Evaluating the Culture-awareness in Pre-trained Language Model

Stanford CS224N Custom Project

Ryan Li
Department of Computer Science
Stanford University
`lansong@stanford.edu`

Yutong Zhang
Department of Computer Science
Stanford University
`yutongz7@stanford.edu`

Zhiyu Xie
Department of Computer Science
Stanford University
`zhiyuxie@stanford.edu`

Abstract

Pretrained large language models (LLMs) have significantly advanced various applications but continue to encounter challenges regarding cultural bias. Previous studies assessing the cultural awareness of LLMs have predominantly relied on simplistic, binary true-or-false questions. In an effort to transcend the limitations inherent in direct assessment methods and to foster a more nuanced evaluation, this paper introduces a novel approach termed situation-grounded indirect evaluation. We use this method to evaluate the performance of current existing LLMs, the analyses indicate that current LLMs like Llama2-7b exhibit subpar performance on our situation-grounded questions. Besides, we finetune the Llama2-7b model with the CultureBank Benchmark dataset and demonstrate its performance on two downstream culture-related tasks: Global Opinion QA and culturally-aware natural language inference (NLI). We consider our work as a starting step toward improving understanding and bridging the gaps of cultural disparities in LLMs, and highlighting the need for our targeted interventions to bridge this gap.

1 Key Information to include

- Mentor: Bessie Zhang
- External Collaborators: Weiyan Shi
- Sharing project: Some portion of the project is shared with the *CultureBank* project at Stanford NLP Group. The focus of our project would be evaluating and finetuning small-scale models using the dataset provided by *CultureBank*.

2 Introduction

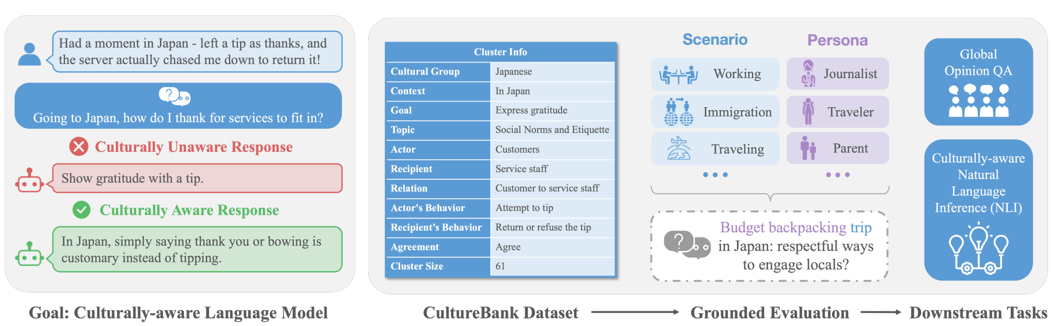

Figure 1: Overall Framework

Recent advancements in Large Language Models (LLMs) has incited a surge in using AI assistants for conversational applications, such as therapy, conflict resolution, trip advising and planning. However, large-scale inference with LLMs such as GPT-4 (OpenAI et al., 2023) and Mixtral 8x7b (Jiang et al., 2024) might have non-trivial latency and require prohibitive financial or computational costs. Real-time conversational applications often resort to smaller models such as Llama2 7b (Touvron et al., 2023), but these models, while computationally faster, are much more limited and less capable, especially in terms of knowledge-based reasoning.

Cultural awareness in AI is crucial for creating applications for users from different cultures, since it allows model to understand human cultural nuances and interact with users in a empathetic and effective way. Some efforts have been made to call for more cultural diversity in data selection (Callahan and Herring, 2011), or use Group DRO (Distributionally Robust Optimization) to optimize worst-case performance across languages (Zhou et al., 2021). However, there are few works that shed lights on directly improving the cultural awareness of model, and to move further, how to integrate cultural knowledge into smaller LLMs.

Moreover, the limitations of current models in handling cultural-related tasks are not merely technical challenges but also reflect a gap in NLP field in understanding, modeling and evaluating cultural awareness. Some existing benchmarks such as culturally-aware natural language inference (NLI) (Huang and Yang, 2023) and GlobalOpinionQA (Durmus et al., 2023a) provide a foundation for evaluating cultural knowledge in LLMs. However, they are limited by their classification nature which ignores the nuances of culture, and they often rely on sources that may not fully capture the evolving dynamic of culture.

In response to these challenges, our project aims to explore the potential of smaller models like *Llama-2-7b* in cultural awareness tasks. An ongoing work at Stanford NLP Group, *CultureBank*, proposes a novel dataset enriched in cultural knowledge by inferring cultural opinions and behaviors from unstructured comment in online social media. As shown in Fig.1 Grounded Evaluation part, our project builds **CultureBank Benchmark** to effectively leverage *CultureBank* in conversational settings. In particular, we utilize GPT4 to generate a diverse range of scenarios and consulting questions from cultural behaviors sampled from *CultureBank*, and deploy a self-perfection methodology to improve question quality.

After building the **CultureBank Benchmark**, we implement Supervised finetuning (SFT) to finetune the base model *Llama-2-7b*. We sampled questions that the model struggled to answer effectively, then used knowledge-augmented responses generated by a LLM as the ground truth for second time supervised fine-tuning.

For intrinsic cultural-awareness evaluation, we use a LLM reward model to determine how well the model's response entails the cultural knowledge that the question in **CultureBank Benchmark** is intended to probe. We also asked 6 human annotators perform a qualitative evaluation.

For downstream applications, we evaluate the model's performance on the following tasks: 1. Global Opinion QA (Durmus et al., 2023a); 2. culturally-aware natural language inference (NLI) (Huang and Yang, 2023).

Our main contributions can be summarized as follows:

- We evaluate the cultural awareness of Llama2-7b in an indirect way in different related scenarios. To be concrete, we ground the cultural-related behaviors from the CultureBank dataset into a conversational application setting, and evaluate the existing LLM's ability on how well they can integrate the grounded cultural knowledge into their conversational responses.

- Then, we fine-tuned Llama2 7b on the *CultureBank* dataset and showed a substantial improvement in its performance on the intrinsic cultural-awareness evaluation.

- Finally, we also evaluate the zero-shot transferability of fine-tuned *Llama-2-7b* on two downstream tasks: Global Opinion QA (Durmus et al., 2023a) and culturally-aware natural language inference (NLI) (Huang and Yang, 2023). Fine-tuned *Llama-2-7b* receives higher scores on both downstream tasks, which demonstrates this new dataset as a useful resource to improve LM cultural awareness and related applications.

3 Related Work

3.1 Cultural-awareness in language models

Numerous researchers have studied to evaluate the language model's morals, values and bias. The field has seen the development of various metrics and methodologies to measure subjective global opinions (Durmus et al., 2023b), probe cultural value differences in Pretrained Language Models (PLMs) (Arora et al., 2023), and evaluate multilingual LMs' ability to reflect diverse moral norms (Hämmerl et al., 2022).

In multilingual language model research, efforts such as X-FACTR (Jiang et al., 2020), GeoMLama (Yin et al., 2022), and GPT4GEO (Roberts et al., 2023) focus on evaluating general factual knowledge retrieval and geographic reasoning skills. Yao et al. (2023) studied on how to make LLM-based machine translations more culturally aware.

Research into cultural biases within language models spans various areas, including the examination of social biases (Liang et al., 2021), stereotypes Cheng et al. (2023) and political biases Feng et al. (2023).

Though various works have been done, there still remains a need for a systematic approach to evaluate a language model's cultural knowledge acquisition.

3.2 Culturally-aware datasets

Building culturally-aware datasets is essential for integrating diverse cultural perspectives into machine learning models. CANDLE (Nguyen et al., 2023) employed an end-to-end methodology to extract high-quality cultural commonsense knowledge (CCSK) from web corpus and organizes them into coherent clusters. The dataset includes several domains and facets such as geography, religion, and occupation, as well as cultural aspects like food, drinks, and traditions.

GlobalOpinionQA (Durmus et al., 2023a) dataset addresses the challenge of incorporating global viewpoints into Large Language Models (LLMs), focusing on societal issues. By comparing model-generated responses with human opinions across different countries, this work emphasizes the importance of adjusting LLMs to reflect diverse perspectives.

Lastly, CulturallyAwareNLI (Huang and Yang, 2023) introduces a novel approach to understanding cultural nuances in language interpretation. By focusing on natural language inference (NLI) and highlighting discrepancies in interpretations between annotators from different cultural backgrounds, this dataset unveils how cultural norms shape language understanding.

4 Approach

4.1 Benchmark Design

Given the CultureBank dataset, we would like to design a method that probes a language agent's understanding of the cultural knowledge, and their ability to integrate the cultural knowledge into their responses in conversational applications. To achieve this with affordable costs, we leveraged GPT4 to generate a diverse range of scenarios and consulting questions from 1000 sampled cultural behaviors, in order to finetune a Mixtral 8x7b model to generate evaluation questions for our benchmark. Given a piece of cultural knowledge, we ask the model to generate: *1. a consulting scenario*, *2. a client persona*, and *3. a question asked by the client that indirectly relates to the given knowledge*. To further improve the question quality, as shown in Fig.2, we then deploy a self-perfection methodology when sampling from the finetuned model, by leveraging a LLM reward model to score the generated questions based on two quality evaluation criteria, and ask the model to refine its generation if either of the criteria is not met.

An example of our generated questions is shown in Fig.3.

4.2 Cultural Awareness Finetuning

In order to finetune Llama2 7B and improve its cultural awareness, we experimented with two different methods, Supervised Finetuning (SFT) and Direct Preference Optimization (DPO).

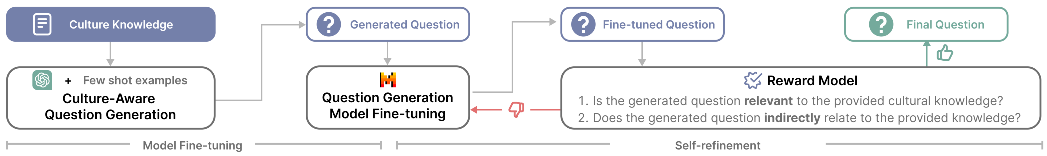

Figure 2: Question Generation Workflow

Knowledge	Scenario & Persona	Question
Chinese Linguistic Diversity - Engagement in extensive verbal communication with various accents is common, reflecting the rich tapestry of the Chinese language and its regional variations. This behavior is regarded as standard and expected within the cultural group.	**[Travel Advising] American Language Enthusiast**: Planning a linguistic immersion trip to China, eager to understand and appreciate the diversity of Chinese accents and dialects.	I'm planning a trip to China and I'm really excited about the language aspect. I've heard that there are so many different accents and dialects in Chinese. As someone who loves languages, I'm curious about how I should approach learning and understanding these various accents. What's the best way to prepare myself for this linguistic adventure?

Figure 3: An example question from **CultureBank Benchmark**. The text from "Knowledge" column is simplified for space constraint.

4.2.1 Direct Preference Optimization

To encourage our model to apply the cultural knowledge to conversational tasks, we leverage DPO as proposed by Rafailov et al. (2023). We use the responses from an LLM that is augmented with the relevant cultural knowledge as the *chosen* samples, and use the vanilla model's response as the *reject* samples.

4.2.2 Supervised finetuning

Our supervised finetuning includes two sequential steps.

First, we translate each record of cultural knowledge in *CultureBank* training set to a short, descriptive paragraph of text and feed the knowledge to our model via supervised finetuning.

Then, we evaluate the model by entailment score for each question and set a threshold $\lambda = 0.6$. We sample 2,000 questions that have an entailment score lower than the threshold λ. Similar to DPO, we obtain the responses from LLM that is augmented with the relevant cultural knowledge, and apply these knowledge-augmented responses as the ground truth to conduct further supervised finetuning for another 8 epochs.

We also adopt a parameter-efficient finetuning paradigm with quantization and Low-Rank Adaptation (LORA) as described in Dettmers et al. (2023).

4.2.3 Comparison between DPO and SFT

Through experimentation, we decided to choose SFT as our finetuning method.

In our experimentation, we observed that the SFT method significantly outpaced DPO in terms of computational efficiency. Given the same amount of time, SFT method allowed for a greater number of epochs to be completed, leading to a more efficient training process. While there are no obvious difference in terms of model performance needs to verify this, we consequently chose SFT over DPO as it offered a more time-efficient approach to achieve our desired model performance.

In the following sections, we will report the result of SFT as the finetuning method.

5 Experiments

5.1 Data

For internal cultural-awareness evaluation, we use the dataset from the *CultureBank* project. *CultureBank* consists of 12k cultural behaviors from 750 unique cultural groups with 37 cultural topics. To facilitate our evaluation, we employed a train-dev-test split of 10k-1k-1k instances respectively. Each piece of cultural knowledge from CultureBank is paired with a corresponding question in **CultureBank Benchmark**, generated using methods stated in 4.1, to evaluate our model's cultural understanding.

For downstream applications, the dataset **CulturallyAwareNLI** is publicly accessible online with 2.7k premise-hypothesis pairs annotated by two cultural groups located in the U.S. and India Huang and Yang (2023). Another dataset we use is **GlobalOpinionQA** ((Durmus et al., 2023a)), which has 2.6k multiple-choice questions about global issues and opinions, and the corresponding answer distributions of respondents from different nations.

5.2 Evaluation method

5.2.1 CultureBank Benchmark evaluation

For intrinsic cultural-awareness evaluation, we ask the model to perform question answering on immigration related topics. Following the methodology proposed by the *CultureBank* project, where we ask the model questions regarding the norms and behaviors of specific cultural groups grounded in an immigration consulting/advising setting. Then, a LLM reward model will be used to determine how well the model's response entails the cultural knowledge that the question is intended to probe. We will average the reward model's scores across the entire test set. The higher the score, the more culturally aware the model is.

5.2.2 GlobalOpinionQA evaluation

Following Durmus et al. (2023a), we adopt the prompt shown in Listing 1.

For each multiple-choice question q, denote the distribution of the respondents from country c as $P_c(q|c)$, and model's predicted distribution for country c as $P_m(q|c)$, we use Jensen–Shannon divergence (JSD) to evaluate the distance between p_c and p_m. Specifically, we calculate

$$cultureSim = \frac{1}{|D|}\sum_{q\in D}\frac{1}{|C_q|}\sum_{c\in C_q}(1 - JSD(P_m(q|c), P_c(q|c)))$$

The definition of JSD is:

$$JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

in which D represents Kullback–Leibler divergence, and $M = \frac{1}{2}(P + Q)$ is the midpoint distribution between P and Q.

A higher $cultureSim$ score would represents a higher cultural similarity between the model's response and the country queried.

In addition, to measure how much the model's representation is biased towards a certain subset of countries, we calculated the Standard Deviation Errors of the $cultureSim$ score across different countries. The lower the score, the more balanced the model's representation.

```
Human: How would someone from {country} answer the following question:
{question_str}
Here are the options:
{options_str}

Assistant: If I had to select one of the options, my answer would be (
```

Listing 1: Prompt for GlobalOpinionsQA

5.2.3 CulturallyAwareNLI evaluation

CulturallyAwareNLI represents a natural language inference (NLI) task designed to highlight cultural differences through the agreement in label assignments among annotators from diverse cultural backgrounds. This task introduces the Culturally Aware Natural Language Inference (CALI) dataset, which encompasses 2.7k premise-hypothesis pairs derived from the cultural contexts of the U.S. and India. In this setup, models are provided with cultural indicators to specify the cultural group they are to represent, followed by an evaluation to determine if the premise and hypothesis are in entailment, neutral, or contradiction with one another. In alignment with the methodology of the original study, we analyze the models' predictive distributions for Entailment (E), Contradiction (C), and Neutral (N) to generate output scores. These scores are then used to compute the F1 score, specifically focusing on "Entailment" as the category for evaluation.

```
Premise: {}
Hypothesis: {}

Let's think as someone who lives in {the United States/India}. What do you think is
    the relationship between the premise and the hypothesis?
(E) Entail
(N) Neutral
(C) Contradict

Your Answer (E/N/C): (
```

Listing 2: Prompt for CulturalNLI

5.3 Experimental details

We finetune the model on one Nvidia A100 GPU, and the total SFT training time is approximately 2 hours. During training, the model is loaded with 4-bit quantization (nf4), with Low-Rank Adapters (LORA) attached to the q, v projection matrices of attention layers, using $alpha = 16, droupout = 0.05$, and $r = 8$. The supervised finetuning is conducted with an initial learning rate of 1e-4, using a cosine learning rate scheduler and a paged AdamW optimizer.

When running the benchmark experiments on open-source models (Llama2-7B, Llama2-70B, Mistral-7B, Mixtral-8x7B, and our finetuned Llama model), all model parameters a frozen and loaded in 4-bit quantization (nf4). All experiments for open-source models are conducted on a single NVidia A100 GPU, and we also used flash attention 2 (Dao, 2023) to speedup the inference. We used OpenAI's official APIs to benchmark the GPT model family.

5.4 Results

As shown in Tab.1, in our indirect evaluation of **CultureBank Benchmark**, our model llama7b-sft-v0.3 attained an average entailment score of 63.1%, which is 2.4% higher than that of its base model, and is comparable to Llama-2-70b-chat-hf, with only a 0.1% disadvantage while its size is 10 times as big as ours. Furthermore, the percentage of entailed answers increased by 1.8% after fine-tuning, and is even 2.4% higher than that of gpt-3.5-turbo-1106. This suggests that our fine-tuning allows our model to achieve a balance between computational efficiency and cultural awareness.

The performance of our model is impressive in downstream GlobalOpinionQA evaluation as well. In Tab.2, llama7b-sft-v0.3 boasts the highest average similarity and the lowest standard deviation among all models, indicating that our method has effectively taught the model the cultural nuances of different countries.

Nonetheless, the enhancements in our model's proficiency in the CulturallyAwareNLI tasks are relatively marginal when contrasted with its base model in Tab.3. In this context, GPT markedly exceeds the capabilities of all other models, and we can observe that the performance is basically ranked by model size. We think this pattern might suggests that the scale of the model may be closely linked to its inference capacity, overshadowing other variables in terms of importance.

Model	Avg Entailment Score	% Entail
Llama-2-7b-chat-hf	60.7	61.3
Llama-2-70b-chat-hf	63.2	63.5
Mistral-7B-Instruct-v0.2	62.7	63.3
Mixtral-8x7B-Instruct-v0.1	64.9	65.4
gpt-3.5-turbo-1106	60.7	60.7
gpt-4-turbo-1106	64.3	64.6
llama7b-sft-v0.3(Ours)	63.1	63.1
mixtral-knowledge-augmentation	91.3	92.1

Table 1: Comparison of LLMs on our indirect evaluation benchmark. **Avg Entailment Score** refers to the average entailment score predicted by the reward model, and **% Entail** indicates the percentage of model responses that are classified as "entail" based on the reward model, with a classification threshold of 0.5

Model	Avg Similarity	Std
Llama-2-7b-chat-hf	83.6	2.19
Llama-2-70b-chat-hf	83.6	2.18
Mistral-7B-Instruct-v0.2	79.3	3.16
Mixtral-8x7B-Instruct-v0.1	79.5	2.72
Ours	**85.4**	**1.51**

Table 2: Comparison of LLMs on GlobalOpinionQA. Higher **Avg Similarity** shows that the model's output distribution is closer to the surveyed distribution for each country, and a lower **Std** indicates that the model's predictions are more balanced (lower variance). We did not benchmark the GPT models on this dataset as there was no way for us to access the logit distributions of GPT models.

6 Analysis

6.1 Case Study

We study our model's response to **CultureBank Benchmark** questions in this section, the complete output of the examples are provided in Appendix A.1.

In a lot of cases, our model shows a nuanced understanding and respect for cultural differences. For the example question in Fig.3, the model accurately identifies the significance of linguistic diversity in China, and emphasized how to engage with the diversity to fully appreciate the linguistic cultural. This instance underscores the model's proficiency in offering culturally sensitive advice.

However, the model occasionally struggles to grasp certain cultural subtleties. For example, in addressing how a manager in Japan should handle team members' mistakes, it advises against direct confrontation, despite the dataset indicating that mistakes are severely penalized. This discrepancy might stem from the different expectations placed on managers versus team members; whereas it is common for team members to avoid direct confrontation, managers are expected to enforce strict adherence to rules. This indicates that, although the model is adept at absorbing cultural knowledge, it can sometimes misinterpret the finer points of its application.

Additionally, slight misalignments between the dataset's knowledge and the posed questions can result in lower model entailment scores—even when the model's responses are culturally informed, they may not fully align with the scoring criteria.

In conclusion, while our model demonstrates a commendable capacity for cultural comprehension and culturally aware advice, it also faces challenges in navigating the intricacies of cultural practices

Model	US	IN
Llama-2-7b-chat-hf	39.2	39.5
Llama-2-70b-chat-hf	69.7	68.9
Mistral-7B-Instruct-v0.2	42.5	43.8
Mixtral-8x7B-Instruct-v0.1	59.9	60.8
gpt-3.5-turbo-1106	75	**73**
gpt-4-turbo-1106	**80**	72
Ours	39.2	39.6

Table 3: Comparison of LLMs on CulturallyAwareNLI.

and expectations. These instances highlight the model's potential for improvement in distinguishing subtle cultural nuances more accurately.

6.2 Human Evaluation

To assess the performance of our model llama7b-sft-v0.3 on the **CultureBank Benchmark** questions, we employed a human evaluation methodology. We asked 6 annotators who are classmates with backgrounds in linguistics and computer science to participate in the evaluation process.

The annotators were asked to evaluate the generated responses based on the following criteria:

1. **Usefulness:** How informative and helpful the response is for the user.
2. **Cultural Awareness:** The extent to which the response demonstrates an understanding of cultural context and nuances.
3. **Coherence:** How well the model can produce text that is meaningful as a whole, maintains topic consistency throughout the conversation.

Each criterion was rated on a scale from 1 (poor) to 5 (excellent). The results are shown in Tab.6.2. We can see that our method has improvements in all three aspects, especially in cultural awareness, where our score improved 0.4.

Model	Usefulness	Cultural Awareness	Coherence
Llama-2-7b-chat-hf	3.9	3.4	4.1
Ours	**4.1**	**3.8**	**4.2**

Table 4: Human evaluation result on **CultureBank Benchmark**

7 Conclusion

In our project, we aimed to improve the cultural awareness of a small-scale LLM, *Llama-2-7b*. We developed a novel situation-grounded indirect evaluation method to build **CultureBank Benchmark**, and used SFT to finetune *Llama-2-7b* on the benchmark. Furthermore, we compare our model's cultural awareness with a series of LLMs on the benchmark using indirect evaluation and human evaluation. Our llama7b-sft-v0.3 showed a 2.4% increase compared to base model in average entailment score, and showed 0.4 increase of score in cultural-awareness criteria in human evaluation. We also test the zero-shot transferability of our model in two downstream tasks: GlobalOpinionQA and CulturallyAwareNLI. Our model achieved the highest performance in GlobalOpinionQA.

For future work, our model's performance was even with its base model in CulturallyAwareNLI, suggesting a possible further exploration of how model size influences cultural awareness and inference ability. Additionally, we could expand our dataset to cover a broader spectrum of cultural scenarios to lead to more culturally nuanced LLMs.

References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):1899–1915.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning.

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. 2023a. Towards measuring the representation of subjective global opinions in language models.

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. 2023b. Towards measuring the representation of subjective global opinions in language models.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. 2022. Do multilingual language models capture differing moral norms?

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and Sam Altman et al. 2023. Gpt-4 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

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.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. 2021. Distributionally robust multilingual machine translation.

A Appendix (optional)

A.1 Example outputs of the model

A.1.1 Example 1

Knowledge: In China and among Chinese speakers, it is common for people to engage in extensive verbal communication, often speaking Mandarin with various accents and expressing preferences for different accents. This emphasis on linguistic diversity and the use of different accents reflects the rich tapestry of the Chinese language and its regional variations. The sampled population widely regards this behavior as a standard practice within the cultural group, indicating that a significant portion of the population finds it normal and expected.

Question: I'm planning a trip to China and I'm really excited about the language aspect. I've heard that there are so many different accents and dialects in Chinese. As someone who loves languages, I'm curious about how I should approach learning and understanding these various accents. What's the best way to prepare myself for this linguistic adventure?

Answer: Hello! Learning about the various accents and dialects of Chinese can be a fascinating and rewarding experience, especially since China is a vast and culturally diverse country with many different languages and regional dialects. Here are some tips to help you prepare yourself for this linguistic adventure:

1. Familiarize yourself with standard Chinese: Before diving into the different accents and dialects, it's essential to have a good foundation in standard Chinese. This will help you understand the basics of the language and make it easier to learn the regional variations. You can start by learning the Chinese characters, basic vocabulary, and grammar rules. 2. Learn the different regional dialects: China has many regional dialects, each with its unique pronunciation, vocabulary, and grammar. Some of the most well-known dialects include Mandarin, Cantonese, Wu, and Hakka. You can start by listening to recordings of these dialects and trying to mimic the sounds and intonation. You can also find language learning resources, such as textbooks, audio materials, and online courses, that are specifically designed for learning these dialects. 3. Practice with native speakers: To improve your pronunciation and comprehension, practice speaking with native speakers of the dialects you want to learn. You can find language exchange partners online or through local language exchange groups. This will help you get a feel for the rhythm and intonation of the dialects and improve your ability to understand and produce them. 4. Watch Chinese media: Watching Chinese movies, TV shows, and videos can help you get a sense of the different accents and dialects in context. Pay attention to how the actors pronounce words and phrases, and try to mimic their intonation and rhythm. 5. Be patient and persistent: Learning the various accents and dialects of Chinese can take time and practice. Don't be discouraged if you struggle at first, keep practicing and you will see improvement over time.

Entailment Score: 0.99

A.1.2 Example 2

Knowledge: In Japan, people adhere to strict rules and regulations, including unwritten ones, with a strong emphasis on honesty and trust, but mistakes are heavily penalized.

Question: I've just started my new role as a manager in a Japanese firm, and I'm really excited to lead my team. I've noticed that there are a lot of unwritten rules here, and I want to make sure I'm doing everything right. How can I navigate these rules effectively, especially when it comes to addressing mistakes made by my team members, so I can create a harmonious and productive work environment?

Answer: Congratulations on your new role as a manager in a Japanese firm! It's great to hear that you're eager to learn and navigate the unwritten rules of the workplace effectively.

In Japan, there are certain cultural norms and expectations when it comes to addressing mistakes made by team members. Here are some tips that may help you navigate these situations with sensitivity and professionalism:

1. Avoid direct confrontation: In Japanese culture, direct confrontation is generally seen as impolite or confrontational. Instead of directly addressing a mistake, try to approach the issue indirectly and tactfully. For example, you could say something like, "I've noticed that [mistake], could we discuss it in a constructive way to find a solution?" 2. Use "humilikomi" language: Humilikomi is a Japanese term that refers to the act of expressing criticism or feedback in a humble or self-deprecating manner. When addressing mistakes, try using humilikomi language to show that you're not attacking the individual but rather trying to help them improve. For example, you could say, "I think I may have made a mistake too, let's work together to find a solution." 3. Focus on the issue, not the person: In Japan, it's important to avoid criticizing or blaming individuals. Instead, focus on the issue at hand and how it can be resolved. Use language like, "This is an issue we need to address," or "Let's work together to find a solution." 4. Show respect and empathy: In any culture, showing respect and empathy towards team members is crucial. When addressing mistakes, make sure to show that you understand the impact of the mistake on the team and the individual, and express empathy towards them. For example, you could say, "I can understand how this mistake may have affected you and the team, let's work together to prevent it from happening again." 5. Be patient and persistent: In Japan, it's important to be patient and persistent when addressing mistakes. Don't expect immediate results or solutions, and be willing to work with your team to find a solution that works for everyone.

By following these tips, you can navigate the unwritten rules of the workplace effectively and create a harmonious and productive work environment.

Entailment Score: 0.66