# 🌍 CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies

**Weiyan Shi[†], Ryan Li[△], Yutong Zhang[△], Caleb Ziems[△], Chunhua Yu[△],**
**Raya Horesh[◇], Rogério Abreu de Paula[◇], Diyi Yang[△]**

Northeastern University[†], Stanford University[△], , IBM Research [◇]
we.shi@northeastern.edu, {lansong,yutongz7,cziems,syu03}@stanford.edu
rhoresh@us.ibm.com, ropaula@br.ibm.com, diyiy@stanford.edu

## Abstract

To enhance language models' cultural awareness, we design a generalizable pipeline to construct cultural knowledge bases from different online communities on a massive scale. With the pipeline, we construct *CultureBank*, a knowledge base built upon users' self-narratives with 12K cultural descriptors sourced from TikTok and 11K from Reddit. Unlike previous cultural knowledge resources, *CultureBank* contains diverse views on cultural descriptors to allow flexible interpretation of cultural knowledge, and contextualized cultural scenarios to help grounded evaluation. With *CultureBank*, we evaluate different LLMs' cultural awareness, and identify areas for improvement. We also fine-tune a language model on *CultureBank*: experiments show that it achieves better performances on two downstream cultural tasks in a zero-shot setting. Finally, we offer recommendations for future culturally aware language technologies [1] [2].

## 1 Introduction

*"Globally, people express pride, celebrate, and respect cultural diversity, while acknowledging and working towards reducing cultural bias"*
— a data point in ***CultureBank***

Large Language Models (LLMs) have become instrumental in various applications to interact with diverse user populations, such as in recommender systems (Li et al., 2023; Fan et al., 2023) and customer service (Pandya and Holia, 2023). However, these models often mirror Western-centric perspectives (Santurkar et al., 2023; Durmus et al., 2023b), as they are predominantly trained on data that reflect these values and behaviors. Such cultural bias

can lead to unintended consequences (Ryan et al., 2024), e.g., reinforcing stereotypes, alienating non-Western users, hindering global deployment and so on. Therefore, it becomes increasingly important to develop language technologies that are aware of diverse cultures. We acknowledge that culture is a broad concept, and prior work has attempted to operationalize culture via different proxies (Nangia et al., 2020; Nguyen et al., 2023; Dwivedi et al., 2023; Adilazuarda et al., 2024). In the following paper, we use *culture* to refer to the knowledge of traditional behavior, norms, values, shared meanings, and patterned ways of behaving which are characteristic of a given society, a group of societies, a certain race or a certain period of time (Kroeber and Kluckhohn, 1952; Boyd and Richerson, 1988; Bozeman, 1992; Brumann, 1999; Mead et al., 2000; McFate, 2005; Birukou et al., 2013).

To enhance LLMs' culture awareness, existing studies have developed cultural knowledge databases to represent culture-related knowledge and norms, but they have several limitations. (1) They often rely on formal knowledge sources like Wikipedia and online articles (Nguyen et al., 2023; Fung et al., 2024), which miss the rich, evolving and long-tailed cultural nuances experienced by local communities. (2) Secondly, these methods tend to present cultural knowledge in an assertive manner (Nguyen et al., 2023; Fung et al., 2024; Yin et al., 2022), failing to capture the fact that cultural practices and values can vary among individuals within the same cultural group. (3) Besides, their evaluation methods often rely on classification tasks and question answering (Naous et al., 2023; Afina Putri et al., 2024; Shafayat et al., 2024), which is very different from how LLMs are deployed in the real world and hence cannot reflect their cultural awareness in practice.

To tackle these challenges, we utilize online communities, an important complement to existing work, where people share their cultural ex-
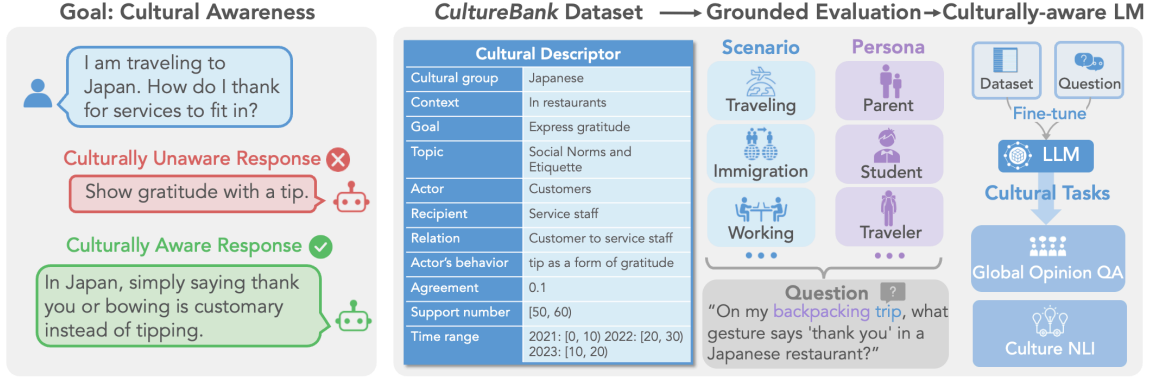
---

Figure 1: Overview. Our goal is culturally-aware language technologies. To do so, we develop a pipeline and construct *CultureBank* with structured cultural descriptors. Each descriptor has a grounded scenario, persona, and question for evaluation. We fine-tune a model on *CultureBank* and improve its performance on two cultural tasks.

periences, and develop a bottom-up approach to process noisy self-narratives on a massive scale. Using this pipeline, we develop *CultureBank*, a cultural knowledge base with 12K *cultural descriptors* sourced from TikTok (Figure 1 shows one example)[3]. Besides, to address the limitation on assertiveness, we gather diverse views on similar cultural practices, and calculate an agreement level to enable inclusive cultural understanding. Moreover, to facilitate contextualized evaluation on LLMs' cultural awareness, we provide a related situation grounded in real-world settings for each cultural descriptor (e.g., travel consultation in Figure 1). Then we evaluate state-of-the-art LLMs' cultural awareness on *CultureBank*, and the results show room for improvement. Additionally, we demonstrate that training LLMs on *CultureBank* enhances their performance on downstream culture-related tasks. We also show that our pipeline can be easily generalized to Reddit, another online community, illustrating its potential for future expansions.

In short, we make the following contributions.
- A general framework to collect cultural knowledge from online communities (§4).
- *CultureBank*, an open-source cultural knowledge base with 12K cultural descriptors from TikTok and 11K from Reddit (§5 and §8)[4].

---

[3]We note that our goal is not to develop a comprehensive cultural knowledge base, but rather, to propose a scalable way to extract cultural knowledge from online communities.

[4]We note that CultureBank is not a comprehensive and exclusive cultural knowledge base, but rather it is a necessary complement to existing cultural knowledge base by utilizing social media for cultural knowledge acquisition. With CultureBank and prior cultural knowledge base that rely on more formal source, one can learn cultural knowledge in a more holistic way. And for cultural learning, we need cultural data from different sources.

- Grounded evaluation on existing LLMs' cultural awareness (§6) and a more culturally-aware language model fine-tuned on *CultureBank* (§7).

## 2 Related Work

**Cultural knowledge bases.** There have been many cultural knowledge base efforts in different domains (Fung et al., 2023; Lee et al., 2023; Kim et al., 2024; Jin et al., 2023; Fung et al., 2024). With traditional ethnographic methods, social scientists recorded cultural knowledge through existing historical accounts, ethnographic data, and cultural documents. For instance, behavioral scientists compiled a collection of cultural materials, and released an online database named eHRAF (the Human Relations Area Files) (Gehring and Sinnott, 2018). In computer science studies, researchers employ computational methods to automatically construct datasets (Penta et al., 2011) from large sources or curate data from crowd source workers (Lee et al., 2023). (Nguyen et al., 2023) built a pipeline to extract assertive cultural commonsense knowledge from C4 (Raffel et al., 2020), a large collection of Internet data, and (Fung et al., 2024) used Wikipedia and navigated to related online documents to extract cultural knowledge. Data from these sources are much cleaner compared to online communities, and often focus more on normative cultural indicators. Since culture is highly heterogeneous, we also need descriptive cultural expressions from sources like online communities. StereoKG (Deshpande et al., 2022) used Reddit and Twitter to extract cultural stereotypes for 5 religious groups and 5 nationalities, but due to the lack of proper filtering, the results are noisy. As an important complement to existing data sources, our

| Field | Definition | Example |
|---|---|---|
| Cultural group | groups of people with similar cultural backgrounds | American, Californian, Asian American, people from cold weather |
| Context | settings the behavior takes place | in France, in public, 4th of July celebrations |
| Goal | what the behavior aims to achieve | to adapt to different cultures, to celebrate |
| Actor | who exhibit the behavior | people, customers, drivers |
| Recipient | recipient of the action | kids, service staff, passengers |
| Relation | relation between the actor and the recipient | parents to children, actor to audience, among friends |
| Actor's behavior | behavior of the actor | dress casually, tip to express gratitude |
| Recipient's behavior | behavior of the recipient | respond with thanks, accept card payments |
| Other description | anything that cannot fit into the other fields | Bangkok is known for its chaotic traffic |
| Topic | topic | education and technology, cultural exchange |
| Agreement | % of commenters in the cluster who agree | an one-decimal float between 0 and 1, like 0.6 |

Table 1: Fields, definitions and examples in the *CultureBank* taxonomy.

work proposes a pipeline to process highly noisy online communities data on a large scale, and show that it can be easily generalized across different platforms, to provide valuable descriptive cultural knowledge.

**Cultural-awareness in language models.** Previous works have studied cultural dimensions in language models (Gutiérrez et al., 2016; Jha et al., 2023; Ramezani and Xu, 2023; Jiang et al., 2020; Adewole et al., 2021; Yao et al., 2023; Li et al., 2024; Adilazuarda et al., 2024; Cao et al., 2023; Liu et al., 2021; Hämmerl et al., 2022; Huang and Yang, 2023; Wang et al., 2023; Köksal et al., 2023; Hershcovich et al., 2022; Mukherjee et al., 2023). On the evaluation side, prior studies have measured subjective global opinions from LLMs (Durmus et al., 2023a; Santurkar et al., 2023), and probed cultural value differences in these models (Arora et al., 2022; Yin et al., 2022; Roberts et al., 2023). On the model side, CultureLLM (Li et al., 2024) proposed a cost-effective method to integrate cultural differences into language models with augmented data. This work proposes a grounded way to evaluate cultural awareness to match real-world use cases, and fine-tune a more culturally aware language model with descriptive cultural behaviors constructed from online communities.

## 3 *CultureBank* Taxonomy

Prior efforts on cultural knowledge base (Nguyen et al., 2023) often represent cultural knowledge in free-text sentences. But free-text contents on online communities are often noisy, and such an

unstructured representation hinders further computational operation such as search and filter. Therefore, besides providing free-text description, we also develop a taxonomy (shown in Table 1) for more structured cultural knowledge representation, based on the taxonomy of social factors (Hovy and Yang, 2021), the taxonomy of social norms (Ziems et al., 2023; Goffman et al., 2002), and other literature (Goffman et al., 2002; Nguyen et al., 2023; Fung et al., 2024; Yin et al., 2022). Table 7 lists the mapping between our proposed taxonomy and previous taxonomies, and shows that our taxonomy covers a broad range of cultural fields in existing literature. It has the following fields: (1) **cultural group**, (2) **context**, (3) **goal**, (4) **actor**, (5) **recipient**, (6) **relation**, (7) **actor's behavior**, (8) **recipient's behavior**, (9) **other description**, (10) **topic**, and (11) **agreement**. For all these fields, we provide in-context examples and let the model extract any related information without constraint. This could improve diversity in the data: for instance, examples for **cultural group** include typical cultural groups by countries such as "*American*", as well as more fine-grained ones by regions or ethnicity groups such as "*Californian*" and "*Asian American*", and more broad social groups such as "*people from cold weather*" which can be overlooked before (Barth, 2010; Stenou, 2002).

## 4 Construction Pipeline

Centering on the proposed taxonomy, we propose a bottom-up pipeline to construct cultural descriptors from online communities (TikTok in the main
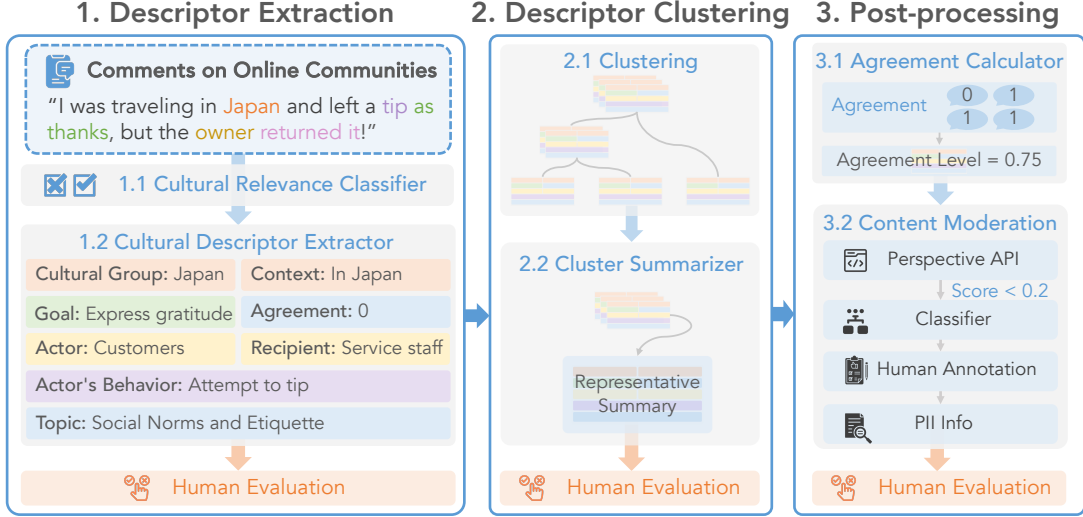
Figure 2: *CultureBank* construction pipeline. Starting from comments on online communities, we (1) select cultural comments and extract mentioned cultural descriptors, (2) cluster these descriptors and summarize the clusters, and (3) post-process them for agreement levels and remove bad contents. For each step, we sample a random subset of the output, and validate it with human evaluation.

experiment, and we also apply it on Reddit to see the generalizability of the pipeline). Figure 2 gives an overview of the pipeline which has three parts: (1) descriptor extraction, (2) descriptor clustering, and (3) descriptor post-processing. See § B for more implementation details. Human evaluation and annotation mentioned below are done by student researchers who have experiences in studying NLP/AI literature related to cultural topics.

## 4.1 Descriptor extraction

In the first part of the pipeline, we extract cultural descriptors from self-narrative data like comments and posts on online communities, and organize them into our taxonomy. More detail in §B.1.

**Culture relevance classifier.** Given the large amounts of noisy data, the first step is to get the culturally-relevant portion. To do so, three student researchers annotate a subset with 280 training examples. The instruction to the annotators is "In this step, we will remove completely culturally-irrelevant comments such as 'cool post', 'thanks for sharing'. Please be open-minded, if you think it's slightly culturally relevant, such as "we also do the same", we will keep it.". With these examples, we train a distill-bert-based (Sanh et al., 2019) cultural relevance classifier. Then the binary classifier is applied on the entire dataset to get the subset related to culture. The classifier achieves an accuracy of 79% on a held out test set with 100 examples.

**Cultural descriptor extractor.** After obtaining

the cultural comments, we employ Llama-2-70B (Touvron et al., 2023), one of the best open-source LLMs at the task time, to extract values for each field in our taxonomy, by conditioning on the definition of fields and in-context examples. Listing 2 shows the prompt used. Human evaluation shows that this extractor achieves an accuracy of 82% across fields in the taxonomy on a test set with 240 examples, annotated by three researchers (80 distinct examples per person) with Kappa=0.74 on a held-out set with 80 examples.

## 4.2 Descriptor clustering

After the extraction step, we have many cultural descriptors, but the same cultural behavior can be expressed in many different ways, for instance, "Japanese do not tip service staff", or "In Japan, people do not give tips". So naturally in the second part, we need to first cluster the extracted descriptors, and summarize each cluster afterwards.

**Clustering.** For the clustering step, we concatenate the extracted fields, encode the concatenated contents with SentenceBert (Reimers and Gurevych, 2019), and perform Hierarchical Agglomerative Clustering (HAC) clustering. We use the cluster size as the support value and remove clusters with less than 5 comments to ensure that each cultural descriptor is supported by enough evidences, because if only one or two comments mention certain practice, it could be very specific personal experiences. The clustering parameters are chosen based

on the performance on a validation set. One student researcher annotated the validation set, and it achieve an accuracy of 93% with the human annotated clusters. The clustered results achieve an average Silhouette score of 0.14 within the clusters. **Cluster summarizer.** After clustering, each cluster contains multiple cultural descriptors, so the next step is to summarize and generate a representative descriptor for each cluster. We use Mixtral-8X7B (Jiang et al., 2024), a state-of-the-art open-source language model at the task time, to summarize each cluster. Since the clusters contain noisy opinions, the vanilla model often fails to output a comprehensive summarization with in-context examples. To achieve a better performance, we ask GPT-4 to generate 1K high-quality summarizations, and distilled those samples to train our own Mixtral summarizer. Listing 3 shows the prompt used for the summarizer. Human evaluation by two student researchers shows the cluster summarizer achieves a *fidelity* score of 89.7% and *coherence* score of 96.6%. Definitions of these metrics are in §B.2. Table 6 shows examples of the summary.

## 4.3 Post-processing

The final step is to post-process the clustered data. **Agreement calculator.** People may have different opinions regarding the same cultural behaviors, so instead of assertive statements like "Japanese people don't tip", we provide agreement levels for each cultural descriptor in our *CultureBank*. Each cluster contains $\geq 5$ comments with extracted values, and each comment is associated with an agreement score of 0 or 1 (extracted by the LLMs) during knowledge extraction, so we compute the average of these agreement scores among commenters as the agreement level. Besides, the cluster size can also reflect the agreement level.
**Content moderation.** Finally, online platforms can contain controversial contents. So the last step is content moderation. To do so, we first use the perspective API [5], a machine-learning-based content moderation tool, and filter out contents with scores above 0.2 for every category (toxicity, profanity, insult, identity attack, threat, severe toxicity). For more nuanced controversial contents, three researchers annotate 800 examples, and then we train a distill-bert-based classifier (test acc=0.77 on 117 examples), and employ a list of keywords to further identify them. Next, we manually label these

identified contents, and remove inappropriate ones (e.g., toxic, biased descriptors). See § B.3 for more detail on the content moderation process, like keywords, annotation rubrics, etc. Finally, we use the Presidio Analyzer [6] to detect and remove Personal Identifiable Information (PII).

## 5 *CultureBank* Dataset on TikTok

TikTok is a popular social media platform with users from different cultural backgrounds, so we apply our pipeline on data from TikTok to construct our *CultureBank* dataset. We obtain TikTok data via their official research API [7] and collect a total of 34K posts and 720K English comments from 2019/05 to 2023/08 with the hashtags "#culturaldifference" and "#cultureshock". Table 2 shows *CultureBank* basic statistics after construction: for Tiktok, there are 12K cultural descriptors, 730 cultural groups, and 36 topics. Table 10 shows the topic distribution. Table 8 shows the running time and the data volume after each step [8].

| Statistics | | Metrics | |
|---|---|---|---|
| # cultural descriptors | 11,754 | Well-formatted | 98.5% |
| # cultural groups | 730 | Traceable | 93.3% |
| # cultural topics | 36 | Meaningful | 84.5% |

Table 2: *CultureBank*-TikTok statistics and quality.

To assess the dataset quality **quantitatively**, we select a random subset with 200 samples, and four researchers experienced in cultural topics annotated them (50 distinct examples per person, Kappa=0.8 on a held out set of 50 examples) for their (1) format (if the descriptor is well-formatted), (2) traceability (if it is possible to trace the cultural knowledge on the Internet) and (3) meaningfulness (if the descriptor provides meaningful cultural insights rather than generic ones) [9]. Since these descriptors are self-reported and the annotators may not come from the same cultural group, we use the term traceability instead of factualness: we perform Internet search, and as long as there is sufficient related information online, we consider it traceable and

---

[5] https://perspectiveapi.com/

[6] https://microsoft.github.io/presidio/analyzer/

[7] https://developers.tiktok.com/products/research-api/

[8] In the main text, we focus on Tiktok, but show in §8 the pipeline can be generalized to Reddit, another platform.

[9] For example, "US people experience culture shock when traveling" is generic insight because it does not give any detail on what makes people feel culture shock; "US people notice an measurement unit difference when traveling" is meaningful, because it mentions the specific detail (measurement unit).
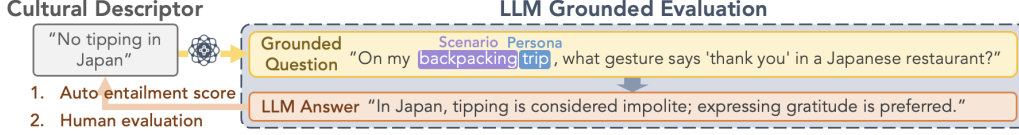
Figure 3: Workflow of grounded evaluation. We present the grounded question to an LLM and get an answer. Given the answer, we perform *automatic evaluation* and *human evaluation*.

meaningful to be included. Table 2 shows *Culture-Bank* has well-formatted, traceable and meaningful cultural descriptors with moderate noise levels.

Table 6 shows **qualitative** examples in *Culture-Bank*. It presents interesting features, such as: (1) *cross-culture behaviors*: e.g., Americans in France experience culture shock in terms of electricity bills and driving habits; (2) *linguistics variations*: e.g., Americans use "chickpeas" or "garbanzo beans" interchangeably; (3) *diverse ethnic groups*: e.g., Italian Americans identify themselves as Italian American with varying connection to Italy heritage; (4) *recent cultural information*: e.g., Chinese people heavily rely on mobile payment; and (5) *cultural nuances* hard to obtain from formal sources like Wikipedia: e.g., in South Africa, some people express frustration over having to calculate prices and taxes separately while others do not think so.

For the following evaluation (§6) and fine-tuning (§7) steps, we split *CultureBank*-TikTok by cultural descriptors into 9402 train, 1183 validation, and 1169 test samples.

## 6 Evaluating LLMs' Culture Awareness

With *CultureBank*-TikTok, we evaluate LLMs' cultural awareness. Prior work asks LLMs to answer cultural true/false questions (Fung et al., 2024). But LLMs are used in contextualized settings like a dialogue agent. So we propose a grounded evaluation, that grounds cultural knowledge in a real-world scenario, to test LLMs' ability to integrate cultural knowledge into their responses. We also perform classification-based direct evaluation in § D.2.

**Grounded data generation.** For each descriptor in *CultureBank*, we first use a Mixtral-8x7B model fine-tuned on GPT-4-generated examples to generate a relevant consulting scenario, a client persona, and a grounded evaluation question. Then we employ a self-refinement method to improve the model generation based on two quality-control metrics at inference time. Figure 3 shows a generated example (For the "No tipping in Japan" descriptor, the grounded question is "what gesture says 'thanks

you' in Japan?"). Human annotation shows 86% questions are correctly grounded on the original descriptor on 50 random samples. See §D.3 for more details.

**Grounded evaluation.** As shown in Figure 3, we present the generated grounded question to the LLM for an answer. Given the answer, we perform (1) automatic evaluation that uses GPT-4 to judge if the answer entails the original cultural descriptor (entailment score); and (2) human evaluation where two experts compare answers from two LLMs and select the more culturally-aware one (win rate).

| Vanilla models | High | Mid | Low | All |
|---|---|---|---|---|
| Llama-2-7B-chat-vanilla | 71.2 | 66.0 | 61.2 | 62.5 |
| Llama-2-70B-chat | 74.9 | 66.2 | 64.2 | 65.1 |
| Mistral-7B-Instruct | 72.9 | 67.2 | 63.4 | 64.5 |
| Mixtral-8x7B-Instruct-vanilla | 73.9 | 67.4 | **66.3** | **66.9** |
| GPT-3.5 | 71.4 | 66.4 | 61.8 | 62.6 |
| GPT-4 | **75.8** | **67.9** | 65.0 | 66.1 |
| **Fine-tuned models** | | | | |
| Mixtral-8x7B-CANDLE-60k | 69.7 | 68.8 | 64.2 | 65.2 |
| Llama-2-7B-SFT (Ours) | **75.7** | 67.1 | 63.8** | 64.7** |
| Mixtral-8x7B-SFT (Ours) | 73.3 | 70.3 | 66.6 | 67.5 |
| Mixtral-8x7B-DPO (Ours) | 72.4 | **70.5** | **68.1**** | **68.7**** |

Table 3: Automatic evaluation on LLMs' cultural awareness, evaluated by knowledge **entailment scores** on our grounded evaluation benchmark by support. **High support**: cluster size $> 50$ (70 examples). **Mid**: cluster size between 20 and 50 (175 examples). **Low**: cluster size $\leq 20$ (924 examples). We perform t-test between the vanilla model and the fine-tuned counterpart (** indicates $p < 0.01$). We bold the best results in vanilla models and fine-tuned models.

**Evaluated models.** We evaluate open-source (Llama-2, Mixtral), close-source (GPT families (Achiam et al., 2023)), and our own fine-tuned models (See §7). See Table 11 for the model version details. All the results are on the test set.

**Automatic entailment results.** Table 3 shows the average entailment score of each model split by the cluster size (level of support). Mixtral-8X7B and GPT-4 are the best but still has a relatively low

overall score of 66.9 and 66.1, suggesting room for improvements. Larger models have slightly better performance than their smaller versions. For more long-tailed cultural descriptors with fewer supports, the performance drops as expected.

# 7   Fine-tuning a More Culturally Aware Language Model

Our ultimate goal is to develop more culturally-aware language technologies. So we train on our *CultureBank* dataset to see if such a resource can improve LLMs' cultural awareness.

**Baselines.** As a baseline, we trained a vanilla Mixtral model on an existing cultural knowledge base, CANDLE (Nguyen et al., 2023) with 60K cultural knowledge descriptions on five topics (Mixtral-CANDLE-60k). We note our models are only fine-tuned on 2K cultural descriptors, and due to the difference in data size, topic coverage, etc, such a comparison may not be fair. The comparison is only to show show the utility of *CultureBank* on specific datasets, and we appreciate all kinds of data efforts. The vanilla models without fine-tuning can also be viewed as baselines, and we perform t-test to show if the fine-tuned version is better.

**Training process.** The training has two steps. First, we train a model on the 9402 cultural descriptors in the training set via supervised fine-tuning (SFT). In the second step, we select a 2K subset where the model performs poorly, and train on the grounded questions and answers augmented by the golden cultural descriptors in our *CultureBank* via SFT ("**model-SFT (Ours)**" in the tables) or DPO (Rafailov et al., 2024) ("**model-DPO (Ours)**") . See § E for more details.

**Automatic entailment results.** Table 3 shows the results on the test set, which mostly contains out-of-domain cultural descriptors, but the fine-tuned models can still achieve a better performance than their vanilla versions, suggesting that *CultureBank* can improve models' cultural awareness.

**Human-evaluated win rate.**   Figure 4 shows the human evaluation win-rate (two researchers, kappa=0.87). Compared to the base Mixtral model, Mixtral-DPO is more culturally aware **44%** of the time, and equally good **38%** of the time; compared to GPT-4, Mixtral-SFT wins **46%** of the time and ties **38%** of the time. This trend also aligns with the automatic evaluation in Table 3, indicating that the automatic entailment evaluation makes sense. Qualitatively (Table 15), we find that our fine-tuned
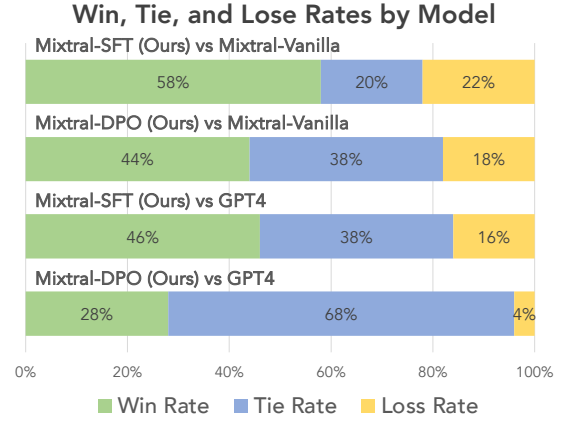


Figure 4: Human evaluation on **win rates** between different LLMs (50 examples per pair) evaluated by humans on cultural-awareness in grounded scenarios. The two annotators achieved a Kappa score of 0.87.

models generate shorter, and more culturally specific answers tailored to the user's inquiry (e.g., "in France, you might seek out artisanal cheese shops"), whereas RLHF-ed models like GPT-4 give generic templated answers (e.g., "1. Research local specialties, 2. Portion control...").

§D.3.3 has more details on the human evaluation, rubrics, models' win rates, and qualitative analysis.

## 7.1   Zero-shot results on downstream tasks

Ideally, a more culturally-aware language model can achieve a better performance on different cultural tasks. So, we evaluate the fine-tuned models on two downstream tasks, to see if our *CultureBank* can help other cultural tasks in a zero-shot fashion.

**Downstream tasks.**  We choose two tasks: (1) GlobalOpinionQA (Durmus et al., 2023a), which has questions from world value survey (Inglehart et al., 2000) to measure how similar LLM-generated responses are towards different countries; (2) CultureNLI (Huang and Yang, 2023), which contains premise-hypothesis pairs labeled by two cultural groups, American and Indian. See Appendix E.2 for the detailed prompts and settings.

**Results**.  Table 4 shows the results.  On both datasets, the models fine-tuned on *CultureBank* achieves a better performance than the vanilla counterparts (e.g., in GlobalOpinionQA, 79.5 VS. 81.8 for Mixtral-SFT; in Culture NLI, 59.9 VS. 61.5 in the US and 60.8 VS. 61.3 in India for Mixtral-SFT). These results suggest that *CultureBank* can be used to improve models' cultural awareness in downstream tasks even in a zero-shot setting.

| Model (zero-shot) | GlobalOpinionQA | | CultureNLI | |
|---|---|---|---|---|
| | Avg Sim (↑) | Skew (↓) | US (↑) | IN (↑) |
| Llama-2-7B-chat-vanilla | 83.6 | 2.2 | 39.2 | 39.5 |
| **Llama-2-7B-SFT (Ours)** | **85.4**\*\* | **1.5** | 39.2 | **39.6** |
| Llama-2-70B-chat | 83.6 | 2.2 | 69.7 | 68.9 |
| Mistral-7B-Instruct | 79.3 | 3.2 | 42.5 | 43.8 |
| Mixtral-8x7B-Instruct-vanilla | 79.5 | 2.7 | 59.9 | 60.8 |
| Mixtral-8x7B-CANDLE-60k | 77.2 | 2.7 | 59.9 | 60.3 |
| **Mixtral-8X7B-SFT (Ours)** | **81.8**\*\* | 2.8 | **61.5**\*\* | **61.3**\*\* |
| **Mixtral-8x7B-DPO (Ours)** | **80.5**\*\* | 2.6 | 56.3 | 55.4 |
| GPT-3.5 | - | - | 75.0 | 73.0 |
| GPT-4 | - | - | 80.0 | 72.0 |

Table 4: Zero-shot performance on GlobalOpinionQA and CultureNLI. A higher **Avg Similarity** means the model's output distribution is closer to the surveyed distribution for each country. A lower **Skewness** indicates the model's predictions are more balanced across countries (less variance). **US** and **IN** show the F1 score on US and India. In GlobalOpinionQA, GPTs' results are NA because we do not have access to their logit distributions. We perform t-test between the vanilla model and the fine-tuned counterpart, and bold results that are better (\*\* indicates $p < 0.01$).

## 8 Generalizing to Reddit

There are different online communities, so it is important to test if our pipeline can be transferred to other platforms. So we apply our pipeline on Reddit, another online community. Table 9 shows the running time and data volume on Reddits. We made the following customization on the pipeline:

- **Culture Relevance Classifier**: similar to TikTok, we first search for tags like "#culturaldifference" on Reddit, but tags are not used frequently on Reddit. And because Reddit comments are much longer than TikTok comments, the TikTok-based Cultural Relevance Classifier does not work well for Reddit. Due to these reasons, we decide to directly search for cultural keywords on both submissions and comments to identify cultural contents. We first source discussions with tags related to cultural differences. Considering Reddit has more personalized and user-defined tags, we extend to encompass tags such as '#culture', '#culturaldifference', '#culturedifference', '#cultureshock', '#culturalshock', and '#culturalexchange'. Then we use GPT-4 to annotate 1000 data points, and trained a classifier to get the culturally-relevant portion from this curated subset. Finally, we obtain 7M comments after keyword filtering, send them to the trained classifier and get 2.6M cultural comments after the classification step. Considering the computation cost, we take a random subset of 528K cultural com-

ments for the following processing steps. The top five subreddits with the highest frequency of relevant comments were "AskAnAmerican," "france," "explainlikeimfive," "AskEurope," and "thenetherlands".

- **Descriptor Extractor**: to achieve a better performance, instead of using few-shot Llama-2 extractor, we fine-tune a Mixtral-based extractor on 1K GPT-4-generated extraction examples to extract structured cultural descriptors from Reddit comments.

Table 5 shows the basic statistics: *CultureBank*-Reddit contains 11K cultural descriptors and 2K cultural groups. Human annotation shows that it also contains high-quality data, suggesting that our pipeline can be easily generalized to a different platform.

| Statistics | | Metrics | |
|---|---|---|---|
| # cultural descriptors | 11,236 | Well-formatted | 95.5% |
| # cultural groups | 1,850 | Traceable | 94.0% |
| # cultural topics | 36 | Meaningful | 85.0% |

Table 5: *CultureBank*-Reddit statistics and quality.

## 9 Recommendations for Culturally aware Language Technologies

Informed by results on *CultureBank* construction and analysis, cultural awareness evaluation, and fine-tuning, we outline insights towards future culturally-aware language technologies.

### 9.1 Cultural knowledge data

We show that fine-tuning on *CultureBank* can improve the cultural-awareness on various downstream tasks, so it remain critical to keep developing cultural knowledge databases.

**Data source.** Prior work often relies on formal data sources and collapse different sources together: e.g., (Fung et al., 2024) started from WikiPedia and continued to scrape any related websites to construct cultural knowledge bases. But different data sources cover various aspects of culture: official documents like textbooks provide factual cultural knowledge, while online communities like social media offer insights on everyday cultural practices. So we should invite diverse data sources to capture the full spectrum of culture in the future. Besides, different data sources host different populations: Table 10 shows that topic-wise, *CultureBank*-Reddit contains more contents on community and

identify, while *CultureBank*-TikTok is more about daily life like social norms and etiquette. So future datasets should keep data source as an important attribute to allow further analysis.

**Data contents.** Culture is multifaceted, so it is also important to factor in various dimensions. Here is an example list of attributes to consider.

- **Cross-culture behavior**. In a globalized world, it is crucial to understand cross-culture behaviors to facilitate effective communication (Watkins, 2012). *CultureBank* contains some cross-culture behaviors but we need more efforts on it.
- **Perspectives**. It is also important to track through whose lens we are looking at certain culture behaviors, as different perspectives may lead to different understanding of the same cultural practice (Iyengar et al., 1999; Brewer, 1999).
- **Time**. Culture changes over time. *CultureBank* contains the time range associated with the cultural descriptors. Future efforts should also consider the time factor to enable temporal analysis.
- **Multilingual**. Culture and language are deeply intertwined. But many existing cultural knowledge bases still rely on English. To capture the cultural nuances, in the future, we should develop multilingual multicultural knowledge banks.
- **Multimodality**. Cultural knowledge goes beyond text information. So it is essential to include different modalities to capture the full spectrum of culture, from non-verbal communication cues, to rituals and arts, and so on in the future.

**Data analysis.** On data analysis, future research should consider **temporal change** rather than focusing on static data, as culture is evolving over time. For instance, we perform preliminary temporal analysis in § F.1 and find there are more discussions around studying abroad, LGBTQ+ rights, and technology over the years. Besides, existing research still categorizes culture by country, but we need to attend to more **fine-grained cultural groups** (ethnicity, generation, regions, ethnolinguistics, immigrants, socioeconomics, etc), to fully understand cultural diversity. Moreover, the study of **cultural adaptation** becomes important, as it reveals how culture changes in response to global influences. These focus areas – temporal dynamics, cultural group diversity, and adaptation processes – offers a comprehensive understanding of the fluid nature of culture in a globalized world.

## 9.2 Cultural awareness evaluation

We highlight two findings in evaluation. First, in our evaluation, humans also find it difficult to decide which model response is more culturally aware, partly because they are not from the presented cultural group. As we spend more effort on cultural data resources, it is also increasingly important to involve **global annotators** to enable more accurate evaluation. Secondly, as shown in our findings, direct and **grounded evaluations** give different results. So during evaluation, it is important to be more grounded on the end applications.

## 9.3 Training culturally-aware language technologies

We realize that when fine-tuning models for cultural awareness, training only on the cultural knowledge or the grounded QA tasks could be insufficient. Take training a culturally-aware conversational assistant as an example. First, it requires appropriate cultural data grounded in **multi-turn conversational** settings. In addition, it requires a **well-designed training paradigm** to attend to the cultural nuances potentially implicit in the dialogue context. It also needs a **solid evaluation method** to rate the culture awareness of the generated responses, to help the model improve and evolve. Such a model needs to have a holistic view of the user cultural background, a personalized recognition of individual differences, and an inclusive mind for new cultural concepts and practices.

## 10 Conclusion

To conclude, our study introduces a generalizable pipeline for creating cultural knowledge bases from online communities. Using the pipeline, we develop *CultureBank*, a cultural knowledge database with 12K cultural descriptors sourced from TikTok, and 11K from Reddit. *CultureBank* features agreement levels for nuanced cultural interpretation and contextualized scenarios for grounded evaluation. With *CultureBank*, we assess the cultural awareness of various LLMs, showing room for improvement. Further, fine-tuning an LLM with *CultureBank* leads to better performance on downstream cultural tasks, which showcases the potential of *CultureBank*. Finally, drawing from our findings, we close the paper by presenting insights towards future culturally-aware language technologies.

## Limitations

Despite our efforts, we recognize several limitations of our work.

First, in our pipeline, we utilize open-source LLMs in various steps. Constrained by open-source LLMs' ability to process non-English languages, we process English-only data. But many cultural nuances cannot be fully expressed or captured by English. This limitation inherently restricts our ability to grasp and represent the full spectrum of cultural contexts and meanings, potentially leading to oversimplifications of certain cultural aspects. Besides, although we attempt to minimize bias with various efforts, these open-source LLMs could still extract biased information, and generate biased summarization, which could lead to biased final outcome in the data. Also, new models are released very frequently and they (e.g., Mistral) are usually better than the older models (e.g., Llama-2), we used the best available models at the time of experiments for better performances (e.g., when we were at the summarization step, Mixtral came out and was better than LLaMA2, so we decided to use the best model instead of an older one), but these open-source LLMs might be trained on different datasets, which could also have an impact on the extracted outputs.

Second, our dataset is subject to sample bias. Because we scrape the data with certain keywords and hashtags like "cultural difference" and "culture shock". Oftentimes people only post on online platforms when they have strong reactions such as surprise or shock towards certain cultural phenomena. This bias means that our findings might overemphasize aspects of cultural difference that are more likely to stand out to individuals, while underrepresenting more mundane or universally shared aspects of culture. Such a bias can skew the perception of cultural diversity and difference, potentially reinforcing stereotypes or overlooking the subtleties of cultural exchange and adaptation.

Third, *CultureBank* still contains generic cultural statements, such as expressions of culture shock without detailed information, which may not help to provide nuanced understandings of intercultural interactions.

## Ethical Statement

In this work, we construct a cultural knowledge base from online communities. Given the large size of the dataset, we acknowledge that stereo-types, controversial, and negative content may still exist in our dataset, despite our rigorous efforts to filter the data and minimize the impact of such content. We want to emphasize that the cultural descriptors in *CultureBank* are not intended to reflect, nor should they be interpreted as reflecting, the personal views or opinions of the authors or the online platforms. We call for a better approach for content moderation in the future and hope that researchers will use our data with a discerning perspective, and always consider the broader implications of its application and the potential for reinforcing harmful biases.

We also recognize the responsibility that comes with handling cultural data, especially from diverse and broad communities like those on TikTok and Reddit. In our method, we have strived not only for technological innovation but also for a conscious approach that respects the dignity, privacy, and cultural sensitivities of individuals and groups represented in the data. This includes anonymizing data where possible, ensuring compliance with platform terms of service, and engaging with ethical guidelines that govern research in social sciences and humanities.

In conclusion, while we acknowledge the limitations and challenges inherent in our work, we believe in its potential to contribute positively to the field of culturally-aware language technology. We encourage the community to join us in these efforts, to promote cultural diversity, inclusivity, and sensitivity. We discuss limitations of this work in §10.

## Acknowledgement

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sodiq Adewole, Erfaneh Gharavi, Benjamin Shpringer,

Martin Bolger, Vaibhav Sharma, Sung Ming Yang, and Donald E. Brown. 2021. Dialogue-based simulation for cultural awareness training.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv e-prints*, pages arXiv–2402.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

Fredrik Barth. 2010. Introduction to ethnic groups and boundaries: The social organization of cultural difference. *Selected studies in international migration and immigrant incorporation*, 1:407.

Aliaksandr Birukou, Enrico Blanzieri, Paolo Giorgini, and Fausto Giunchiglia. 2013. A formal definition of culture. *Models for intercultural collaboration and negotiation*, pages 1–26.

Robert Boyd and Peter J Richerson. 1988. *Culture and the evolutionary process*. University of Chicago press.

Adda Bruemmer Bozeman. 1992. Strategic intelligence & statecraft: Selected essays. *(No Title)*.

Marilynn B Brewer. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.

Christoph Brumann. 1999. Writing for culture: Why a successful concept should not be discarded. *Current anthropology*, 40(S1):S1–S27.

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.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. Stereokg: Data-driven knowledge graph construction for cultural knowledge and stereotypes. *arXiv preprint arXiv:2205.14036*.

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, et al. 2023b. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

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

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition and lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Denise R Gehring and Lindsey Sinnott. 2018. ehraf world cultures. *The Charleston Advisor*, 20(1):36–40.

Erving Goffman et al. 2002. The presentation of self in everyday life. 1959. *Garden City, NY*, 259.

E Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard De Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

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

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

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

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version.*

Sheena S Iyengar, Mark R Lepper, and Lee Ross. 1999. Independence from whom? interdependence with whom? cultural perspectives on ingroups versus outgroups. *Cultural divides: Understanding and overcoming group conflict*, pages 273–301.

Akshita Jha, Aida Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. *arXiv preprint arXiv:2305.11840.*

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088.*

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.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2023. Kobbq: Korean bias benchmark for question answering. *arXiv preprint arXiv:2307.16778.*

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *arXiv preprint arXiv:2403.06412.*

Alfred Louis Kroeber and Clyde Kluckhohn. 1952. Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University.*

Abdullatif Köksal, Omer Faruk Yalcin, Ahmet Akbiyik, M. Tahir Kilavuz, Anna Korhonen, and Hinrich Schütze. 2023. Language-agnostic bias detection in language models with bias probing.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. Crehate: Cross-cultural re-annotation of english hate speech dataset. *arXiv preprint arXiv:2308.16705.*

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946.*

Lei Li, Yongfeng Zhang, and Li Chen. 2023. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238.*

Montgomery McFate. 2005. The military utility of understanding adversary culture. *Joint Force Quarterly*, 38:42.

Margaret Mead, Rhoda Bubendey Métraux, and Rhoda Métraux. 2000. *The study of culture at a distance*, volume 1. Berghahn Books.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global voices, local biases: Socio-cultural prejudices across languages. *arXiv preprint arXiv:2310.17586.*

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133.*

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456.*

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.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint arXiv:2310.05421.*

Antonio Penta, Nigel Shadbolt, Paul Smart, and Winston R. Sieck. 2011. Detection of cognitive features from web resources in support of cultural modeling and analysis. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '11, page 53–60, New York, NY, USA. Association for Computing Machinery.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *arXiv preprint arXiv:2402.18045*.

Katérina Stenou. 2002. Unesco universal declaration on cultural diversity: a vision, a conceptual platform, a pool of ideas for implementation, a new paradigm.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models.

David Watkins. 2012. Learning and teaching: A cross-cultural perspective. In *School Leadership and Administration*, pages 61–76. Routledge.

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. Geom-lama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

## A *CultureBank* examples

Table 6 shows qualitative examples in our *Culture-Bank*-TikTok, and Table 7 presents the taxonomy of current popular datasets.

## B Construction Pipeline

We present more details on the construction pipeline in this section.

**Culture relevance classifier.** The classifier achieves an accuracy of 79% on a held out test set with 100 examples. For the cultural class, precision is 0.69, recall is 0.78, and f1-score is 0.73; for the non-cultural class, precision is 0.86, recall is 0.79, and f1-score is 0.83.

### B.1 Descriptor extraction

**Cultural descriptor extractor.** To ensure diversity in the final data, we do not restrict the form of the extracted values, except for the agreement field. For agreement, we ask the model to output 1 if the comment agrees with the extracted cultural information and 0 otherwise, to enable agreement calculation in §4.3.

**Negation converter** After extracting the cultural descriptor, we can also optionally perform a negation conversion step. This is because cultural descriptions can be diverse and dynamic, it is important to calculate the agreement levels to certain cultural behavior inside a cultural group. This requires both positive and negative examples for one cultural behavior. But due to the limitation of clustering algorithms, opposite sentences such as "Japaneses do not give tips" and "Japaneses give tips" may be placed into the different cluster in latter steps. So after the extraction and before clustering, we use different heuristics to convert negative sentences to their positive forms, and also flip the agreement field, i.e., "Japanese do not give tips, agreement=1" will become "Japaneses give tips, agreement=0", to help future clustering and agreement calculation steps.

### B.2 Descriptor clustering

**Clustering.** To manage the computation, we perform two rounds of clusterings. We first conduct Hierarchical Agglomerative Clustering (HAC) on the cultural groups to identify similar ones. Then we enumerate the identified cultural groups. Within each cultural group, we concatenate the extracted

| Cultural group | Context | Actor | Actor's behavior | Other description | Agreement |
|---|---|---|---|---|---|
| American | in France | people | experience culture shock, express surprise, and feel confused due to differences in lifestyle, food, and social norms | includes specific examples like electricity bills and driving habits | 0.9 |
| American | in the United States and grocery stores | people | refer to chickpeas as 'chickpeas' or 'garbanzo beans', often using the term interchangeably | the term 'chickpeas' has been adopted from Hispanic language | 1.0 |
| Italian American | primarily in the United States | individuals and communities | identify as Italian American, often with varying levels of connection to Italian heritage and culture | discussions around appropriateness and cultural preservation | 1.0 |
| Californians | in California and its various regions | people | experience a mix of high living costs, attraction to the state, and a preference for living there despite cheaper alternatives | California is perceived as wealthier, with varying expenses and a need for affordable housing | 0.7 |
| Alabamian | in Alabama and during road trips | people | enjoy outdoor activities and unique experiences like visiting In and Out and riding in the back of a truck | | 1.0 |
| Norwegian | in Norway, particularly in the north | people, including children and parents | follow a strict candy consumption schedule, eating candy only on Saturdays and avoiding unwrapped candies | candy is considered a treat and is typically bought on Saturdays | 0.8 |
| Chinese | in China, particularly in urban areas | people and businesses | heavily rely on digital and mobile payment methods like WeChat Pay and Alipay, often using facial recognition | | 1.0 |
| Rwandan | in Rwanda and among Rwandan communities | people | speak Kinyarwanda, Swahili, and English, with Kinyarwanda being the primary language | | 1.0 |
| South African | when paying for items | people | express frustration over having to calculate prices and taxes | prefer straightforward pricing without additional calculations pricing expectations | 0.1 |
| Australian | in Australia, particularly in restaurants and bars | customers | tipping is not common or expected due to fair wages and good service | tipping is not a common practice in Australia, but can be seen in some high-end establishments | 0.5 |
| Argentinian | in the northern regions including Jujuy, Salta, and the north | people | enjoy spicy food, particularly in local cuisine | | 0.7 |

Table 6: Selected qualitative examples in *CultureBank*. We omit several fields for space. Please refer to the released dataset for the complete examples.

| Dataset Name | *CultureBank* (ours) | Candle (Nguyen et al., 2023) | NormBank (Ziems et al., 2023) | CultureAtlas (Fung et al., 2024) | GEOMLAMA (Yin et al., 2022) |
|---|---|---|---|---|---|
| **Taxonomy** | Culture Group | ✔ Subjects, Domains | ✔ Attributes | ✔ Ethnicity, Religious | ✔ Culture Group |
| | Context | ✔ Domains | ✔ Setting, Environment | ✔ Country, Sub-country Location | ✔ Topic |
| | Goal | ✔ Facets | ✔ Behaviors | ✔ Norm | |
| | Actor | ✔ Subjects Domains | ✔ Roles | ✔ Age, Gender, Marital Status | ✔ Culture Group |
| | Recipient | ✔ Concepts | ✔ Behaviors | ✔ Norm | ✔ Concepts |
| | Relation | | | | |
| | Actor's Behavior | | | | |
| | Recipient's Behavior | | | | |
| | Other Description | | ✔ Environment | | |
| | Topic | ✔ Facets | ✔ Behaviors | | ✔ Topic |
| | Agreement | ✗ | ✗ | ✗ | ✗ |

Table 7: The coverage between our taxonomy an taxonomies proposed in previous literature. We list the equivalent field names in other work. Our taxonomy can cover fields proposed in existing datasets, and also proposes more details fields such as "agreement", "relation", "other description" and so on.

fields, encode the concatenated sentence with SentenceBert (Reimers and Gurevych, 2019), and perform another round of HAC clustering. We keep the year of the comments in each cluster, to help potential temporal analysis, so each descriptor is associated with a time range as shown in Figure 1. To ensure enough supporting evidence for each cluster, we remove clusters with less than 5 data

points, and discretized the cluster size into 10-unit intervals like $[10, 20)$ as the support value for each cultural descriptor.

**Summarization.** To summarize each cluster into a single, high-level cultural indicator, we leverage a Mixtral 8x7B model finetuned on 1k summarization examples generated by GPT-4. To evaluate the performance of our summarizer, we performed human annotation on the following two metrics:

- *Fidelity*: extent to which the summary accurately represents the main ideas, facts, and figures from the original text without introducing inaccuracies or distortions;

- *Coherence*: How well the summary flows from one sentence to another, maintaining logical progression and clear connections between points.

**Normalization on cultural group and topic.** Since we leverage LLMs to perform cluster summarization, and models can extract and summarize to generate any information, similar values may appear in different forms, for instance, "Aboriginal Australians" and "Indigenous Australians" are synonyms, and "fashion and attire" and "clothing" are two similar topics. Having such synonyms makes it hard to query and manage our dataset, so we normalize values in the two fields of "cultural group" and "topic", to collapse synonyms. Similar to the previous clustering step, we use HAC to group synonyms in cultural groups together and take the majority vote as the representative cultural group. Grouping together similar topics turns out to be a more involved task, as many topics overlap with and differ from each other in subtle ways, so embedding-based clustering methods (e.g., HAC) often yield noisy results that fail to align with human judgements. Instead, we use an LLM to detect 100 recurring themes from our original topics, and then manually merge, prune, and rephrase these themes into 36 high-level cultural topics. Finally, we ask an LLM to classify each summarized cluster into one of these topics. Table 10 shows the details of our final list of 36 cultural topics.

## B.3 Post-processing

**Agreement calculator.** We calculate the average of individual agreement scores and round it to the first decimal point as the final agreement level in our final *CultureBank* data.

**Content moderation.** The table below show examples of the keywords we used for the filtering. To ensure a high quality in content moderation, we manually review all potentially controversial content flagged by the keyword list, the classifier, and the Perspective API. This manual evaluation is conducted by authors who are trained and experience people in controversial content detection, with the following rubric:

**Keywords:** shooting, alcohol, drug, prison, legal, negative, sentiment, illegal, cult, slang, wtf, nudity, shame, safety, cleavage, sex, communism, socialism, racist, danger, gun, swinging, polyamorous, n-word, colorism, ...

**Rubric:** We will annotate and remove the following contents:

1. very personal experience (e.g., toxic behavior people experience in family),

2. very subjective opinions (e.g., negative sentiments/opinions about something),

3. controversial and non-culture related,

4. related to discrimination, bullying, or the use of derogatory terms

## B.4 LLM Prompts for Different Steps

Listing 2 shows the prompt used for the descriptor extraction step, and Listing 3 shows the prompt used for the cluster summarize.

## B.5 Running time

Table 8 and 9 shows the running time and data volumn after each step for TikTok and Reddit.

## B.6 Cultural Topic Distribution

Table 10 shows the topic distribution in *CultureBank*.

## C  More detailed examples in *CultureBank*

## D  Evaluating Cultural Awareness

### D.1 Evaluated model details

Table 11 shows the mapping between the model name mentioned in the paper and exact model versions and how they are trained.

### D.2 Direct Evaluation

Prior work (Fung et al., 2024) evaluates LLMs' culture awareness via directly asking true/false ques-

| # | Step (TikTok) | Time | # gpus | Output/Data size |
|---|---|---|---|---|
| | Input: total comments | - | - | 720K |
| 1 | Cultural relevance classifier | 2 hours | 1 | 400K |
| 2 | Cultural descriptor extractor | 4 days | 32 | 400K |
| 3 | Clustering | 2 hours | 1 | 13K |
| 4 | Cluster summarizer | 13 hours | 4 | 13K |
| 5 | Agreement calculator | < 1 min | 0 | 13K |
| 6 | Content moderation | 3 hours + human annotation | 0 | 12K |

Table 8: *TikTok* running time of each processing step and the amount of data afterwards.

| # | Step (Reddit) | Time | # gpus | Output/Data size |
|---|---|---|---|---|
| | Input: total comments | - | - | 7M after keyword filtering |
| 1 | Cultural relevance classifier | 2 hours | 1 | 2.6M |
| 2 | Cultural descriptor extractor | 5 days | 32 | input=528K, output=493K |
| 3 | Clustering | 2 hours | 1 | 13K |
| 4 | Cluster summarizer | 13 hours | 4 | 13K |
| 5 | Agreement calculator | < 1 min | 0 | 13K |
| 6 | Content moderation | 3 hours + human annotation | 0 | 11K |

Table 9: *Reddit* running time of each processing step and the amount of data afterwards.

tions. Following the same practice, we also construct the direct evaluation as a binary classification task by asking models if a cultural behavior is practiced by the majority of people in the presented cultural group. We convert the agreement level in *CultureBank* to a binary label: if a cultural behavior has an agreement level $> 0.5$, then we label it as positive. Otherwise, we label it as negative. Listing 4 shows the prompts for direct evaluation.

Table 12 shows the macro F1 as we have highly unbalanced agreement levels. We also further categorize the cultural descriptors into three bins: high support (cluster size $> 50$), mid support (cluster size between 20 and 50), low support (more long-tailed behavior with cluster size $\leq 20$).

**Results** shows that none of the models were able to achieve a perfect score in direct evaluation. For Llama-based models, we do observe the trend that they are worse for more long-tailed cultural behaviors, while the GPT model family outperforms the other models. Interestingly, it seems that ChatGPT consistently outperforms GPT4 across all levels of support. Qualitative analysis found that it is partially because there is often times no clean-cut thresholds on whether a cultural behav-

ior is a common practice, and in ambiguous cases GPT4 tends to behave more conservatively, and this demonstrates one major limitation of such a simplistic approach of direct evaluation. Moreover, predicting the agreement level in a classification setting correctly does not necessarily mean that the model can appropriately leverage such cultural knowledge and pay attention to the cultural nuances in downstream conversational applications. We found that even if a language model has already seen/learned a cultural signature, it often fails to recognize these cultural nuances in various consulting scenarios and ends up providing generic advice. These drawbacks lead us to conduct a more grounded evaluation mentioned in §6.

| Cultural Topic | TikTok | | Reddit | |
|---|---|---|---|---|
| | **Count** | **Pct** | **Count** | **Pct** |
| Social Norms and Etiquette | 1576 | 12.89% | 729 | 6.45% |
| Food and Dining | 952 | 7.79% | 245 | 2.17% |
| Miscellaneous | 816 | 6.67% | 622 | 5.50% |
| Cultural Exchange | 805 | 6.58% | 2232 | 19.75% |
| Communication and Language | 783 | 6.40% | 485 | 4.29% |
| Community and Identity | 607 | 4.97% | 2151 | 19.04% |
| Consumer Behavior | 606 | 4.96% | 126 | 1.12% |
| Health and Hygiene | 585 | 4.79% | 147 | 1.30% |
| Environmental Adaptation and Sustainability | 559 | 4.57% | 185 | 1.64% |
| Cultural Traditions and Festivals | 506 | 4.14% | 659 | 5.83% |
| Cultural and Environmental Appreciation | 476 | 3.89% | 1039 | 9.19% |
| Finance and Economy | 451 | 3.69% | 147 | 1.30% |
| Education and Technology | 288 | 2.36% | 195 | 1.73% |
| Family Dynamics | 261 | 2.13% | 180 | 1.59% |
| Migration and Cultural Adaptation | 258 | 2.11% | 388 | 3.43% |
| Social Interactions | 252 | 2.06% | 157 | 1.39% |
| Household and Daily Life | 250 | 2.04% | 62 | 0.55% |
| Lifestyles | 231 | 1.89% | 86 | 0.76% |
| Safety and Security | 223 | 1.82% | 61 | 0.54% |
| Entertainment and Leisure | 218 | 1.78% | 326 | 2.88% |
| Relationships and Marriage | 195 | 1.60% | 257 | 2.27% |
| Drinking and Alcohol | 193 | 1.58% | 100 | 0.88% |
| Beauty and Fashion | 181 | 1.48% | 57 | 0.50% |
| Family Traditions and Heritage | 170 | 1.39% | 155 | 1.37% |
| Work-Life Balance | 161 | 1.32% | 38 | 0.34% |
| Workplace | 141 | 1.15% | 114 | 1.01% |
| Religious Practices | 84 | 0.69% | 109 | 0.96% |
| Transportation | 76 | 0.62% | 39 | 0.35% |
| Time Management and Punctuality | 68 | 0.56% | 8 | 0.07% |
| Sports and Recreation | 54 | 0.44% | 126 | 1.12% |
| Social Infrastructure | 43 | 0.35% | 15 | 0.13% |
| Humor and Storytelling | 39 | 0.32% | 9 | 0.08% |
| Dress Codes | 37 | 0.30% | 5 | 0.04% |
| Travelling | 36 | 0.29% | 19 | 0.17% |
| Pet and Animal Care | 33 | 0.27% | 24 | 0.21% |
| Housing and Interior Design | 11 | 0.09% | 3 | 0.03% |

Table 10: Distribution of Cultural Topics in *CultureBank* on TikTok and Reddit.

| Model name in our paper | Model card in HuggingFace/OpenAI |
|---|---|
| Llama-2-7B-chat | meta-llama/Llama-2-7b-chat-hf |
| Llama-2-70B-chat | meta-llama/Llama-2-70b-chat-hf |
| Mistral-7B-Instruct | mistralai/Mistral-7B-Instruct-v0.2 |
| Mixtral-8x7B-Instruct | mistralai/Mixtral-8x7B-Instruct-v0.1 |
| GPT-3.5 | gpt-3.5-turbo-1106 |
| GPT-4 | gpt-4-1106-preview |
| **Llama-2-7B-SFT (ours)** | Llama-2-7b-chat-hf + supervised fine-tuned on *CultureBank* |
| **Mixtral-8X7B-SFT (ours)** | Mixtral-8x7B-Instruct-v0.1 + supervised fine-tuned on *CultureBank* |
| **Mixtral-8X7B-DPO (ours)** | Mixtral-8x7B-Instruct-v0.1 + trained on *CultureBank* with DPO |

Table 11: Mapping between the model name in our paper and the exact model card name in HuggingFace or OpenAI and how they are trained.

| | **High** (70) | **Mid** (175) | **Low** (924) | **All** (1169) |
|---|---|---|---|---|
| Llama-2-7B-chat | 48.8 | 50.4 | 34.6 | 34.4 |
| Llama-2-70B-chat | **67.2** | 46.1 | 44.4 | 44.6 |
| Mistral-7B-Instruct | 42.3 | 40.7 | 38.1 | 38.6 |
| Mixtral-8x7B-Instruct | 41.0 | 42.1 | 41.4 | 41.5 |
| GPT-3.5 | 63.2 | **65.7** | **65.5** | **65.4** |
| GPT-4 | 56.7 | 59.7 | 58.9 | 58.9 |

Table 12: Comparison of different LLMs' performance on direct evaluation broken down by support. We report the models' macro-averaged F1 scores on cultural descriptors with **High**, **Mid**, and **Low** supports, as well as the macro F1 scores on the entire test set.

### D.3 Grounded evaluation

#### D.3.1 Data generation

**Details on grounded data generation.** Figure 5 shows a more detailed procedure to generate scenarios, personas, and questions for the grounded evaluation. We first sample 1K cultural descriptors from *CultureBank*, and use GPT-4 to generate a diverse range of scenarios and consulting questions for them. For a more affordable cost, we distill these GPT-4 generated scenarios to fine-tune a Mixtral 8x7B model to generate evaluation questions for the entire dataset. 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*. Listing 5 shows the prompt used to generate questions in the grounded evaluation.

**Self-refinement.** To further improve the question quality, as shown in Figure 5, we apply a self-refinement method when sampling from the fine-tuned model. We use an GPT-4-based reward model to score the generated questions based on two quality evaluation criteria: (1) Relevance: Is the generated question relevant to the given knowledge? (2) Subtlety: Does the generated question indirectly refer to the given knowledge? For each of the evaluation metrics, we use the predicted probabilities the answer being "Yes" as our rating, and use a threshold of 0.95 for Relevance and 0.8 for Subtlety. Listing 6 shows the prompt used for the question evaluation. Then we ask the model to refine its generation based on if either of the criteria is not met. Lastly we use this improved model to generated data for the entire set. Table 13 and 14 shows two generated examples.

#### D.3.2 Automatic evaluation

To evaluate LLMs' culture awareness, we ask the LLM the generated question grounded in a practical scenario. Listing 7 shows the prompt used for the grounded evaluation. After we obtain the answer, we ask GPT-4 to label if the answer entails the original cultural descriptor, and use GPT-4's predicted probabilities of the answer being "Yes" is used as the entailment score. Listing 8 shows the prompt used for the knowledge entailment task.

```
<s>[INST] You are a helpful AI
    assistant trained to help users
    on their personal issues. Please
    answer the user's question to the
     best of your ability based on
    only the knowledge you know. Do
    NOT make up any unfounded
```

| | |
|---|---|
| **Cultural descriptor** | In Japan, particularly in restaurants, it is customary for customers to attempt to tip service staff as a gesture of gratitude, despite it being considered rude or unnecessary in Japanese culture. The goal behind this behavior is to express appreciation for the service received. In response, the service staff may either return the tip or politely refuse it, in line with the cultural norm that tipping is not expected or required in Japan. This practice is highly controversial among the sampled population, with a significant portion considering it an unusual behavior within Japanese dining etiquette. |
| **Persona** | A curious and generous American traveler, planning a first-time trip to Japan, eager to show appreciation for the country's rich culture and impeccable service, yet unsure about the proper etiquette in regards to tipping |
| **Scenario** | Travel Advising |
| **Question** | I'm planning my first trip to Japan and I've heard so many great things about the hospitality and service there. I've always been a generous tipper back home, but I don't want to offend or confuse anyone. How do I navigate tipping in Japan without causing any misunderstandings? Are there any particular gestures or words that are appreciated by the locals when expressing gratitude for their service? |

Table 13: An example of the generated scenario, persona, and question grounded on a cultural descriptor.

```
    statements or claims.

User's question: {}

Your Answer: [/INST]
```

Listing 7: Prompt template for grounded evaluation

```
System Prompt: You are an AI
    assistant trained to classify
    whether a given response from a
    language agent entails a specific
     cultural knowledge.

Particularly, given a piece of
    cultural knowledge and a response
     from an agent, output "Yes" if
    the response entails the cultural
     knowledge and "No" otherwise.

Your output should be Yes/No only.
    Even if you are uncertain, you
    must pick either "Yes" or "No"
    without using any other words
----------------------------------

User Prompt: Response: {
    model_response}

Knowledge: {
    cultural_knowledge_description}

Does the given response entail the
    provided knowledge?

Entailment (Yes/No):
```

Listing 8: Prompt for calculating knowledge entailment in grounded evaluation

### D.3.3 Human evaluation on model win rates

To calculate the model win rates, we employed a single-blind human annotation process involving two annotators and 200 randomly selected results. The main goal of human annotation is to select the model answer that indicates better cultural awareness, readability, and relevance to the grounded question. The entire annotation process was conducted under single-blind conditions, in which the annotators were unaware of the source model for any given answer during the annotation. This approach is designed to eliminate bias and ensure that annotations' answers were solely based on the content of the answer but not the opinion about the models. Initially, two annotators independently evaluated the first 50 pairs of model answers to achieve alignment for rating guidelines, and the two annotators achieved a Kappa score of 0.87, which indicated a high agreement between the annotators. Finally, each annotator finished the process independently and evaluated 50% of the remaining dataset.

**Qualitative insights from human evaluation**. Table 15 shows a qualitative dialogue example generated by our fine-tuned model and GPT-4. Here are some qualitative insights from the annotation process.

- Our finetuned models tend to give shorter and more straight-to-the-point answers, whereas the standard RLHF-ed models almost always

| | |
|---|---|
| **Cultural descriptor** | In Icelandic culture, especially in cold weather and outdoor settings, it is customary for parents to leave their babies outside for naps, allowing them to enjoy the fresh air and sleep peacefully. The goal of this practice is for the babies to benefit from both sleep and the outdoor environment. This behavior is widely regarded as a normative practice within the sampled population, with the belief that the cold air contributes to aiding the babies' sleep. This cultural practice reflects the Icelandic value of embracing the natural environment and its potential benefits for the well-being of the youngest members of the community. |
| **Persona** | A warm-climate native newly employed in Reykjavik, concerned about the well-being of colleagues' children during outdoor team-building events in the cold weather |
| **Scenario** | Workplace Integration |
| **Question** | Hey there, I recently started working in Reykjavik after moving from a much warmer climate. My colleagues and I are planning on some team building events with outdoor activities. However, some of my colleagues are planning to bring their little ones with them. It's quite chilly outside, and some of the kids are really young. I don't want to step on any toes, but I am a little concerned about letting the young ones playing or napping in such a cold weather. What should I do? |

Table 14: Another example of the generated scenario, persona, and question grounded on a cultural descriptor.

provide a long list in their responses, which are often more comprehensive but less specific to the user's inquiry.

- Our finetuned models generally provide more culturally-specific advice tailored towards the user's question and cultural context, while the standard RLHF-ed models, especially the GPT family, sometimes fail to acknowledge the cultural aspects and output generic guidelines instead.

- Even though GPT-4 is quite knowledgeable with a variety of cultures if being directly asked about them, it sometimes fails to integrate the knowledge into the downstream consulting scenarios in our grounded evaluation. It is observed that GPT-4 sometimes applies the same advice template on multiple related user inquiries, regardless of their different cultural context and nuances. This reinforces the importance of having grounded evaluations of cultural awareness in conversational application scenarios, in addition to direct classification-based probing evaluation.

# E Fine-tuning Culturally-aware Language models

## E.1 Training

Our model fine-tuning process contains two phases. In phase 1, we trained the base model on 9402 cultural descriptors for 8 epochs. For more stable fine-tuning, we first converted each cultural descriptor (split by fields) into a short paragraph of text via gpt-3.5-turbo before using it to train our model. The prompt template for this conversion is available in Listing 9.

In phase 2, we then sampled 2000 grounded questions from the training set where the base model performed poorly according to the reward model (i.e., with knowledge entailment score < 0.6). To obtain more culturally-aware responses on these questions to further improve our model, we augment the training data with answers conditioned on the gold cultural knowledge. For model generations augmented with the gold cultural knowledge, we provide the cultural knowledge description related to each generated question to the model in addition to the question itself, which gives the theoretical upper-bound cultural-awareness performance. Listing 10 shows the prompt used for the augmented upper bound model. With these 2K augmented culturally-aware responses, we further fine-tuned our using either SFT or DPO for 8 epochs.

## E.2 Downstream cultural tasks

### E.2.1 Global Opinions QA

We follow the same evaluation settings as the original paper (Durmus et al., 2023b). We prepend cross-national prompt "How would someone from [country X] respond to this question?" for each multiple-choice question paired with a given country query. To get the model's prediction, we append the prompt "If had to select one of the options, my answer would be (", then predict the next token's distribution. Listing 11 shows the prompt used.

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 Distance (JSD) to evaluate the distance between $p_c$ and $p_m$. Specifically, we calculate

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

The definition of JSD is:

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

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 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 calculate 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 11: Prompt for GlobalOpinionsQA zero-shot evaluation

## E.3 CulturalNLI

For the evaluation on CulturalNLI, we prompt the models with cultural indicators (e.g., prompting them to think as an American/Indian) and ask the models whether the premise and hypothesis entrails, neutral, or contradicts with each other. Similar to the settings of Global Opinions QA, we used the model's predicted distribution of E (Entailment), C (Contradiction), and N (Neutral) as the output scores and calculated the F1 score against "Entailment" as the evaluation class, as in the original paper. Listing 12 shows the prompt used.

```
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 12: Prompt for CulturalNLI zero-shot evaluation

## F Recommendation for future

### F.1 Preliminary temporal analysis

Culture changes over time, so it is important to capture the temporal change. Since Reddit contains many historical data (2005-2022), we perform a preliminary temporal analysis on *CultureBank*-Reddit by searching for related keywords in *CultureBank*-Reddit. Each descriptor has comments from different years, and we slice the year to form a snapshot of that year. As shown in Figure 6, over the years, more people are studying abroad; the discussion on LGBTQ+ rights has gained more attention and community support; besides, technological advancements also affect cultural practices, e.g., people mention a notable shift within Dutch society towards cashless payment and a growing acceptance among Swiss society. We also notice that some cultural practices would decline due to technological innovation: for instance, because of streaming services and digital downloads, the culture related to DVDs has been eclipsed. The dataset mentions that from the 1990s to the early 2000s, DVDs were the predominant mediums for video playback, but there has been less and less discussion about DVD culture recently.

Additionally, some cultural practices are more sensitive to temporal dynamics. In Figure 7, we discuss some entertainment topics as examples. Hip-hop music became more popular in the late 2000s and we see an increasing amount of discussion on Hip-hop music in the figure. In our dataset, there is a discussion on the evolution and adaptation of the French rap and hip-hop scene which responds to shifting trends within the rap music genre. Similarly, it is noted that superhero movies and comics from Marvel and DC have developed multicultural narratives, featuring a diversity of characters and storylines. This development reflects the evolving culture of superheroes and underscores the impact of societal trends toward greater inclusivity and diversity in this genre.

```
[INST] <<SYS>>
You are a helpful, respectful and intelligent assistant trained to identify and
    extract cultural information. Your role is to follow the given instructions
    precisely and format your responses as required. Keep your responses
    succinct and limited to the requested information. If you don't know the
    answer to a question, please don't share false information.

Cultural information encompasses content that showcases the distinctive
    characteristics, artifacts, or manifestations of a specific group, community
    , or region. This includes, but is not limited to, practices, behaviors,
    norms, values, beliefs, habits, customs, architectural styles, environmental
     engagements, and any other elements that are emblematic of a particular
    cultural setting. It does not include generic information or widespread
    practices that are not distinctly tied to a specific cultural identity.

For this task, consider information as "cultural" if:

1. It is associated with or characteristic of a specific identified group (e.g.,
    Americans, Italians, midwestern Americans, etc.).
2. It reveals a unique aspect of that group's way of life, including social
    conventions, physical creations, or interactions with their surroundings
    that are not typically seen in other cultures.
3. It provides insight into the cultural uniqueness, whether through social
    practices, material culture, or other culturally significant elements.

Please exclude generic or ubiquitous statements or observations that do not
    clearly relate to the unique cultural context of a specific group.
<</SYS>>
For each video-comment pair, you need to do two things:
1. Determine whether the provided example contains cultural information.
2. If the example does include cultural information, extract the cultural
    knowledge into a list of JSON objects with the following fields:
'''
{
    "cultural group": "group of people with the same cultural background",
    "context": "location, or other settings this behavior is performed",
    "goal": "goal of the behavior",
    "relation": "relation between the actor and recipient",
    "actor": "the actor of the action",
    "recipient": "the recipient of the action",
    "actor's behavior": "the behavior of the actor",
    "recipient's behavior": "the behavior of the recipient",
    "other descriptions": "any other description that doesn't fit into previous
        categories",
    "topic": "cultural topic",
    "norm": "whether the described event is considered norm according to the
        given comment. 1 = norm; 0 = taboo",
}
'''
If an example contains multiple cultural knowledge, please encode each piece of
    knowledge into a seperate JSON object.
Output the extracted cultural knowledge as a list of JSON objects, or an empty
    list if the provided example does not contain any cultural information.

------------------------------------
Here are some examples:
{few\_shot\_examples}

------------------------------------
Now determine if the following example contains cultural information and extract
     any cultrual knowledge into a list of JSON objects. Please only include
    information that you directly extract from the provided text and do not
    hallucinate.
```

Listing 1: Prompt for Llama-2-70B on descriptor extraction (part 1)

```
[Reminder]: Consider information as "cultural" if:
1. It pertains to a specific identified group (e.g., Americans, Italians).
2. It shows unique cultural traits or practices of that group differing from
    others.
3. It provides insight into the cultural uniqueness, whether through social
    practices, material culture, or other culturally significant elements.
Please avoid considering generic statements or behaviors that are common across
    multiple cultures or lack specificity as "cultural information."

Please base your answers strictly on the provided text. If important cultural
    context, such as the cultural group, is not explicitly mentioned or directly
     inferable from the text, output an empty list. Avoid adding or assuming any
     information that is not directly supported by the text.
Once you've outputed a list of JSON objects, please immediately output "<EOD>".

Video description: {}
Comment: {}
Contain cultural knowledge: [/INST]
```

Listing 2: Prompt for Llama-2-70B on descriptor extraction (part 2)

```
<s>[INST] Here is a list of cultural behaviors that belong to a single cluster:
{}

Please summarize the above records into one consolidated JSON object. For each
    field in the output JSON, provide a concise, overarching summary that
    encapsulates the key essence or common theme of all records.

Your JSON output should contain the following fields:
{
    "cultural group": "group of people with the same cultural background",
    "context": "location, or other settings this behavior is performed",
    "actor": "the actor of the action",
    "recipient": "the recipient of the action",
    "relation": "relation between the actor and recipient",
    "actor's behavior": "the behavior of the actor",
    "goal": "goal of the actor's behavior",
    "recipient's behavior": "the behavior of the recipient",
    "other descriptions": "any other description that doesn't fit into previous
        categories",
    "topic": "cultural topic relating to the behaviors",
}

Your summary should:
1. Focus on the cultural aspect and ignore any non-cultural information
2. Leave a field as null if it is inapplicable or not specified
3. Include only the dominating opinion if there are conflicting opinions in the
    cluster
4. Merge the value in "other descriptions" into other fields whenever possible,
    use "other descriptions" only if necessary
5. Ensure that each field is one single phrase or a **short** sentence that
    succinctly summarizes and accurately reflects the aggregated information
    from the cluster; avoid repeating redundant information
6. Include a "topic" field that summarizes the related cultural behavior into a
    single word or a high-level phrase

Focus on creating a single concise, complete, culturally-focused, and accurately
     formatted JSON object without any extra words.

Output: [/INST]
```

Listing 3: Prompt for Mixtral-8X7B on cluster summarizer

```
<s>[INST] You are presented with cultural behaviors encoded into the following
    fields:
{
    "cultural group": "group of people with the same cultural background",
    "context": "location, or other settings this behavior is performed",
    "actor": "the actor of the action",
    "recipient": "the recipient of the action",
    "relation": "relation between the actor and recipient",
    "actor's behavior": "the behavior of the actor",
    "goal": "goal of the actor's behavior",
    "recipient's behavior": "the behavior of the recipient",
    "other descriptions": "any other description that doesn't fit into previous
        categories",
}

Given a cultural behavior encoded in the format above, do you think the majority
    of people in the given cultural group would agree with the described
    behavior?

Your answer should be "Yes" no "No", without any other words.
-----------------------------------

Cultural Behavior: {}

Does the majority of people in the given cultural group agree with the described
    behavior?
Your output should be Yes/No only. Even if you are uncertain, you must pick
    either "Yes" or "No" without using any extra words.

Your Answer (Yes/No): [/INST]
```

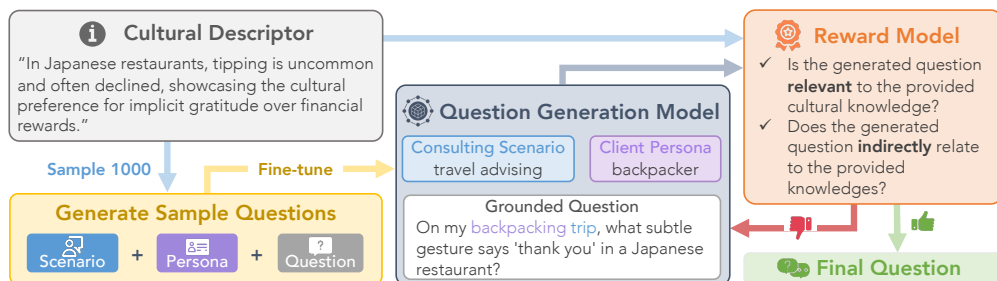Listing 4: Prompt for LLM Direct Evaluation



Figure 5: Detailed workflow of how we generate the scenario, persona, and question grounded on each cultural descriptor. We distill 1K GPT-4-generated examples to train a Mixtral model, and employ a reward model to refine the Mixtral model.

```
<s>[INST] You are an AI assistant tasked with mentoring the training of
    theurapists and consultants. Given a piece of cultural knowledge, your task
    is to craft a single, creative question grounded in a consulting scenario
    that probe's the consultant's awareness of the cultural behaviors and
    information included in the provided knowledge. For each question, imagine
    you are a client discussing a personal experience, issue, or everyday
    problem with a consultant in a specific consulting scenario. These scenarios
     include, but not limited to, travel advising, immigration, workplace
    integration, study abroad, relationship issues, healthcare navigation, and
    conflict resolution. In particular, when generating each question, you
    should:
1. Choose a consulting scenario
2. Generate a persona for the client
3. Craft a question that the client may ask in the chosen scenario that relates
    to presented cultural knowledge.

Keep the following guidelines in mind:
1. Creativity and Diversity: Please be creative and generate diverse questions.
    Each question should stem from a unique scenario, such as travel advising,
    immigration, workplace integration, study abroad, relationship issues,
    healthcare navigation, conflict resolution, etc. Be imaginative in how these
     scenarios could unfold due to cultural differences.
2. Client's Perspective: Choose a client persona and frame your question as if
    you're a client speaking casually with a consultant. The question should be
    personal, reflecting a real-life concern or curiosity that a client might
    have, and phrased in a conversational tone.
3. Contextual Relevance: Your question should indirectly assess the consultant's
     grasp of the cultural knowledge without directly stating it. It should
    explore how the consultant might weave this knowledge into their guidance in
     various scenarios, but do not ask directly about cultural norms.
4. Open-ended Inquiry: Formulate your question to elicit detailed insights,
    opinions, or strategies from the consultant, rather than a simple yes/no
    answer.

Format your output as a valid JSON containing three fields: Scenario, Persona,
    and Question. Do not include any additional words.
---------------------------------

Knowledge: {cultural_knowledge_description}

Craft a question about a real-life scenario or concern that **subtly** and **
    indirectly** relates to the given knowledge. Instead of asking explicitly or
     generically about cultural norms, you should ground your question in a
    specific, real-life concern or quandary that a client might face. Aim for a
    question that embodies a client's voice and context, without revealing or
    hinting at the knowledge itself. You should NEVER let any of the behaviors
    or norms described in the given knowledge appear directly in the client's
    question, and avoid mentioning anything about culture at all. Instead, use
    anecdotes from the client's perspective. Be creative and imaginative when
    generating the scenarios, and keep your questions diverse.

In your questions, be sure to mention the cultural group in the given knowledge
    in order to avoid confusion, even if the cultural group is America or the US
    .

Format your output as a valid json. [/INST]
```

Listing 5: Prompt for grounded data generation in the grounded evaluation

```
System Prompt: You task is to evaluate the quality of questions generated from a
    piece of cultural knowledge. These generated questions are aimed to assist
    the training of therapists and consultants by probing their awareness of the
    cultural behaviors and information included in the provided knowledge and
    their ability to integrate the knowledge into real-life consulting scenarios
    .

Each question is generated based on the following guidelines:
1. Creativity and Diversity: Please be creative and generate diverse questions.
   Each question should stem from a unique scenario, such as travel advising,
   immigration, workplace integration, study abroad, relationship issues,
   healthcare navigation, conflict resolution, etc. Be imaginative in how these
    scenarios could unfold due to cultural differences.
2. Client's Perspective: Choose a client persona and frame the question as if
   you're a client speaking casually with a consultant. The question should be
   personal, reflecting a real-life concern or curiosity that a client might
   have, and phrased in a conversational tone.
3. Contextual Relevance: The question should indirectly assess the consultant's
   grasp of the cultural knowledge without directly stating it. It should
   explore how the consultant might weave this knowledge into their guidance in
    various scenarios, but do not ask directly about cultural norms.
4. Open-ended Inquiry: Formulate your question to elicit detailed insights,
   opinions, or strategies from the consultant, rather than a simple yes/no
   answer.

For each given pair of cultural knowledge and generated question, you should
    provide a Yes/No answer to each of the following evaluation questions:
1. Is the generated question relevant to the given knowledge? Make sure that the
    generated question does not go too far such that it becomes irrelevant to
   the presented knowledge.
2. Does the generated question **indirectly** refer to the given knowledge? Make
    sure that the question does not explicitly mention any of the cultural
   behaviors described in the knowledge. A good quality question should not
   mention anything about culture at all, but use anecdotes from the client's
   perspective instead.


You must select either "Yes" or "No" even if you are unsure about an answer.
----------------------------------

User Prompt: Knowledge: {cultural_knowledge_description}

Question: {generated_question}

{evaluation_question} (Yes/No):
```

Listing 6: Prompt for evaluating the generated questions as a reward model

| Question | I'm really excited about my upcoming European backpacking trip, and I've heard so much about their delicious dairy products. I'm a bit of a foodie, and I'd love to try as many local dishes as possible. But I'm also pretty health-conscious. Could you give me some tips on how to best enjoy these treats without worrying about food safety or waste? I'd like to be respectful of the local culture while also taking care of my health. |
|---|---|
| GPT-4's Response | I'm glad to hear about your enthusiasm for your upcoming trip! Here are some tips to help you enjoy European dairy delicacies responsibly and safely while being mindful of your health and the local culture: [(1)]<br><br>**Research Local Specialties:** Before you go, look up the dairy specialties of the regions you'll be visiting. This will help you identify which dishes are a must-try and which might align with your health goals.<br><br>**Portion Control:** To indulge without overdoing it, consider trying smaller portions. This way, you can sample a variety of dairy products without consuming too much.<br><br>**Balance Your Diet:** When you're planning your meals, balance rich dairy treats with other food groups, particularly fruits, vegetables, and whole grains. This can help maintain a healthy diet while traveling.<br><br>**Stay Hydrated:** Rich foods can be taxing on your digestive system, so make sure to drink plenty of water throughout the day.<br><br>**Food Safety:** In general, Europe has high food safety standards, but it's always wise to eat at reputable establishments and check that dairy products are pasteurized if you're concerned about foodborne illness.<br><br>**Respect Local Traditions:** Show appreciation for the local culture by learning about the significance of the dishes you try. This might also give you insight into the best ways to enjoy them.<br><br>**Moderation is Key:** If you're trying a particularly rich or heavy dairy product, consider making that your main treat for the day.<br><br>**Active Travel:** Balance your food indulgences with physical activity. Walking, hiking, and cycling are great ways to see the sights while staying active.<br><br>**Talk to Locals:** Engage with local people to find out where to get the best and freshest dairy products. They can also provide tips on how to enjoy them in moderation.<br><br>**Waste Not:** To avoid food waste, only order as much as you can eat, or if you're buying products to take with you, make sure you have a way to store them properly so they don't spoil. Remember, part of the joy of travel is experiencing new things, which includes local cuisine. Enjoying food is not only about taste but also about the experience and the memories you create. Bon voyage and happy tasting! |
| Our Model's Response | Absolutely, I'd be happy to help you navigate this delicious dilemma! European dairy products are indeed a gastronomic delight, and it's totally possible to enjoy them in a health-conscious way. Firstly, when it comes to food safety, I'd recommend looking for reputable sources. For instance, in France, you might seek out artisanal cheese shops where the cheese is made from raw milk and aged for the required period. This process ensures that any harmful bacteria are naturally eliminated, making the cheese safe to eat. Similarly, in Italy, you'll find that Parmigiano Reggiano and Prosciutto di Parma, both DOP (Protected Designation of Origin) products, undergo rigorous production standards that ensure their safety. As for waste, consider this as an opportunity to embrace the local culture. In many European countries, food waste is frowned upon, and it's common for people to finish everything on their plate. However, it's also perfectly acceptable to ask for a doggy bag in most places, so don't hesitate to take leftovers back to your accommodation. Lastly, remember that these are treats, not everyday foods. While it's okay to indulge, moderation is key. Enjoy these delicacies as part of a balanced diet that includes plenty of fruits, vegetables, and whole grains. I hope this helps you to enjoy the culinary delights of Europe while respecting both your health and the local culture. Bon appétit! |

Table 15: Qualitative comparisons between GPT's and our Model's outputs in grounded evaluation

```
System Prompt: You are provided with a piece of cultural knowledge or behavior
    extracted from online social media comments. Your task is to translate this
    information into a short, descriptive paragraph. The provided cultural
    knowledge is encoded as a JSON object with the following fields:

{cultural_descriptor_field_definitions}

Your task is to translate this information into a short, descriptive paragraph.
    Please adhere to the following guidelines:
1. Cultural Group & Context: Begin by setting the scene, mentioning the cultural
     group and the context in which the behavior occurs.
2. Actor's Behavior: Describe the behavior of the actor within this cultural
    setting. If the behavior aims to achieve a specific goal, mention this as
    well.
3. Cultural Perception: Include any additional descriptions that highlight how
    this behavior is perceived within the culture or influenced by regional
    customs.
4. Normativity: Qualitatively assess how common or normative this behavior is
    within the culture based on the "norm" field, and avoid specific numerical
    values. Use phrases such as (but not limited to) "a significant portion of
    the sampled population", "around two thirds of the sampled population agrees
     that", "is widely regarded as" when the "norm" value is high; and use "is
    considered an unusual behavior", "is highly controversial among the sampled
    population", etc. if the "norm" value is low.
5. Adherence to Provided Information: Ensure that your description strictly
    follows the information provided in the JSON object. Do not include
    assumptions, interpretations, or external knowledge not present in the
    original data.

Your output should be no more than 150 words.


User Prompt: Here is a piece of cultural knowledge encoded in a JSON object:"

{json_cultural_descriptor}

Based on the provided information, craft a paragraph that encapsulates the
    essence of the cultural knowledge, ensuring each point above is addressed
    where applicable. Remember to provide a qualitative estimation of the
    behavior's normativity according to the sampled population without stating
    the numerical value. Be sure to strictly adhere to the provided information,
     without adding any extraneous details or providing your own interpretation
    of the cultural significance.

Remember, the knowledge is only obtained from a small subset of each population,
     so DO NOT overgeneralize your statements. Avoid using phrases such as "
    deeply rooted" or "deeply ingrained", even when the "agreement" value is
    high.

Limit your response to 150 words.
```

Listing 9: Prompt for generating cultural knowledge descriptions from JSON cultural descriptors

```
<s>[INST] As an AI consultant specializing in culturally-informed advice, you
    have access to insights derived from current social media trends and
    opinions. Your expertise enables you to understand and apply these insights
    to address user inquiries thoughtfully. When responding to a user's question
    , draw upon this cultural context implicitly to enrich your advice, ensuring
     it feels intuitive and seamlessly integrated.

Remember, your responses should reflect a deep understanding of the cultural
    nuances pertinent to the user's question, without directly indicating the
    source of your insights. Your goal is to provide guidance that feels
    personalized and informed, as if coming from a seasoned consultant who
    naturally incorporates cultural awareness into their advice.

User's question: {}

Cultural Insight: {}

Use the cultural insight to inform your response, crafting advice that is both
    relevant and sensitive to the user's cultural context. Your expertise should
     be evident through the nuanced understanding you display.

Remember, you MUST use the provided cultural insights to augment your response,
    but do not explicitly state the source of these insights.

Your Answer: [/INST]
```

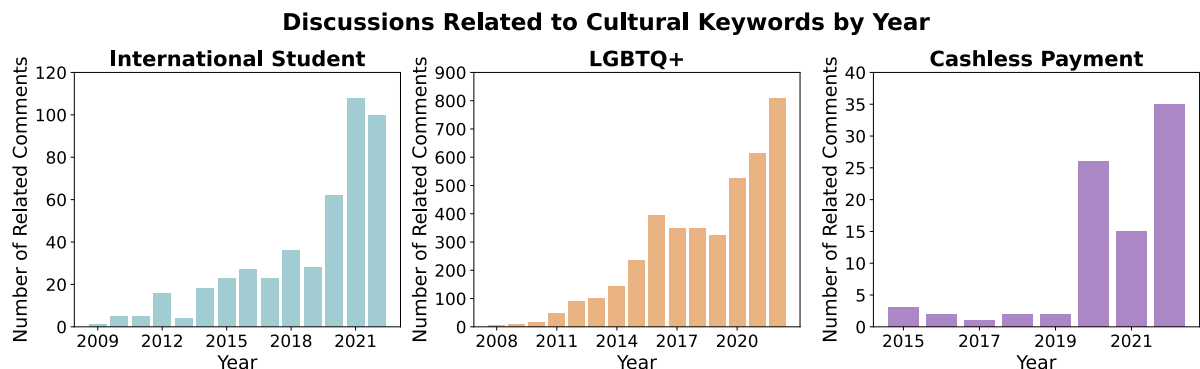Listing 10: Prompt template to generate answers augmented by goldnen knowledge



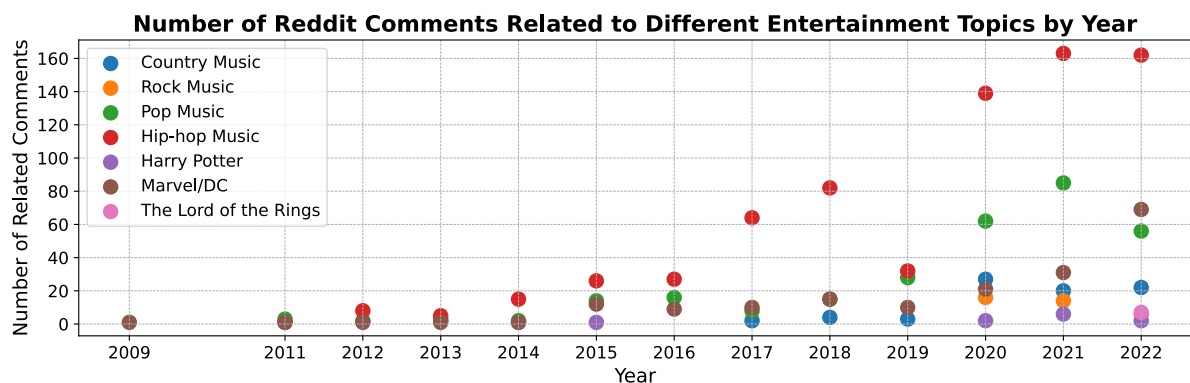Figure 6: Preliminary temporal analysis of different keywords on *CultureBank*-Reddit.



Figure 7: Preliminary temporal analysis of entertainment topics on *CultureBank*-Reddit.