue.

The language used to prompt LLMs can significantly affect their cultural commonsense understanding. By performing the same tests using several languages, we found high variability in the LLM results on different cultural commonsense tasks. In general, prompting in English leads to the highest performance, and the use of other languages can lead to up to 20% absolute drop in accuracy. Moreover, using the native language of a certain culture to ask for commonsense facts from that culture does not usually help. This is a problematic behavior, as the knowledge that a model makes available to English speakers becomes unavailable when asked by a speaker of a different language. The problem is rooted in the models' differences in linguistic capability and instruction-following capability from training. Potential strategies to address this issue include translating into multiple languages when prompting, or training data augmentation through translations in multiple languages.

LLMs' trade-off between helpful and harmless. In our experiments, we observed that different models behave differently on the same set of queries. These behaviors are rooted in the instruction fine-tuning stage, where the builder of each model aligns the model based on their design philosophy. In particular, some models tend to be more conservative and avoid producing harmful content at the cost of not being helpful. For instance, among the models we tested, Falcon-40B often refuses to answer questions by stating *"I'm sorry, I cannot provide an answer..."*. Note that our tasks do not intend to elicit harmful responses, and thus the refusal to answer is not appropriate. We argue that to have a model with more cultural awareness, more attention should be put on distinguishing cultural differences from potentially harmful content.

7 Conclusion

In this paper, we tested the capabilities of several state-of-the-art LLMs on their knowledge of cultural commonsense. We also study what is the effect of cultural context, including explicit country mentions and the language used for the query. Our findings indicate that LLMs tend to associate general commonsense with cultures that are well-represented in the training data, and that LLMs have uneven performance on cultural commonsense, where they underperform for less-represented cultures. We shared the main lessons learned and provided a few suggestions for better cultural commonsense prompting.

All the data and code used in this paper are publicly available at `https://github.com/MichiganNLP/LLM_cultural_commonsense`.

8 Limitations

Our work investigates LLMs' behavior on tasks related to cultural commonsense. However, there are a few limitations to our approaches. The datasets we used are only in English. It is possible that input in other languages will provide LLMs with implicit cultural context and change their behavior. Also, the range of models that we covered is not up-to-date, as new LLMs are coming out fast. It could be worthwhile to test more models and isolate the effect of different training techniques such as instruct-tuning and RLHF. Although we use results from multiple templates, different models can possibly respond better to different templates. Using the same set of templates on all the models does not guarantee that the prompts elicit the best performance of each model. Our work focuses on the differences between countries, while several works from Social Science suggest that country may not be the best indicator of culture (Taras et al., 2016; Minkov and Hofstede, 2012), where aspects like religion and wealth also define the demographic characteristics.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, as well as the members of the Language and Information Technologies lab at the University of Michigan for the insightful discussions during the early stage of the project. This project was partially funded by a grant from LG AI and by a Microsoft Foundational Model grant. Soujanya Poria was additionally supported by an AcRF MoE Tier-2 grant (Project no. T2MOE2008 and Grantor reference no. MOE-T2EP20220-0017). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of LG AI, Microsoft, or AcRF MoE.

References

Anurag Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. An atlas of cultural commonsense for machine reasoning. *ArXiv*, abs/2009.05664.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra-Aimée Cojocaru, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *ArXiv*, abs/2311.16867.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Annual Meeting of the Association for Computational Linguistics*.

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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*, abs/2005.14165.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Esin Durmus, Karina Nyugen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *ArXiv*, abs/2306.16388.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2020. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *ArXiv*, abs/2102.00894.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for*

Computational Linguistics: ACL 2023, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2020. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. *ArXiv*, abs/2009.12677.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Michael Minkov and Geert Hofstede. 2012. Is national culture a meaningful concept? cultural values delineate homogeneous national clusters of in-country regions. *Cross-Cultural Research*, 46(2):133–159.

Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2022. Extracting cultural commonsense knowledge at scale. *Proceedings of the ACM Web Conference 2023*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Annual Meeting of the Association for Computational Linguistics*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *ArXiv*, abs/1909.01066.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. An adversarial winograd schema challenge at scale.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Conference on Empirical Methods in Natural Language Processing*.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *ArXiv*, abs/2210.09150.

Júnia Coutinho Anacleto Silva, Henry Lieberman, Marie Tsutsumi, Vânia P. A. Neris, Aparecido Augusto de Carvalho, José H. Espinosa, Muriel de Souza Godoi, and Sílvia Helena Zem-Mascarenhas. 2006. Can common sense uncover cultural differences in computer applications? In *IFIP AI*.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002.

Open mind common sense: Knowledge acquisition from the general public. In *OTM Conferences / Workshops*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975.

Vas Taras, Piers Steel, and Bradley L Kirkman. 2016. Does country equate with culture? beyond geography in the search for cultural boundaries. *Management International Review*, 56:455–487.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu 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*, abs/2307.09288.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. In *North American Chapter of the Association for Computational Linguistics*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Conference on Empirical Methods in Natural Language Processing*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

A Additional Results

A.1 Multilingual Models and Encoder-Decoder Models

We test mT0 and bloomZ-7b-mt on the cultural commonsense QA task. Both of them are models pretrained for the multi-lingual setting, and the mT0 is an encoder-decoder model based on mT5. The results in 8 match our previous findings that the native language does not always help. These two models also perform the worst in Iran similar to the other models tested.

Model	US	China	India	Iran	Kenya
mT0-13b	0.536	0.536 | 0.416	0.488 | 0.504	0.288 | 0.608	0.312 | 0.432
BLOOMZ-7b	0.392	0.600 | 0.392	0.616 | 0.592	0.368 | 0.000	0.464 | 0.248

Table 8: GPT3.5's performance on Commonsense QA tasks on additional models

A.2 Few-shot setting

We have tested cultural commonsense QA tasks in a few-shot setting and compared the results with the zero-shot. The few-shot examples are randomly selected from the dataset, using the same format as the zero-shot setting but with the ground truth answer provided for the examples.

Examples	US	China	India	Iran	Kenya
0-shot	0.78	0.78	0.79	0.51	0.57
2-shot	0.78	0.75	0.81	0.54	0.66
5-shot	0.81	0.78	0.82	0.54	0.66

Table 9: GPT3.5's performance on Commonsense QA tasks with few-shot examples