Global Gallery: The Fine Art of Painting Culture Portraits through Multilingual Instruction Tuning

Anjishnu Mukherjee[1] **Aylin Caliskan**[2] **Ziwei Zhu**[1] **Antonios Anastasopoulos**[1]
[1]Department of Computer Science, George Mason University
[2]The Information School, University of Washington
{amukher6,zzhu20,antonis}@gmu.edu aylin@uw.edu

Abstract

Exploring the intersection of language and culture in Large Language Models (LLMs), this study critically examines their capability to encapsulate cultural nuances across diverse linguistic landscapes. Central to our investigation are three research questions: the efficacy of language-specific instruction tuning, the impact of pretraining on dominant language data, and the identification of optimal approaches to elicit accurate cultural knowledge from LLMs. Utilizing the GeoMLaMA benchmark for multilingual commonsense knowledge and an adapted CAMeL dataset (English-only) for evaluation of nuanced cultural aspects, our experiments span six different languages and cultural contexts, revealing the extent of LLMs' cultural awareness. Our findings highlight a nuanced landscape: while language-specific tuning and bilingual pretraining enhance cultural understanding in certain contexts, they also uncover inconsistencies and biases, particularly in non-Western cultures. This work expands our understanding of LLMs' cultural competence and emphasizes the importance of integrating diverse cultural perspectives in their development, aiming for a more globally representative and equitable approach in language modeling.[1]

1 Introduction

Large language models (LLMs) are capable of performing well across a wide variety of tasks (Bommasani et al., 2022; Srivastava et al., 2023) owing to their ability of generating coherent text that draws from a large corpus of pre-training data. However, some tasks like performing open-ended social reasoning involve questions (Parrish et al., 2022) which due to being under-specified or requiring a certain level of critical thinking elicit an opinionated answer from the LLM that affects different social groups, sometimes in undesirable ways (Bender et al., 2021). The role of culture is undeniable when looking at factors that determine people's beliefs and behavior in social settings. Cultural differences exist across countries and they interplay with the dominant language spoken, influencing both individual traits and group behavior. However, it is a well-known problem that multilingual LLMs are trained on corpora that are not equally representative of all parts of the world, but are rather more "western aligned" (Weidinger et al., 2022). This leads to potential issues of misrepresentation of culture and a lack of understanding of cultural knowledge in text generated by LLMs. Other work which studies this question brings out the lack of precision in cultural representations (Ramezani and Xu, 2023), problems of skewing distributions or amplifying biases existing in society (Jakesch et al., 2023), erasing underrepresented nuances (Hutchinson et al., 2020) and also the impact of low resourced languages and cultures they are spoken in

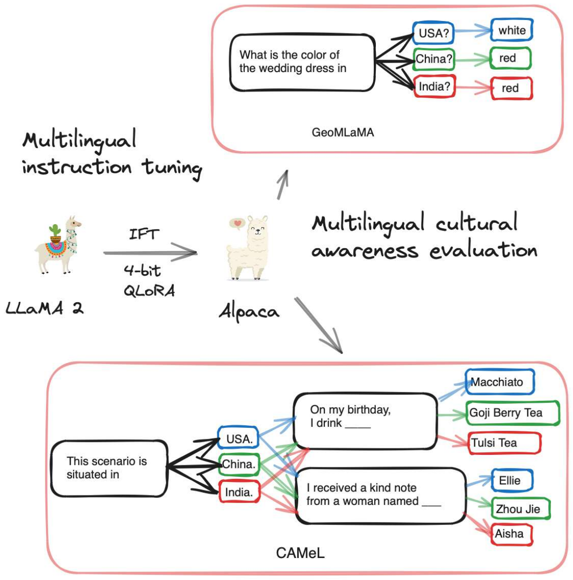
What is the color of the wedding dress in
USA?
China?
India?
white
red
red
GeoMLaMA
Multilingual instruction tuning
IFT
4-bit QLoRA
LLaMA 2
Alpaca
Multilingual cultural awareness evaluation
This scenario is situated in
USA.
China.
India.
On my birthday, I drink ____
Macchiato
Goji Berry Tea
Tulsi Tea
I received a kind note from a woman named ___
Ellie
Zhou Jie
Aisha
CAMeL

Figure 1: We instruction-tune the English-based LLaMA 2 in 5 languages (Hindi, Mandarin Chinese, Persian, Swahili, Greek) and evaluate both general cultural awareness as well as fine-grained cultural understanding in multilingual settings.

[1] Code and data are available: https://github.com/iamshnoo/culture-llm.

(Wibowo et al., 2023). While these studies have laid the groundwork, there remains a gap in understanding how language-specific instruction tuning might correlate with cultural knowledge.

Our work aims to first test the hypothesis that instruction tuning on data in a specific language might improve the cultural awareness of the LLM for the related culture. We also look at the impact of continued pre-training in a bilingual setting and how that has an influence on multilingual cultural understanding. Specifically we design the following research questions for this purpose:

RQ1 : Does instruction tuning on language-specific data enhance cultural knowledge?

RQ2a : Does pretraining on language-specific data enhance cultural knowledge?

RQ2b : What is the optimal approach for eliciting cultural information from LLMs?

For this purpose, we translate the instructional data used for training the Alpaca model (Taori et al., 2023) to five languages (Hindi, Mandarin Chinese, Persian, Swahili, Greek) other than English to cover the six distinct cultures. We then use this data to train low rank adapters (LoRA, Dettmers et al. (2023)) followed by evaluation on a benchmark for multilingual cultural knowledge in each language to measure the impact of instruction tuning.

We also explore whether LLMs understand tangible cultural nuances like food, beverages, clothing, etc by framing a dataset about different social situations with cultural targets based on a previous study (Naous et al., 2023). We use this to ask:

RQ3 : Do LLMs understand the nuances of culture and what disparities exist across tangible cultural aspects?

Overall, our findings show shortcomings in not only how culture is understood by LLMs, but also in current existing approaches at overcoming them.

2 Data and Methods

Our study explores the cultural understanding of large language models (LLMs) through two primary strategies: enhancement of an existing benchmark and the creation of a new, culturally-focused benchmark. Our methodology involves translating instructional data from the Alpaca dataset (which does not include culturally relevant information) into five additional languages and conducting supervised fine-tuning on various LLaMA 2 model (Touvron et al., 2023) sizes using these translations (a sample of which is manually verified by speakers of the dominant language). With this approach we aim to test whether SFT in itself might improve cultural processing even though we are not explicitly training on culture specific data. This enables a broad examination of how LLMs handle cultural nuances across different linguistic contexts.

Base Model	LoRA	Prompt
English	{lang} Alpaca	{lang}
English	English Alpaca	English
{lang}	{lang} Alpaca	{lang}
{lang}	Non-Alpaca LoRA	{lang}

Table 1: The four experimental combinations we test for RQ1 and RQ2. *lang* refers to language-specific variants of Alpaca or a language-specific prompt.

2.1 Data

We work with 2 different datasets to understand cultural awareness at global and granular scales. We use Data Portraits (Marone and Van Durme, 2023) primarily to estimate whether these evaluation sets may overlap with the pretraining corpus, and find that they likely do not.

GeoMLaMA The GeoMLaMA benchmark (Yin et al., 2022) is central to our study on the cultural awareness of language models. Originally containing culturally diverse fill-in-the-blank sentences, we have converted it into a question-answer (QA) format. This adaptation makes it suitable for evaluating decoder-only models. Key features of this benchmark include:

- **Multilingual Scope:** Covers five countries (USA, China, India, Iran, Kenya), each with its dominant language (English, Mandarin Chinese, Hindi, Persian, Swahili). We further expand our investigation by integrating a Greece/Greek variant of the dataset. This addition provides a broader spectrum for analysis, especially for languages that are lower resourced.
- **QA Format:** Consists of 900 multilingual questions (150 questions for each of six languages) with one gold correct and multiple incorrect answers, facilitating a clear assessment of the model's cultural understanding.
- **Cultural Diversity:** Questions cover a range of 17 broad cultural topics (eg. broom usage, climate, driver seat, measurement unit, etc) and are presented in both the country's dominant language and other languages, allowing for a comprehensive cross-cultural evaluation.

For **RQ1**, this dataset allows us to examine whether instruction tuning in a language specific to a given culture leads to better understanding and representation of that culture in language models. For **RQ2a** and **RQ2b**, the GeoMLaMA dataset's multilingual nature helps assess the impact of pretraining language models on language-specific data. Note that, the GeoMLaMA paper defines a metric for the benchmark that is slightly different from raw accuracy (as it accounts for country priors as well), which is what we also use in our experiments.

CAMeL Our study also incorporates the CAMeL dataset, initially introduced by Naous et al. (2023), to conduct detailed cultural analysis. Originally designed to compare Arabic and Western cultural norms, we have adapted the dataset to align with the countries featured in the GeoMLaMA benchmark. This adaptation involves collecting new data from speakers of the dominant language and selecting sentence templates that are broadly applicable across various cultures. Our modified version of the CAMeL dataset is tailored to specifically address **RQ3**, which focuses on the language models' granular understanding of cultural elements.

Key aspects of our adapted CAMeL dataset are:

- **Cultural Adaptation:** We have enriched CAMeL to reflect six cultures from the GeoMLaMA benchmark, involving data collection from speakers of the dominant language and culturally diverse sentence templates.
- **Cultural Categories and Prompts:** The dataset contains nine categories, such as gendered pairs, each with ten unique prompts and around fifty targets, covering a range of cultural elements like food, names, clothing, and literature.
- **QA Scenarios for Granular Analysis:** We create five types of multiple-choice QA scenarios from CAMeL, designed to assess the models' depth of cultural understanding and their ability to distinguish between various cultural elements.

In our experiments, we deploy various multiple-choice QA scenarios derived from both datasets, ensuring that choice order is randomized to mitigate positional bias in large language models (Pezeshkpour and Hruschka, 2023). This approach allows us to comprehensively address each research question, ensuring that our findings are robust and well-supported by empirical evidence.

2.2 Language-specific finetuning data

To investigate the effect of language-specific instruction tuning on cultural awareness, we begin with the 52k instruction-following demonstrations used for training the Alpaca model (Taori et al., 2023), referred to as the cleaned Alpaca dataset. These instructions, originally in English, are translated into six languages (English, Chinese, Hindi, Persian, Swahili, and Greek) using an automatic translation system from the NLLB project (Team et al., 2022). These translations correspond to the dominant languages of six cultures (American, Chinese, Indian, Iranian, Kenyan, and Greek) under study, resulting in the Alpaca-X dataset, where 'X' denotes the respective language for eg., Alpaca-en is the original English Alpaca data, and Alpaca-hi is the translated Hindi version. It is important to note that all these datasets are content-equivalent, only differing in terms of language.

Language	Input	Instruction	Output	Avg
Chinese (zh)	0.78	0.79	0.75	0.77
Greek (el)	0.82	0.83	0.78	0.81
Hindi (hi)	0.84	0.85	0.82	0.84
Persian (fa)	0.83	0.84	0.80	0.82
Swahili (sw)	0.80	0.80	0.77	0.79

Table 2: Reference-free quality estimation for translations from English using CometKiwi shows the high quality of the translated data used for instruction tuning. The metric is on a scale of 0-1, with 1 being perfect translation.

Reference-free Quality Estimation of Machine Translation We use CometKiwi (Rei et al., 2022) for estimating the quality of the the translated Alpaca-X datasets. Each dataset has three columns - input, instructions and output, corresponding to the Alpaca data format. Only considering rows where all the columns have some translatable data (does not contain code, is not empty string), we look at the Quality Estimation (QE) scores in Table 2 which lie between 0 and 1, with 1 representing perfect translation. All the values are around 0.8 on average indicating a high quality of translated data. Note that there are minor variations in the quality between the columns which have longer documents (eg. Output) vs the columns which have shorter documents (eg. Input, Instruction). These high scores for QE align with our initial feedback from speakers of the dominant language who were provided a small random sample (around 200 sentences) of the translated data.

2.3 Supervised Instruction Finetuning

We use 4-bit QLoRA (Dettmers et al., 2023) to train using supervised finetuning (SFT), low-rank adapters (LoRA) for the base models using our

Alpaca-X data, with hyperparameters detailed in the Appendix Table 13. These adapters, specific to each language, can be integrated into the base model in a plug-and-play manner. The base model combined with a language-specific adapter trained on Alpaca-X data is also referred to as an Alpaca-X model, for simplicity of notation. For eg., Alpaca-hi data is used to train an adapter for the Alpaca-hi model. Note that SFT, supervised finetuning, and instruction tuning is used interchangeably throughout the rest of the document.

3 Experimental Settings

We divide our experiments into two distinct categories – first looking at how instruction tuning and pretraining play a role in cultural understanding, and then going deeper into different aspects of cultural nuances.

3.1 Studying the effects of language specific instruction tuning

Our experiments are designed to isolate the impact of different components (base model, LoRA, evaluation prompt) on the cultural awareness of LLMs.

Experimental Setups To address our research questions, we have devised the following experimental setups:

1. **For RQ1 (Language-Specific Instruction Tuning):** We compare the LLaMA 2 model with an English-specific adapter (Alpaca-en) against Alpaca-X models, where 'X' denotes other languages. This comparison helps determine the effectiveness of language-specific instruction tuning in enhancing cultural understanding.
2. **For RQ2a (Language-Specific Pretraining):** We explore bilingual base models for Chinese (Yi)[2] and Swahili (Uliza)[3], each with its respective LoRA, to gauge the impact of language-specific pretraining on cultural knowledge.
3. **For RQ2b (Quality of Fine-Tuning Data):** An ablation study contrasts a non-Alpaca-X adapter, developed from high-quality bilingual data, with our Swahili Alpaca adapter. This helps assess the influence of fine-tuning data quality on cultural understanding.

Distribution of token counts for pretraining and instruction tuning In the context of our experiments, the token counts for pre-training and instruction tuning vary significantly. The Alpaca-X models are the only components developed in-house, while the pretrained bilingual models and Swahili LoRA are sourced from open-source repositories.

The LLaMA 2 model underwent pre-training with a substantial 2 trillion tokens. For the Swahili base model, a continued pre-training phase incorporated 0.32 billion Swahili tokens. In contrast, the Chinese base model involved pre-training with a combined total of 3 trillion Chinese and English tokens. The Alpaca dataset used for instruction tuning is relatively small, consisting of 52,000 instructions, which translates to approximately 5.817 million tokens for the English Alpaca, or 0.005 billion tokens. The token count for the non-Alpaca LoRA, used in one of our ablation studies, remains unknown because the open source repository it is adapted from does not specify details. This disparity in token counts highlights the differences in data scale between pre-training and instruction tuning phases which might have some effect on our results that we cannot control.

Further, note that all these controlled experiments are performed using the first dataset (GeoMLaMA) without going into granular details, because our focus was to study the effects of instruction tuning instead of different cultural aspects.

3.2 Granular Analysis of Cultural Aspects

In this section, we outline a series of experiments utilizing the CAMeL dataset to conduct an in-depth analysis of cultural aspects. These experiments are designed to evaluate the model's nuanced understanding of cultural elements.

Setting 1: A multiple-choice question is framed with one option representing the answer from the corresponding culture and five options from the other cultures (all other settings are restricted to four options). This setting aims to assess the model's comprehension of individual cultural aspects rather than a general overview.

Setting 2: No options from the correct culture are provided in the multiple-choice questions, options are randomly sampled uniformly from incorrect cultures. This approach is intended to determine the model's default cultural inclination when the correct option is absent.

Setting 3: Three options from the correct culture are provided alongside one option from a randomly selected incorrect culture. A model with accurate cultural understanding should consistently avoid

[2] Huggingface link for Yi

[3] Huggingface link for Uliza (pre-trained model)

choosing the incorrect option. This setting tests the model's alignment with the findings from the previous settings.

Setting 4: Each question includes four options from the correct culture, but three of these are from a different category than what the question addresses to test for precision of understanding. For instance, in a question about names, three options might be food items, with only one being a name. The model's ability to discern between categories within the same culture is evaluated here. If the model understands culture minutely enough to be able to differentiate between the categories we are asking about, then it would never pick an option from the incorrect category. But if it only has a fuzzy understanding of culture, then it might end up choosing any of the given options as all of them are "culturally correct" in a global sense.

Setting 5: Questions regarding gendered categories are used, where half of the options are correct for the grammatical gender but incorrect for the culture, and the other half are correct for the culture but incorrect for the grammatical gender. This setting tests whether the model prioritizes cultural accuracy over grammatical gender accuracy in its responses. For instance, in a question about American female names with options as Liam, David, Aisha and Divya, we expect the model to choose one between Aisha and Divya over the two male names. Ideally, the model response should stick to the correct grammatical gender, because typically female clothing is worn by females and male clothing by males, and similarly usually females have female-associated names and vice versa. But if the model responses stick to the correct culture and ignore gender, then the model does not necessarily understand the details of the gendered cultural aspect even if it is broadly culturally correct. Note that we do not refer to gender identity in this analysis, but only focus on grammatical gender.

3.3 Evaluation technique

Our evaluation method draws inspiration from existing approaches that aggregate token log-probabilities for prompt completion. Specifically, the techniques used by Trinh and Le (2019) and Wang et al. (2023b) utilize variations of this concept of using aggregated token log-probabilities in determining the most likely prompt completions.

CAPPr Further building on this approach, we utilize CAPPr (K. Dubey, 2023), a tool that implements the aforementioned idea by selecting the completion most likely to follow a given prompt. CAPPr achieves this by calculating the log-probability of each token in a completion, considering both the prompt tokens and the preceding tokens within that completion. This process involves averaging the log-probabilities to derive the inverse perplexity of the completion. Subsequently, these averaged log-probabilities are exponentiated to obtain a completion probability. This procedure is repeated for each potential completion to form a normalized probability distribution over the set of completions, which for our use case represents the different options in a QA setting.

Example Let's take the prompt "This is a " and the possible completion as "cat". (Note that the prompt ends in a whitespace, which is the default in the library and also what we follow when formatting our prompts.) These are concatenated to form the text "This is a cat". This sentence is passed through the tokenizer to obtain encodings, which are then passed through the model to obtain the logits for the entire sequence. Then, log-softmax is applied to these logits and input IDs are sliced out to get log-probabilities for completion tokens.

Evaluating using log probabilities instead of generations In the evaluation method we are using, the model is not ***generating*** text. We format our prompts as "Answer in one word. Choose between the given options. ###Question: Which side of the car is the driver seat in the United States? ###Options: (a) left (b) right ###Answer: " in the Alpaca data format, and then concatenate each of the options (left/ right) one by one to this sequence. For each complete sequence, we encode it and pass it through the model to obtain logits and then take a log softmax over the entire sequence, before splicing the input to get log probabilities for the completion. The probability for the next token being left or right in this case is not negligible, given the formatting, and it is noticeable from the log probabilities that we obtain (as compared to log probabilities for some other irrelevant option like "anywhere" or "middle"). The probability for the answer with the highest probability is usually always more than random for the majority of questions in our data.

Why we are not evaluating generations instead We initially tried to find a reasonable solution to the evaluation of generation problem. Our best approach (the method that gave the highest "accu-

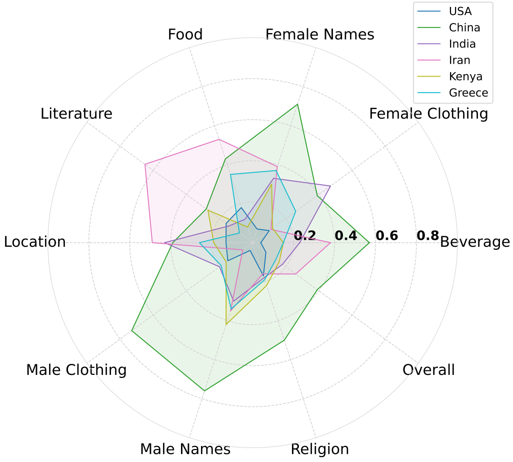

Figure 2: The 70B LLaMA 2 model shows strong performances for China and Iran across cultural concepts for different cultures.

racy" scores on our datasets) involved generating answers with a beam search for a given combination of parameter values for top-p, top-k, temperature, followed by extraction of the relevant part of the answer using a QA system trained on different existing multilingual QA datasets (MNLI, etc.) and then looking for similarities between the extracted answer and the gold answer using a 2 level approach (first looking for exact matches and then looking for similarities based on BERTScore (Zhang et al., 2020)). But even though this approach would lead to "higher scores" on the tasks we define, the numbers were found to be misleading because they do not correlate with human evaluations of the generations which is why we do not report the results from that approach in our paper, even though it would have made for an interesting discussion about evaluation approaches for LLMs.

4 Results

Our findings show that LLMs do *not* understand the specific details that define culture even when we try different approaches like SFT, bilingual pretraining, and prompting in the dominant language.

4.1 RQ1 : Does SFT on language specific data enhance cultural knowledge? (No)

Our investigation into whether SFT on language-specific data enhances cultural knowledge involves a series of experiments, detailed in Appendix 6, Table 7. This section focuses on key results pertinent to our hypotheses for the RQ.

Better than BERT, but only in English Analysis of the GeoMLaMA performance (Table 3) compares the English base model combined with language-specific Alpaca and language-specific prompts against the English base model with English Alpaca and English prompts across three different model sizes ranging from 7B to 70B. Larger model sizes do not necessarily mean better cultural knowledge. We restrict our analysis to non-USA countries where English is not the dominant language. Results indicate that instruction tuning in English slightly outperforms encoder models like BERT and XLMR, a trend not always observed when SFT is applied in other languages. This discrepancy may be attributed to the predominance of English in LLaMA 2's pretraining data, making it the language that is most coherent for the model. We note that while some amount of cultural data is definitely present in the pretraining data, our SFT instructional data does not include cultural content.

SFT lang	China	India	Iran	Kenya	Greece
Results from GeoMLaMA benchmark					
(mBERT)	0.30	0.41	0.21	0.30	-
(XLMR-L)	0.37	0.37	0.37	0.32	-
Prompt language: english					
eng (7)	0.50	0.39	0.24	0.31	0.34
eng (13)	0.54	0.42	0.31	0.28	0.34
eng (70)	0.46	0.45	0.28	0.28	0.38
Prompt language: {lang}					
{lang} (7)	0.25	0.39	0.31	0.31	0.28
{lang} (13)	0.32	0.36	0.28	0.34	0.28
{lang} (70)	0.39	0.33	0.14	0.34	0.34

Table 3: Instruction tuning on language specific data does not consistently enhance cultural knowledge across languages and cultures. The numbers 7, 13 and 70 correspond to the model sizes in billions of parameters. The metric is the GeoMLaMA benchmark metric (country priors are subtracted from calculated accuracies) on a scale of 0-1 with higher being better.

No clear enhancement due to SFT, with a lot of variability across cultures The hypothesis that SFT with language-specific data substantially improves cultural knowledge is not conclusively supported by our findings. There is notable variability across different cultures. For instance, in China and Iran, English-based fine-tuning seems to be more effective, while in India, Hindi fine-tuning competes closely. English emerges as the most effective language for eliciting cultural knowledge across various cultures. However, the second-most effective language is not consistently the dominant one. For example, prompting in Chinese yields better results than Swahili for Kenyan cultural questions

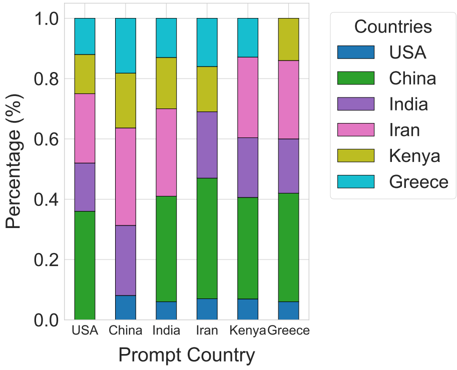

Figure 3: The distribution of countries chosen by the 70B LLaMA 2 model without the question explicitly mentioning the chosen country shows a large percentage favouring China and Iran.

(Table 7). This observation might be influenced by the larger representation of Chinese compared to Swahili in the pretraining data set.

4.2 RQ2a : Does pretraining on dominant language data enhance cultural knowledge? (Yes)

Our study also probes the influence of pretraining language distribution on cultural understanding. Specifically, we contrast the LLaMA 2 model, primarily pretrained on English data, with bilingual base models for Chinese (Yi) and Swahili (Uliza). The performance comparisons are in Table 4. Note that pretraining data potentially contains more, non-translated, culturally specific data which we do not control for, so the effects observed in this section should not be directly compared to the previous section about instruction tuning.

Pretraining is useful in improving cultural understanding along with the instruction tuning and language specific prompting In the context of English language queries, both Yi and Uliza models do not surpass the performance of LLaMA 2. However, for queries related to China, when prompted in Chinese, the Yi model demonstrates superior performance compared to LLaMA 2 and also achieves parity with LLaMA 2's English performance. Similarly, for Kenyan cultural queries, the Uliza model, when prompted in Swahili, matches LLaMA 2's performance.

The quality of pretraining data matters for increased awareness across cultures Notably, the Yi model generally outperforms LLaMA 2 in English for cultures outside China (USA, Iran, Kenya, Greece), as shown in Table 7. This suggests that high-quality, filtered pretraining data, particularly when used for continued pre-training, play an important role in enhancing a model's cultural awareness across different cultures.

Model	Size	China	Kenya
Prompt language : English			
LLaMA 2 + eng Alpaca	7	0.50	0.31
	13	0.54	0.28
	70	0.46	0.28
Yi + eng Alpaca	6	0.43	-
	34	0.39	-
Uliza + eng Alpaca	7	-	0.25
Uliza + {swa, eng} LoRA	7	-	0.31
Prompt language : Chinese/Swahili			
LLaMA 2 + zh/swa Alpaca	7	0.25	0.31
	13	0.32	0.34
	70	0.39	0.34
Yi + zh Alpaca	6	0.39	-
	34	0.54	-
Uliza + swa Alpaca	7	-	0.31
Uliza + {swa, eng} LoRA	7	-	0.41

Table 4: Pretraining on language specific data helps to improve cultural awareness. Bilingual non-alpaca finetuning along with bilingual continually pretrained model gives the most culturally appropriate responses when prompted in the respective dominant language.

Regarding the "difference in quality of data" between the Yi and LlaMA 2 models, while we can't quantify this without access to the pretraining data, our comment is based on direct feedback from a lead developer of the Yi model. They indicated that "higher quality pre training data" was a key factor in Yi's superior performance, particularly for Mandarin Chinese, compared to LlaMA 2.

4.3 RQ2b: Optimal Approach for Eliciting Cultural Knowledge

In an ablation study focusing on the quality of finetuning data, we examine a non-Alpaca LoRA derived from carefully curated Swahili data.[4] The results indicate a clear superiority of the curated LoRA over our Swahili Alpaca, as it surpasses both Alpaca and also LLaMA 2 for English prompts.

This finding underscores that the most effective approach for eliciting accurate cultural knowledge involves a bilingual base model pre-trained on high-quality, language-specific data. Additionally, supplementing this model with a LoRA, instruction-tuned on curated instructional examples and prompted in the respective language, further enhances its cultural understanding. Such a com-

[4] Huggingface link for Uliza (finetuned model)

bination of high-quality pretraining, targeted instruction tuning, and language-specific prompting emerges as the optimal strategy for achieving deep cultural insight. This also implies SFT with curated instructional examples performs better than SFT using generic machine-translated data.

4.4 RQ3 : Do LLMs understand granular tangible cultural aspects? (somewhat)

This part of our study, centered on English, aims to delve into the nuanced cultural understanding of models, using the CAMeL dataset. We chose English for several reasons: the complexity of translating proper nouns, the redundancy of translated nouns representing the same concept, and previous findings indicating superior English performance unless using a language-specific pretrained model.

Prior distributions of cultural aspects of countries affect cultural understanding at the granular level Our analysis reveals that the model displays a pronounced preference for certain cultures, particularly China and Iran, when no correct options are present (Setting 2, Figure 3). This inherent bias significantly affects performance for other cultures when correct options are included.

As illustrated in Figure 2, LLaMA 2 generally exhibits the highest performance across various cultural aspects for China (Setting 1), with some exceptions where Iran leads. However, despite previous research (Weidinger et al., 2022; Durmus et al., 2023) indicating alignment with American values, the model shows a relatively superficial understanding of American culture, as evidenced by its lower performance. A possible explanation could be that the other options provided for the question have higher prior distributions, but there are possibly multiple factors at play here.

In another test (Setting 3), we present a scenario where three options are from the correct culture, along with one option from a randomly selected, incorrect culture. This setup is intended to evaluate the model's ability to discern cultural appropriateness accurately. Our findings reveal a stark contrast in performance based on the cultural context. For questions pertaining to China, the model demonstrates a high degree of accuracy, rarely selecting the incorrect cultural option. In contrast, when presented with questions about male names in the US, the model's performance significantly declines, choosing the incorrect option nearly 70% of the time. This disparity highlights the model's uneven capability in distinguishing between culturally relevant and irrelevant options across different cultural settings. An alternate plausible explanation could be that model has learned that the US is a largely multi-cultural society, which our evaluation approach isn't designed to consider. Such biases could be attributed to the mixed cultural perspectives inherent in the pretraining data, which might emphasize certain cultures over others. Appendix, Tables 8, 9, and 10 provide detailed results.

Category	USA	China	India	Iran	Kenya	Greece
Beverage	0.56	0.31	0.67	0.40	0.50	0.63
Female Clothing	0.60	0.69	0.58	0.81	0.79	0.69
Female Names	0.89	0.87	0.97	0.82	0.92	0.85
Food	0.32	0.40	0.76	0.32	0.69	0.28
Literature	0.21	0.33	0.45	0.20	0.34	0.65
Location	0.81	0.88	0.76	0.72	0.84	0.81
Male Clothing	0.58	0.54	0.85	0.86	0.75	0.74
Male Names	0.94	0.85	0.97	0.85	0.93	0.87
Religion	0.51	0.55	0.81	0.72	0.53	0.66
Overall	0.60	0.61	0.76	0.65	0.70	0.69

Table 5: We measure the percentage of times that LLaMA 2 70B prefers an option from an incorrect category when provided with a single choice from the correct category paired with 3 incorrect ones. Ideally, this should be close to 0 if the model has true understanding.

Complex cultural understanding is lacking even for countries for which the model understands culture broadly Our investigation further explores the model's depth of cultural understanding through a specific testing approach. In this setup, each question offers four culturally appropriate options, but only one option is relevant to the question's category, while the other three belong to different categories. The assessment focuses on the frequency with which the model selects an option from an incorrect category. Notably, even for countries like China and Iran, where the model generally shows a good grasp of broader cultural aspects, the selection of incorrect category options is alarmingly high, as detailed in Table 5. Ideally, the model should have a near-zero selection rate of incorrect categories for countries with strong cultural representation. However, this nuanced understanding appears to be lacking.

A striking example involves the 13B model's interpretation of Chinese female names. In an array of approximately 3K questions, the model consistently showed a preference for beverage names over

Category	USA	China	India	Iran	Kenya	Greece
Female Clothing	0.40	0.83	0.59	0.31	0.44	0.54
Male Clothing	0.51	0.66	0.77	0.42	0.48	0.68
Female Names	0.14	0.90	0.58	0.61	0.66	0.58
Male Names	0.23	0.93	0.57	0.65	0.68	0.65
Overall	0.32	0.89	0.63	0.51	0.56	0.61

Table 6: LLaMA 2 70B prefers being culturally correct than being grammatically gender correct across cultures.

actual female names. For instance, it judged 'Goji Berry tea' as a more probable name than 'Chen' in the given context of 'I met a girl named [fill in] at the park'. While it's conceivable that the model might not always err when presented with non-beverage incorrect options, the fundamental issue remains that it should not select such incongruent options at all. Comprehensive results of this testing are available in the Appendix, Table 11, underscoring the model's limitations in distinguishing between specific cultural categories.

Being culturally accurate is preferred by LLaMA 2 over being grammatically gender accurate, even though it should be the opposite In a nuanced test, we presented options that juxtapose cultural accuracy against grammatical gender accuracy: two options correct in culture but incorrect in gender, and two others correct in gender but incorrect in culture. The results reveal a marked preference for cultural accuracy over gender accuracy, particularly in contexts where the cultural representation in the model's training data is more pronounced (Table 6 and Appendix, Table 12).

This tendency is more evident in countries with a higher cultural prominence in the model's training data. For instance, in questions related to China, the model predominantly selects culturally accurate responses, regardless of grammatical gender correctness. Conversely, for countries like the USA, the model shows a greater propensity to choose options that are correct in terms of gender. This pattern suggests that the prominence of certain cultural or gender concepts in the pretraining corpus along with grammatical gender signals in language significantly influences the model's decision-making process, underscoring the impact of training data composition on the model's understanding of nuanced cultural and gender-related aspects.

5 Related Work

In the context of understanding cultural biases in Large Language Models (LLMs), several studies have made significant contributions, each addressing different aspects of this multifaceted issue. Tao et al. (2023) use the World Values Survey to map GPT models on the Inglehart-Welzel Cultural Map, highlighting the effectiveness of cultural prompting as a mitigation strategy. Durmus et al. (2023) combine datasets from the World Values Survey and the Pew Research Center's Global Attitudes surveys to explore models' alignment with Western values, using various prompting techniques. The SeaEval benchmark (Wang et al., 2023a) demonstrates the challenges multilingual LLMs face in multicultural reasoning, affected by factors like positional bias and language nuances. COPAL-ID (Wibowo et al., 2023) finds that LLMs have a lower understanding of culture-related questions compared to non-culture related ones, especially in multilingual settings. Additionally, Cao et al. (2023) pioneered examining cultural alignment for chatbots, revealing ChatGPT's American-centric alignment.

However, these studies collectively highlight some gaps: a predominant focus on Western-centric perspectives, limited exploration of non-Western cultures, and the need for more comprehensive strategies to incorporate a global spectrum of cultural nuances in LLMs. Our work aims to build upon these findings, addressing these shortcomings by examining LLMs' cultural awareness more holistically and inclusively.

6 Conclusion

This study on the cultural understanding of Large Language Models (LLMs) reveals significant variations in their ability to encapsulate diverse cultural nuances. Our investigations, leveraging the GeoMLaMA benchmark and the adapted CAMeL dataset, demonstrate that while language-specific instruction tuning and bilingual pretraining offer some improvements, they fall short of ensuring comprehensive cultural competence, particularly in non-Western contexts. The findings underscore the need for incorporating a wider range of cultural perspectives in LLM training and development, highlighting the importance of creating models that are not only linguistically adept but also culturally sensitive and globally inclusive.

Limitations

This study, while extensive, is subject to certain limitations which are important to acknowledge:

1. The current methodology conceptualizes culture as a singular entity within a nation-state. This perspective, while useful for structured analysis, might not fully capture the rich diversity and complexity of modern societies, where multiple cultures and languages coexist within a single country. We try to overcome this issue of "culturally stereotyping" our datasets by looking at aggregated statistics of discrepancies in model responses, instead of focusing on individual responses to each prompt. Future research could benefit from exploring more granular approaches that can effectively address this multifaceted nature of cultural identity.
2. We don't control the token distribution for the pretraining process lacks contrasting with the controlled instructional data used in fine-tuning experiments. This could affect result interpretation. Future work should investigate the effects of smaller, high-quality datasets for controlled pre-training across languages.
3. Our experiments use 4-bit QLoRA for instruction tuning, and it's uncertain if results would differ with higher-bit configurations. Further research is needed to explore the impact of varying bit settings.
4. Evaluating large language models is an ongoing challenge within the field, and the methodology chosen for this study, while grounded in established research, has its strengths and limitations. This approach needs to be considered alongside alternative evaluation methods, each with their respective advantages and drawbacks, to suit specific use cases and research objectives.
5. Some of the languages tested (Greek, Persian, Hindi) may not be supported by the tokenizer of LlaMA 2, which is also why we recommend continued pretraining and/or developing custom extended LLaMA tokenizers, for improving language specific cultural awareness.
6. An angle that we do not explore in the paper is finetuning on culturally relevant data, because the definition of what is culturally relevant is nuanced and lacks a clear definition across communities. However, we believe that there are many exciting new works in this area of research which will enable us to soon be able to do so.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive feedback. This work was generously supported by the National Science Foundation under awards IIS-2327143 and IIS-2125466, and the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB23D194. Computational resources for experiments were provided by the Office of Research Computing at George Mason University (URL: `https://orc.gmu.edu`) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST or the National Science Foundation.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai

Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Esin Durmus, Karina Nyugen, Thomas I. 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.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

K. Dubey. 2023. Cappr (version 0.8.7) [computer software].

Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

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

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-

López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2023. Auditing and mitigating cultural bias in llms.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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.

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.

Trieu H. Trinh and Quoc V. Le. 2019. A simple method for commonsense reasoning.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix

Prompt	Model	Size	US	China	India	Iran	Kenya	Greece	Overall
English	LLaMA 2 + English Alpaca	7B	0.28	0.50	0.39	0.24	0.31	0.34	0.34
		13B	0.31	0.54	0.42	0.31	0.28	0.34	0.37
		70B	0.31	0.46	0.45	0.28	0.28	0.38	0.36
	Yi + English Alpaca	6B	0.52	0.43	0.33	0.48	0.50	0.34	0.43
		34B	0.62	0.39	0.42	0.45	0.50	0.44	0.47
	Uliza + English Alpaca	7B	0.21	0.50	0.39	0.17	0.25	0.31	0.31
	Uliza + {Swahili, English} LoRA	7B	0.45	0.39	0.39	0.34	0.31	0.25	0.36
Hindi	LLaMA 2 + Hindi Alpaca	7B	0.28	0.46	0.39	0.34	0.25	0.41	0.36
		13B	0.24	0.36	0.36	0.28	0.31	0.38	0.32
		70B	0.24	0.46	0.33	0.28	0.34	0.38	0.34
Chinese	LLaMA 2 + Chinese Alpaca	7B	0.34	0.25	0.39	0.41	0.41	0.34	0.36
		13B	0.38	0.32	0.39	0.48	0.47	0.38	0.40
		70B	0.38	0.39	0.42	0.48	0.53	0.34	0.43
	Yi + Chinese alpaca	6B	0.38	0.39	0.45	0.34	0.25	0.31	0.36
		34B	0.55	0.54	0.55	0.45	0.44	0.53	0.51
Swahili	LLaMA 2 + Swahili Alpaca	7B	0.34	0.32	0.39	0.17	0.31	0.34	0.31
		13B	0.34	0.29	0.39	0.24	0.34	0.34	0.33
		70B	0.31	0.36	0.39	0.21	0.34	0.38	0.33
	Uliza + Swahili Alpaca	7B	0.31	0.46	0.45	0.28	0.31	0.38	0.37
	Uliza + {Swahili, English} LoRA	7B	0.38	0.32	0.36	0.48	0.41	0.34	0.38
Persian	LLaMA 2 + Persian Alpaca	7B	0.31	0.25	0.27	0.31	0.38	0.38	0.32
		13B	0.31	0.25	0.33	0.28	0.34	0.34	0.31
		70B	0.28	0.36	0.33	0.14	0.25	0.38	0.29
Greek	LLaMA 2 + Greek Alpaca	7B	0.17	0.21	0.27	0.10	0.25	0.28	0.22
		13B	0.21	0.21	0.30	0.17	0.28	0.28	0.24
		70B	0.28	0.21	0.33	0.14	0.22	0.34	0.25

Table 7: RQ1, RQ2: Cultural performance scores of various models on the GeoMLaMA benchmark. Values are between 0 and 1, higher is better.

Data collection from dominant speakers for adapted CAMeL dataset We provided native speakers with a list of words that we procured from different sources on the internet and from large language models as a base collection for each category that they are then asked to verify and correct with more appropriate targets for each category based on their lived experiences.

For the prompts, we follow a similar process, but this time we don't require country specific prompts, only category specific. The final set of prompts is decided by agreement between the authors.

We note that this process has inherent biases for the group of people who perform the tasks, which might implicitly show up in the data in unobserved ways. Also, because two of the categories are about names of people, this may include information about someone's real name, but that would only be so, because it is a common name in some part of their country.

All annotators are demographically located in the USA and are between 25-40 years old. Other than the Hindi annotator who is female, all others identify as male. Also, we note that all annotators are either authors or close friends of authors who did not require any form of compensation.

Note that, the phrase "native" has historical (and sometimes perjorative) connotations with "indigenous", which is not the intended meaning here, so we use "dominant language" throughout the paper.

Category	Size	USA	China	India	Iran	Kenya	Greece
Beverage	7	0.03	0.44	0.16	0.34	0.10	0.09
	13	0.04	0.49	0.16	0.35	0.11	0.09
	70	0.04	0.57	0.23	0.38	0.15	0.15
Female Clothing	7	0.09	0.24	0.42	0.08	0.12	0.13
	13	0.11	0.30	0.46	0.09	0.15	0.14
	70	0.10	0.39	0.47	0.11	0.12	0.26
Female Names	7	0.05	0.50	0.21	0.28	0.19	0.18
	13	0.05	0.52	0.32	0.38	0.24	0.33
	70	0.07	0.71	0.33	0.39	0.30	0.37
Food	7	0.06	0.28	0.06	0.47	0.04	0.23
	13	0.07	0.33	0.08	0.49	0.06	0.21
	70	0.18	0.43	0.12	0.53	0.08	0.35
Literature	7	0.10	0.29	0.10	0.34	0.17	0.07
	13	0.12	0.27	0.12	0.39	0.19	0.06
	70	0.16	0.28	0.14	0.65	0.27	0.08
Location	7	0.09	0.28	0.27	0.39	0.13	0.17
	13	0.07	0.35	0.36	0.43	0.18	0.23
	70	0.13	0.39	0.43	0.49	0.19	0.26
Male Clothing	7	0.08	0.62	0.11	0.06	0.17	0.11
	13	0.09	0.68	0.18	0.10	0.19	0.12
	70	0.15	0.73	0.20	0.06	0.16	0.19
Male Names	7	0.02	0.53	0.22	0.26	0.30	0.20
	13	0.04	0.58	0.30	0.33	0.38	0.30
	70	0.04	0.76	0.30	0.35	0.42	0.34
Religion	7	0.16	0.41	0.09	0.11	0.24	0.11
	13	0.15	0.51	0.14	0.14	0.23	0.14
	70	0.17	0.50	0.18	0.16	0.22	0.19
Overall	7	0.08	0.39	0.18	0.26	0.16	0.14
	13	0.08	0.39	0.18	0.26	0.16	0.14
	70	0.08	0.39	0.18	0.26	0.16	0.14

Table 8: RQ3: Setting1 Results (Default MCQ setting, single correct country choice provided) from the CAMeL benchmark.

Prompt	Size	USA	China	India	Iran	Kenya	Greece
USA	7	0.0	0.36	0.17	0.22	0.14	0.12
	13	0.0	0.33	0.18	0.22	0.15	0.12
	70	0.0	0.36	0.16	0.23	0.13	0.12
China	7	0.1	0.0	0.24	0.27	0.22	0.17
	13	0.09	0.0	0.25	0.29	0.21	0.16
	70	0.08	0.0	0.23	0.32	0.18	0.18
India	7	0.07	0.37	0.0	0.24	0.18	0.14
	13	0.07	0.33	0.0	0.27	0.19	0.13
	70	0.06	0.35	0.0	0.29	0.17	0.13
Iran	7	0.09	0.40	0.21	0.0	0.16	0.14
	13	0.08	0.37	0.24	0.0	0.17	0.14
	70	0.07	0.40	0.22	0.0	0.15	0.16
Kenya	7	0.08	0.37	0.19	0.23	0.0	0.13
	13	0.08	0.34	0.21	0.25	0.0	0.13
	70	0.07	0.34	0.20	0.27	0.0	0.13
Greece	7	0.08	0.37	0.18	0.22	0.15	0.0
	13	0.07	0.33	0.20	0.23	0.16	0.0
	70	0.06	0.36	0.18	0.26	0.14	0.0

Table 9: RQ3: Setting2 Results (Distribution of Countries chosen when correct country is not provided) from the CAMeL benchmark

Category	Size	USA	China	India	Iran	Kenya	Greece
Beverage	7	0.48	0.05	0.22	0.09	0.34	0.37
	13	0.47	0.06	0.24	0.1	0.29	0.32
	70	0.5	0.05	0.19	0.08	0.25	0.27
Female Clothing	7	0.35	0.12	0.04	0.4	0.27	0.29
	13	0.37	0.12	0.04	0.35	0.25	0.28
	70	0.36	0.1	0.06	0.36	0.3	0.19
Female Names	7	0.48	0.07	0.19	0.16	0.21	0.2
	13	0.46	0.11	0.14	0.15	0.21	0.19
	70	0.44	0.04	0.2	0.17	0.19	0.17
Food	7	0.38	0.13	0.37	0.04	0.47	0.13
	13	0.37	0.12	0.34	0.04	0.45	0.17
	70	0.31	0.11	0.32	0.05	0.43	0.14
Literature	7	0.3	0.14	0.34	0.1	0.22	0.44
	13	0.28	0.14	0.29	0.07	0.22	0.45
	70	0.32	0.14	0.32	0.05	0.19	0.43
Location	7	0.42	0.13	0.14	0.11	0.28	0.26
	13	0.48	0.11	0.12	0.12	0.25	0.24
	70	0.45	0.13	0.13	0.08	0.29	0.24
Male Clothing	7	0.3	0.02	0.3	0.39	0.28	0.32
	13	0.32	0.02	0.24	0.31	0.25	0.25
	70	0.28	0.02	0.23	0.35	0.29	0.22
Male Names	7	0.65	0.06	0.19	0.21	0.16	0.18
	13	0.67	0.09	0.15	0.2	0.14	0.2
	70	0.67	0.02	0.17	0.2	0.17	0.17
Religion	7	0.27	0.08	0.32	0.36	0.2	0.28
	13	0.27	0.04	0.28	0.3	0.18	0.29
	70	0.27	0.06	0.28	0.28	0.17	0.23
Overall	7	0.4	0.09	0.23	0.21	0.27	0.28
	13	0.41	0.09	0.2	0.18	0.25	0.27
	70	0.4	0.08	0.21	0.18	0.25	0.23

Table 10: RQ3: Setting3 Results from the CAMeL benchmark (How many times did Llama choose the single incorrect option ignoring the other correct options. This number should ideally be 0 for everything.)

Category	Llama_Size	USA	China	India	Iran	Kenya	Greece
Overall	7	0.68	0.72	0.52	0.67	0.79	0.75
	13	0.73	0.71	0.72	0.72	0.78	0.62
	70	0.6	0.61	0.76	0.65	0.7	0.69
beverage	7	0.61	0.47	0.34	0.44	0.76	0.74
	13	0.66	0.43	0.53	0.64	0.68	0.58
	70	0.56	0.31	0.67	0.4	0.5	0.63
female_clothing	7	0.65	0.83	0.37	0.83	0.85	0.77
	13	0.68	0.81	0.62	0.83	0.78	0.57
	70	0.6	0.69	0.58	0.81	0.79	0.69
female_names	7	0.92	0.98	0.8	0.78	0.96	0.93
	13	0.98	0.99	0.93	0.94	0.97	0.77
	70	0.89	0.87	0.97	0.82	0.92	0.85
food	7	0.52	0.58	0.3	0.34	0.8	0.35
	13	0.47	0.46	0.63	0.33	0.8	0.22
	70	0.32	0.4	0.76	0.32	0.69	0.28
literature	7	0.26	0.37	0.23	0.52	0.41	0.6
	13	0.4	0.49	0.38	0.39	0.42	0.56
	70	0.21	0.33	0.45	0.2	0.34	0.65
location	7	0.8	0.94	0.6	0.66	0.94	0.93
	13	0.94	0.93	0.83	0.75	0.97	0.74
	70	0.81	0.88	0.76	0.72	0.84	0.81
male_clothing	7	0.74	0.65	0.59	0.81	0.8	0.81
	13	0.78	0.55	0.82	0.83	0.78	0.6
	70	0.58	0.54	0.85	0.86	0.75	0.74
male_names	7	0.95	0.94	0.78	0.85	0.93	0.91
	13	0.99	0.99	0.93	0.88	0.93	0.87
	70	0.94	0.85	0.97	0.85	0.93	0.87
religion	7	0.69	0.71	0.63	0.76	0.63	0.71
	13	0.63	0.65	0.78	0.78	0.67	0.62
	70	0.51	0.55	0.81	0.72	0.53	0.66

Table 11: RQ3: Setting4 Results from the CAMeL benchmark (How many times did Llama choose an option from the incorrect category) (it was given 3 incorrect categories, 1 correct category) - Ideally this should be 0 for everything if llama understands what category we are asking about.

Category	Size	USA	China	India	Iran	Kenya	Greece
Female Clothing	7	0.37	0.73	0.53	0.26	0.44	0.43
	13	0.38	0.85	0.59	0.34	0.49	0.41
	70	0.4	0.83	0.59	0.31	0.44	0.54
Female Names	7	0.12	0.85	0.53	0.52	0.65	0.46
	13	0.14	0.8	0.64	0.59	0.69	0.51
	70	0.14	0.9	0.58	0.61	0.66	0.58
Male Clothing	7	0.51	0.64	0.79	0.36	0.54	0.59
	13	0.48	0.68	0.8	0.45	0.52	0.56
	70	0.51	0.66	0.77	0.42	0.48	0.68
Male Names	7	0.21	0.85	0.59	0.55	0.6	0.56
	13	0.22	0.82	0.61	0.62	0.57	0.56
	70	0.23	0.93	0.57	0.65	0.68	0.65
Overall	7	0.3	0.82	0.61	0.44	0.55	0.51
	13	0.3	0.8	0.66	0.52	0.57	0.51
	70	0.32	0.89	0.63	0.51	0.56	0.61

Table 12: RQ3: Setting5 Results for the CAMeL benchmark(How many times did Llama choose correct culture but incorrect grammatical gender?) (2 options were from correct culture but opposite gender, and 2 options were from incorrect culture but correct gender)

Parameter	Value
Random Seed	42
Number of Epochs	1 (for 34B or 70B models), 3 (for 6B, 7B, 13B models)
Bits and Bytes Config	
Load	4 bit
Quantization Type	nf4
DataType	bfloat16
Lora Config	
Lora Alpha	16
Lora Dropout	0.1
R	64
Bias	none
Training Arguments	
Per Device Train Batch Size	6 (1 A100 80GB GPU)
Gradient Accumulation Steps	2
Learning Rate	3e-4
Max Gradient Norm	0.3
Warmup Ratio	0.03
Learning Rate Scheduler	constant
Optimizer	32bit paged AdamW
Max Sequence Length	2048

Table 13: Hyperparameters used for Instruction tuning of the LLaMA 2 models