

CAReDiO: Cultural Alignment of LLM via Representativeness and Distinctiveness Guided Data Optimization

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) more deeply integrate into human life across various regions, aligning them with pluralistic cultures is crucial for improving user experience and mitigating cultural conflicts. Existing approaches develop culturally aligned LLMs primarily through fine-tuning with massive carefully curated culture-specific corpora. Nevertheless, inspired by culture theories, we identify two key challenges faced by these datasets: (1) *Representativeness*: These corpora fail to fully capture the target culture’s core characteristics with redundancy, causing computation waste; (2) *Distinctiveness*: They struggle to distinguish the unique nuances of a given culture from shared patterns across other relevant ones, hindering precise cultural modeling. To handle these challenges, we introduce **CAReDiO**, a novel cultural data construction framework. Specifically, CAReDiO utilizes powerful LLMs to automatically generate cultural conversation data, where both the queries and responses are further optimized by maximizing representativeness and distinctiveness. Using CAReDiO, we construct a small yet effective dataset, covering five cultures, and compare it with several recent cultural corpora. Extensive experiments demonstrate that our method generates more effective data and enables cultural alignment with as few as 100 training samples, enhancing both performance and efficiency.

1 Introduction

As Large Language Models (LLMs) achieve unprecedented advances (Bubeck et al., 2023; OpenAI, 2024; Dubey et al., 2024; Guo et al., 2025), aligning them with human values becomes a focal point to ensure responsible development and enhance user experience (Ouyang et al., 2022; Bai et al., 2022a; Wang et al., 2024). However, existing studies primarily emphasize universal societal values like helpfulness and harmlessness (Askell et al., 2021; Bai et al., 2022a), while overlooking

the cultural pluralism rooted in human values. The globally deployed LLMs are often biased towards Western cultures (Cao et al., 2023; Durmus et al., 2023), due to English corpus’s dominance, which not only dissatisfies users from underrepresented cultural groups but also poses the risk of social conflicts (Ryan et al., 2024). Therefore, *aligning LLMs with nuanced and diverse cultural preferences is both an ethical and practical necessity.*

Early efforts on LLM cultural alignment primarily rely on in-context learning, which conditions LLMs on a target culture through role-playing instruction or native-language prompts, to evoke culturally specific responses (Durmus et al., 2023; Cao et al., 2023; Kwok et al., 2024). Nevertheless, this line of approaches requires highly capable backbone LLMs, e.g., the proprietary ones, with poor robustness and reliability across tasks, and further cause additional inference costs and privacy issues (Saunders et al., 2022). Another paradigm lies in tuning regional LLMs through computation-intensive continual pre-training (Gupta et al., 2023) on large-scale local corpora (Nguyen et al., 2023; Pipatanakul et al., 2023), while cultural alignment can not be fully achieved using the multilingual data (Choenni et al., 2024; Mukherjee et al., 2024). A more promising line is to construct dedicated cultural alignment data for a specified culture (Fung et al., 2024; Shi et al., 2024; Li et al., 2024a,b), which still demands massive and costly data.

Following this line, we ask *Can we achieve effective cultural alignment with fewer data at the least cost?* To answer this question, we investigate culture theories (Triandis et al., 1990; Miyamoto et al., 2018; Fiske and Taylor, 2020) and find that culture diversity stems from the *internal coherence* to enable its continuity and *external distinctiveness* to differentiate themselves from one another (Handwerker, 2002; Becker et al., 2012), e.g., collectivism and individualism are prioritized in Asia and Western cultures respectively (Hofstede

and Hofstede, 2005). Inspired by this, we identify two core challenges in existing cultural alignment datasets. *C1 Representativeness*: these datasets fail to accurately capture the salient constructs of the target culture, leading to irrelevant noise and redundancy, and hence hurting alignment efficiency; *C2 Distinctiveness*: current data struggles to distinguish the unique nuances of a given culture from shared patterns across multiple relevant cultures (e.g., China and Japan), hindering the precise modeling of specific cultural stimuli and preferences.

To handle these challenges, we propose CARE-DiO¹, a novel LLM-empowered framework for automatic cultural data construction, composed of two main components. The first is a cultural data synthesis pipeline, where we adapt universal value test questions to culture-specific versions and introduce Cognitive Conflict Theory (Limón, 2001) to elicit more representative and distinctive data. The second is a data selection strategy, where we quantitatively measure the extent to which each generated training sample meets the two requirements, and conduct further data filtering. In this way, CARE-DiO ensures that each data sample carries a high cultural information load and is sufficiently distinguishable from those in other cultures. Leveraging this framework, we construct a dataset with fewer yet more effective samples, named CARDSet, covering five distinct cultures. Extensive quantitative and qualitative experiments on CARDSet validate that our framework demonstrates its superiority to several recent datasets.

Our contributions are three-fold: (1) We are the first to explore *representativeness* and *distinctiveness* challenges in cultural alignment data. (2) We propose an effective data construction framework, CARE-DiO, to tackle these challenges. (3) We create the CARDSet set using our framework and manifest the effectiveness of our method, which enables alignment with as few as 100 training samples.

2 Related Work

2.1 Alignment of LLMs

To better serve humans and mitigate potential risks, aligning LLMs with human instructions, preferences and values has become increasingly essential (Shen et al., 2023; Wang et al., 2023b; Yao et al., 2023; Wang et al., 2024), with AI safety-related objectives as predominant alignment goal,

such as HHH (*helpfulness, honesty and harmlessness*) (Askell et al., 2021; Bai et al., 2022a), and various safety issues (Ganguli et al., 2022). Various alignment approaches have been investigated: RLHF (Ouyang et al., 2022; Bai et al., 2022a) and RLAIIF (Bai et al., 2022b; Lee et al., 2023) based on the PPO strategy (Schulman et al., 2017); more efficient and stable DPO algorithm (Rafailov et al., 2024) and its numerous variants (Song et al., 2024; Ethayarajh et al., 2024; Azar et al., 2024; Yuan et al., 2024). However, these efforts only emphasize universally shared societal values, overlooking the nuanced preferences across different cultures.

2.2 Cultural Alignment of LLMs

As LLMs are deployed globally, research has focused on evaluating the awareness of cultures in LLMs and proposing cultural alignment strategies.

Definition and Evaluation Culture encompasses values, social norms, interpersonal behaviors and customs, etc (Adilazuarda et al., 2024), across which various benchmarks are constructed. Many studies analyze the values embedded in LLMs using culture-related questionnaires from social sciences, including the World Value Surveys (WVS) (AlKhamissi et al., 2024), Hofstede framework (Cao et al., 2023; Masoud et al., 2023; Kharchenko et al., 2024), European Value Surveys (EVS) (Tao et al., 2024) and GlobalOpinionQA (Durmus et al., 2023). Beyond abstract values, NORMSAGE (Fung et al., 2022) and NormAd (Rao et al., 2024) assess LLMs’ adaptability to specific cultural norms. EtiCor (Dwivedi et al., 2023) tests knowledge of region-specific etiquette in domains such as dining and social interactions. Recently, more comprehensive benchmarks have emerged. CulturalBench (Chiu et al., 2024) is a multiple-choice question set curated and verified by humans. CultureBank (Shi et al., 2024) collects cultural knowledge from social platforms like Tiktok. CultureAtlas (Fung et al., 2024) compiles cultural concepts from Wikipedia, and some resources are synthesized by LLMs (Wang et al., 2023a). Many studies reveal that advanced LLMs show biases towards Western countries, underscoring the importance of cultural alignment besides evaluation.

Alignment Approaches Early explorations in LLM cultural alignment mainly focus on In-Context Learning (ICL; Dong et al., 2022). These approaches instruct LLMs to consider from a particular culture’s perspective (Durmus et al., 2023),

¹Cultural Alignment via Representativeness and Distinctiveness Optimization.

play roles with demographic details (Kwok et al., 2024; Kharchenko et al., 2024) or incorporate cultural description into prompts (Choenni and Shutova, 2024). Additionally, native language prompts show improvements under certain contexts (Durmus et al., 2023; Cao et al., 2023). However, these methods depend on the ICL capability and pre-existing cultural knowledge, less feasible for smaller LLMs (Saunders et al., 2022).

A more robust solution is tuning culture-specific LLMs with carefully crafted datasets. Various regional LLMs have been built upon English-centric models through continued pre-training on large-scale local corpus (Pires et al., 2023; Nguyen et al., 2023; Pipatanakul et al., 2023; Abbasi et al., 2023). Nonetheless, this approach is computationally expensive and cultural adaptation can not be fully achieved with only text in the native language. Recent studies seek culture-related data for cost-efficient alignment. CultureLLM (Li et al., 2024a) uses cultural responses to the World Value Survey to stimulate coherent behaviors. To collect more insightful cultural discussions, CulturePark (Li et al., 2024b) builds a multi-agent framework for cross-cultural communications. CultureSPA (Xu et al., 2024) identifies questions with shifted answers under culture-unaware and culture-aware settings.

Though data applied in existing studies is beneficial for cultural alignment, they have limitations on data representativeness and distinctiveness that are mainly optimized in our paper.

3 Method

3.1 Formalization and Overview

Culture is an important factor of human society, which typically refers to diverse aspects, *e.g.*, values, social norms, interaction manners and customs shared by a group of people. To facilitate the deployment of LLMs satisfying diverse cultural communities, this work aims to align an LLM with pluralistic cultures through fine-tuning. Given a target culture \mathcal{C} and a trainable LLM M , we fine-tune M with a cultural data collection $D_{\mathcal{C}} = \{s_1, s_2, \dots\}$ and an alignment algorithm \mathcal{F} to convert the original LLM to be a culture-specific one $M_{\mathcal{C}}$. On top of this, the fine-tuned model is expected to more effectively meet the nuanced preferences of the target culture and serve local users better.

To achieve the alignment more effectively and efficiently, a critical question raises: *what cultural data should we use for fine-tuning?* Grounded in

the internal coherence and external distinctiveness forming culture diversity as discussed in Sec. 1, we propose a novel framework, CAREDiO, to address the two key challenges: (1) *C1 Representativeness*: we should prioritize core characteristics of the target culture to the insignificant or noisy part. (2) *C2 Distinctiveness*: we should highlight the unique features of the target culture rather than general patterns shared with others, to discover more refined and effective samples for cultural alignment.

CAREDiO framework consists of two key components for automatic cultural data construction:

- **Cultural data synthesis pipeline**: this is empowered by LLMs to automatically synthesize comprehensive data that is representative to the target culture while distinctive from others.
- **Cultural data selection strategy**: we design simple metrics to quantify the two properties of each sample and filter more effective data.

The architecture of the whole framework is illustrated in Fig. 1, with each elaborated as follows.

3.2 Cultural Data Synthesis Pipeline

Considering advanced LLMs pre-trained on a large-scale web corpus have contained rich knowledge across global cultures, our proposed pipeline leverages powerful LLMs to automatically generate cultural conversation data. It includes three steps, aiming to optimize data diversity, representativeness and distinctiveness.

Comprehensive cultural framework As mentioned above, culture is a broad concept involving values, beliefs, norms and customs across various scenarios. To obtain a comprehensive training set, we first develop a cultural framework through integrating diverse definitions of cultures from multiple disciplines such as language, ethics and value. This framework comprises a total of 38 topics across four levels of various granularities: i) cultural values, ii) social norms, iii) behavioral practices and iv) specific customs. Higher levels serve as the cultural foundation across various behavior domains and contexts, while micro levels directly capture the behaviors. More details about this framework can be found in Appendix A.1.

Following this cultural framework, we employ the Self-Instruct approach (Wang et al., 2022) to synthesize k different questions for each topic. To

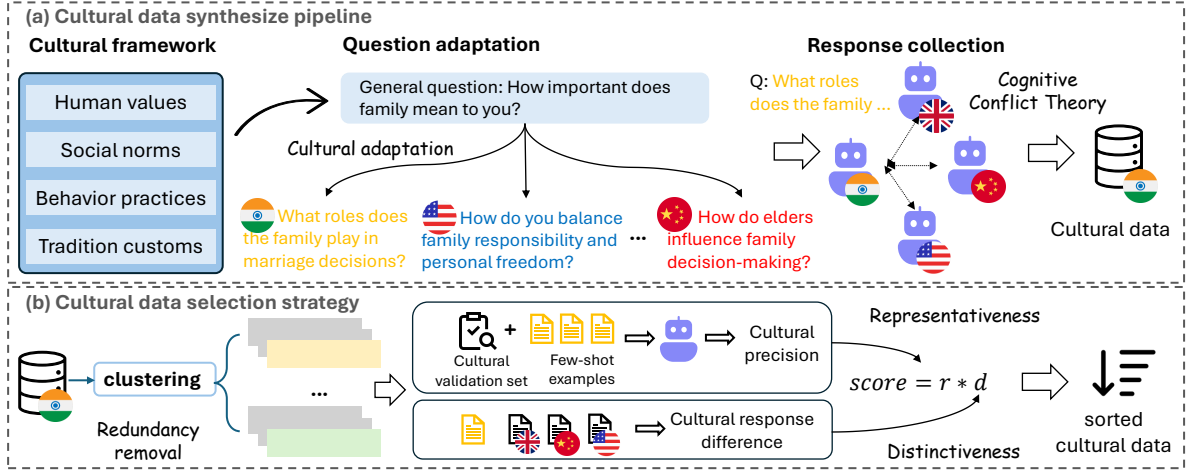


Figure 1: Architecture of the CAREDiO framework, including two modules to optimize representativeness and distinctiveness of data for cultural alignment.

ensure the quality, we instruct the LLMs to generate diverse questions that cover different aspects of each cultural topic and are highly probable to stimulate different answers in the context of different cultures. In addition, to better align with the practical application of LLMs, we consider three types that often appear in practical usage: scenarios-based questions, value-oriented questions and open-ended questions.

Cultural Question Adaptation With the cultural questions generated above, we can directly collect culture-sensitive conversations to construct data for fine-tuning. However, universal questions can be insufficient to reveal unique responses and in-depth thoughts of specific cultures, which desire cultural adaptation. Taking a general question ‘*How important does family mean to you?*’ as an example, the responses across different cultures tend to converge to emphasizing the importance of family in their life, failing to capture the distinctiveness. Whereas, ‘*how do the roles of elders influence family decision-making?*’ for collectivist cultures while ‘*how do you think about that family members should be independent from each other?*’ for individualistic cultures would be more adaptive variations to elicit more distinct perspectives.

To achieve culturally adaptive question refinement, we first leverage a powerful LLM, which is GPT-4o-mini in this paper, to generate culture-sensitive responses for each universal question using role-playing instructions. Given a universal question and responses from multiple cultures, we instruct the LLM to compare these responses, extract the reflected characteristics of each culture and

finally refine the general question to a customized version. This process is completed through chain-of-thought reasoning.

Distinctive Response Generation When generating culture-aware responses to these customized questions, we further introduce a mechanism to enrich the representativeness and distinctiveness of the synthesized data. Inspired by Cognitive Conflict Theory (CCT) (Cosier and Rose, 1977) from social science, cognitive conflicts among cultural communities in the same scenario can provoke people to reflect more on their own culture to answer questions, thus revealing cultural differences. Therefore, we first prompt the LLM to generate responses for each culture in isolation, and then require the LLM to role-play the target culture with responses obtained from several other cultures. This contrastive exposure encourages a more refined and culturally grounded response, better capturing cultural depth and nuance.

3.3 Cultural Data Selection Strategy

Our automatic data synthesis pipeline generates a great deal of cultural data highlighting representativeness and distinctiveness in both queries and responses. However, training on the entire dataset incurs high computational costs. To further improve the efficiency of cultural alignment, we prioritize the samples with the best performance on the two properties. Specifically, we propose approaches to quantify the two features.

Representativeness Measurement We perform clustering on the entire dataset to discover representative samples. We encode all samples into em-

beddings using OpenAI text-embedding-3 API and adopt the Agglomerative hierarchical clustering to merge samples with cosine similarity larger than θ . To ensure the diversity of samples for selection and reduce redundancy, we retain only one central sample per cluster, which has the largest similarity with other samples in the cluster. Then, we compute the representativeness score for each cluster center using the following two ways, denoted as r .

- **Cluster size:** We directly treat the cluster size as a proxy for representativeness. The more data share similar features with the sample, the more representative the feature is of the culture.
- **In-context performance:** With the intuition that a more representative sample can convey richer information about the culture, we leverage the in-context learning capability of the LLM and compute the representative score of each cluster center as the culture assessment score by prompting few-shot samples from that cluster. The assessment is based on small validation sets constructed with samples from CulturalBench (Chiu et al., 2024), a multiple-choice cultural benchmark with ground truth.

Distinctiveness Measurement This score measures how much the answer from the target culture distinguishes from those of other cultures for the same question. Given the response from the target culture, we randomly select answers from four other countries and compute the score as:

$$d = \frac{\sum_{i=1}^4 (1 - \text{cosine}(e, e_i))}{4}, \quad (1)$$

where e and e_i are text embeddings.

Selection Strategy With the two scores r, d for each cluster center calculated, we incorporate a new score s as their multiply for data selection, $s = r * d$. We sort all candidate samples using s and select training data one by one until we reach the pre-defined computational budget.

3.4 Cultural LLM Fine-tuning

Using our constructed dataset that has optimized representativeness and distinctiveness across various cultures, we can fine-tune cultural LLMs via SFT or DPO where the responses generated for other cultures can be regarded as dispreferred ones. To ensure a fair comparison, we follow baselines to use the SFT approaches in this paper.

4 Experimental Settings

Dataset	Types	#Samples	Metrics
GlobalOpinionQA	Questionnaire	2,556	1 - JS Distance
CulturalBench	Multiple-Choice	1,227	Accuracy
CultureBank	Open-ended	1,176	Response Quality
Prism	Open-ended	468	Response Quality

Table 1: Information of multiple benchmarks.

4.1 Datasets and Metrics

We introduce multiple benchmarks for extensive evaluations, each with distinct evaluation protocols and metrics. Statistical information are in Table 1.

• **GlobalOpinionQA** (Durmus et al., 2023): This dataset compiles 2,556 items from cross-national value questionnaires Global Attitudes surveys and World Value Survey. Each item presents an opinion-related question with multiple answer choices, along with the probability distribution of choices across various countries. For evaluation, we compute the model’s predicted probability over options and measure its similarity to the ground truth using 1 – Jensen-Shannon Distance.

• **CulturalBench** (Chiu et al., 2024): This manual dataset contains 1,227 four-choice questions for assessing LLMs’ cultural knowledge, spanning 45 regions and 17 cultural topics. We adopt its CulturalBench-Hard version which transforms each multi-choice item into four binary true/false questions and requires the LLM to evaluate all options correctly. Accuracy is calculated on the ground truth.

• **CultureBank** (Shi et al., 2024): A cultural knowledge base with self-narratives the online community TikTok across diverse scenarios such as work, immigration and travel. It generates a grounded question from each narration and splits 10% (1,176) as the testing set. Response quality is scored on a 1-to-5 scale using GPT-4o, with higher scores indicating better performance.

• **Prism** (Kirk et al., 2025): This dataset includes real conversations between 1,500 diverse participants from 75 countries and 21 LLMs. We filter a subset of questions for evaluation based on two criteria: i) the question is explicitly or potentially related to cultural topics such as relationship management and discussion on abortion; and ii) several cultures exhibit clear differences in responses. We also use GPT-4o to evaluate the culture-awareness of the responses, from 1 to 5.

Models	GlobalOpinionQA	CultureBank	Prism	CulturalBench-Hard
gpt-3.5-turbo	-	4.5331	2.1974	27.38
gpt-4o-mini	-	4.7414	2.2712	46.92
gpt-4-turbo	-	4.8128	2.1916	56.87
gpt-3.5-turbo + Role-Play	-	4.7097	3.8360	34.69
gpt-4o-mini + Role-Play	-	4.8301	4.0457	53.40
gpt-4-turbo + Role-Play	-	4.8164	3.9150	65.28
Llama-3.1-8B-Instruct as backbone model				
Llama-3.1-8B-Instruct	78.40	4.2099	2.0992	29.04
Role-Playing	79.32	4.1312	3.5298	33.42
CultureLLM	79.20	3.8104	3.5192	31.51
CulturePark	78.93	3.4018	3.5248	23.41
CultureSPA	78.16	3.7732	3.5186	32.61
CultureBank	80.99	3.7642	2.9029	9.18
CAReDiO - Cluster	<u>81.28</u>	<u>4.4292</u>	4.1200	<u>34.90</u>
CAReDiO - In context	80.97	4.2776	4.076	33.42
Qwen2.5-7B-Instruct as backbone model				
Qwen2.5-7B-Instruct	81.25	4.2568	2.2718	35.81
Role-Playing	82.20	4.2980	3.4478	33.51
CultureLLM	82.14	3.3038	3.4283	<u>43.35</u>
CulturePark	82.12	4.2796	3.3426	29.00
CultureSPA	83.89	2.7561	3.2622	32.02
CultureBank	82.70	4.2584	3.1528	34.50
CAReDiO - Cluster	84.84	<u>4.3078</u>	<u>4.0202</u>	35.50
CAReDiO - In context	85.24	4.2204	3.9278	35.15

Table 2: Overall performance for our model and baselines. The best results on each dataset are shown in bold, and those of tuning-based methods are underlined. ‘CAReDiO - Cluster/In context’ are variants with different representativeness metrics. The scores are averaged across multiple cultures, with details in Appendix C.1

4.2 Baselines

Three categories of baselines are compared.

(1) **Generally aligned LLMs:** Advanced proprietary GPT-3.5-turbo, GPT-4-Turbo and GPT-4o-mini; widely used open-source LLMs LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

(2) **LLMs with role-playing instructions:** This category uses the same backbone models as above, but incorporates a system prompt to simulate individuals from different cultural backgrounds.

(3) **Fine-tuned culture-specific LLMs:** This category is models fine-tuned using supervised learning with different culturally relevant training data, including CultureLLM (Li et al., 2024a), CulturePark (Li et al., 2024b), CultureSPA (Xu et al., 2024) and CultureBanl (Shi et al., 2024). More descriptions can be found in Appendix B.2.

4.3 Implementation Details

We utilize proprietary LLMs through official APIs, and follow the open-source code to implement other fine-tuning baselines. In Sec. 5.1, we train

LLMs with 1000 pieces of data for each culture, the same for all baselines. We experiment with 5 cultures in this paper: the United Kingdom, Chinese, South Korea, India and Singapore, which can be extended to other cultures. Using Self-Instruct, we synthesize 100 questions for each cultural topic. We cluster samples with a similarity larger than $\theta = 0.7$. Experiments are completed using NVIDIA A100 (80G). We would release the code and synthesized data for reproduction.

5 Results Analysis

5.1 Overall Performance

Table 2 presents a comprehensive comparison of cultural alignment performance between our proposed framework and various baselines. We conduct alignment across five distinct cultures and present the average score here, detailed results for each culture are shown in Appendix C.1.

A primary observation is that leveraging the cultural data synthesized through our framework

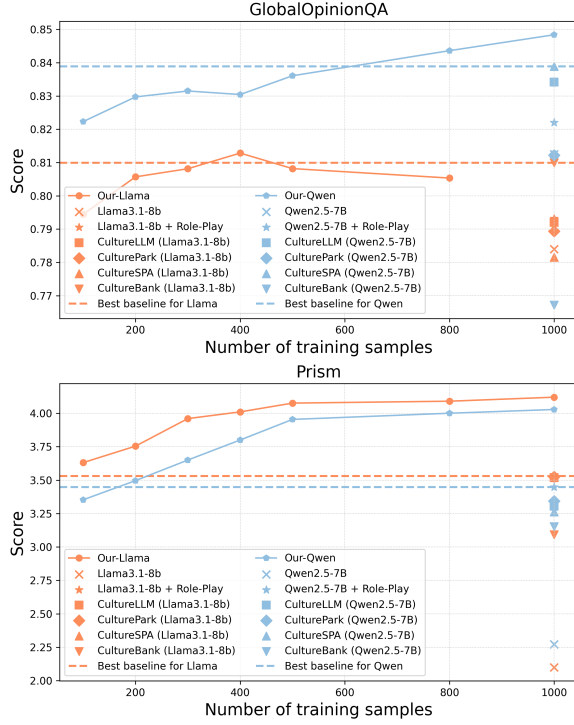


Figure 2: Results for different # of training samples.

significantly enhances cultural alignment in various LLM backbones. As shown in Table 2, this improvement consistently surpasses the effects of the simple role-playing strategy and various fine-tuning baselines across most benchmarks. Furthermore, our models demonstrate even a competitive performance with state-of-the-art LLMs such as gpt-4o-mini and gpt-4-turbo on the Prism dataset.

Notably, our models exhibit greater superiority on open-ended benchmarks like CultureBank and Prism. These benchmarks are constructed from real-world interactions or practical cultural narratives, making them a strong indicator of real-world applicability. Superior performance on these data suggests the potential of our approach to be more adapted to real-world usage scenarios.

In summary, these findings underscore the effectiveness of our cultural data synthesis framework CAREDiS optimizing representativeness and distinctiveness for cultural alignment.

5.2 Number of Training Data

To further evaluate the effectiveness of our synthesized cultural data, We conduct experiments by continuously increasing the training samples from 100 to 1,000. Figure 2 illustrates the comparison against recently proposed cultural corpora.

As more training samples are introduced, our

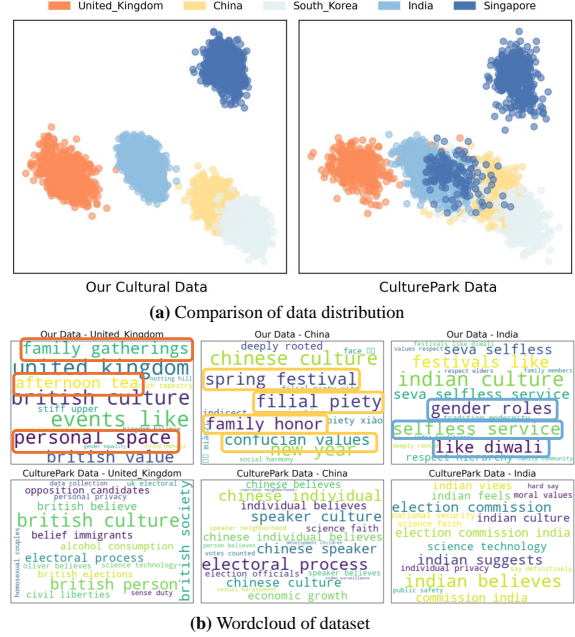


Figure 3: Distribution and word clouds of cultural data.

models continuously acquire new cultural knowledge and gain performance improvements. Moreover, the earlier selected samples contribute more significant performance gains, especially on the Prism dataset. These observations indicate that our dataset is diverse enough to continuously provide learning benefits, while prioritized samples with higher representativeness and distinctiveness are more effective for cultural alignment. This also lies the bed of the efficiency of our approach, which achieves comparable or superior performance to the best baselines using only half the training samples. On the Prism benchmark, our model reaches top performance with as few as 100 samples. This reduction in training overhead is highly valuable for fine-tuning-based methods.

5.3 Qualitative Analysis

We provide data analysis and case studies to reveal why the cultural data constructed by our framework is more effective for cultural alignment in LLMs.

Better distinctiveness & representativeness

First, we embed our synthesized data for distinct cultures and the CulturePark benchmark using OpenAI text-embedding API, then conduct PCA dimensionality reduction. As shown in Figure 3 (a), our data presents superior clustering, with data points from different cultures forming clearly distinct groups, while the CulturePark dataset has obvious distributional overlap across cultures. Even

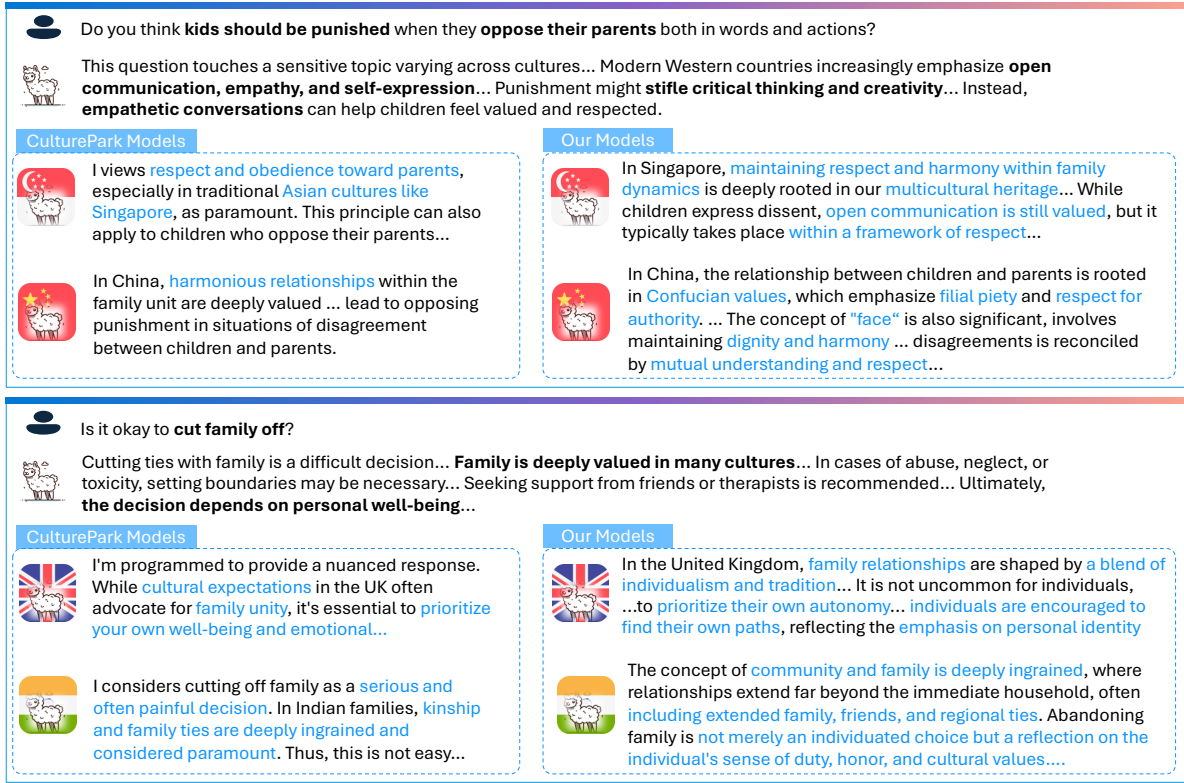


Figure 4: Case studies on cultural alignment.

for cultures that share the Asian background, our data can still capture their nuanced segmentation.

Delving into the specific data of each culture, we compute tf-idf features and visualize the most significant phrases using word clouds in Figure 3(b). Our data exactly encompasses the core concepts and unique elements of these cultures. For example, ‘afternoon tea’, ‘personal space’ reflects British etiquette and privacy norms; ‘filial piety’, ‘family honor’ and ‘spring festival’ captures Chinese tradition and collectivism. In contrast, the CulturePark data emphasizes general value-related topics such as elections, work and so on. The analysis suggests that our data captures representative samples while separating from other similar cultures.

Case Study We present case studies of culturally sensitive topics in Figure 4. Without cultural alignment, the original Llama-3.1-8B-Instruct usually returns general responses that lack cultural specificity. Due to the predominance of English-language training data, its response sometimes demonstrates a bias towards Western perspectives, which underscores the importance of cultural alignment to ensure the inclusivity of AI. We find that culture-specific models exhibit significantly improved adaptation to their respective cultural con-

texts. While models trained on CulturePark data capture coarse cultural characteristics, our models learn more comprehensive and deeper cultural details, thereby providing more appropriate responses. For example, the response in Singapore effectively reflects the cultural emphasis on respect for elders, social harmony and multicultural heritage. Similarly, the Chinese response highlights Confucian ethics and the concept of ‘face’. This qualitative analysis fully demonstrates the value of our approach for cultural alignment to enable LLMs to generate responses that align with deep-rooted cultural values, ensuring both accuracy and appropriateness in human-AI interaction.

6 Conclusion

This paper addresses the critical challenge of representativeness and distinctiveness in cultural alignment data by introducing CAREDiO, an LLM-empowered framework for automatic cultural data construction. It comprises a data synthesis pipeline and a selection strategy to construct cultural data with optimized representativeness and distinctiveness. Using the constructed dataset CARDSet covering five distinct cultures, we demonstrate the superiority of CAREDiO over several recent datasets.

7 Limitations

In this paper, we propose a novel cultural data synthesis framework to generate cultural data rich in representativeness and distinctiveness. Extensive experiments across multiple cultures have verified its effectiveness. Nevertheless, there are several limitations of our work, discussed as follows.

(1) Our synthesis framework currently relies on powerful LLMs to generate cultural data. As a result, it is unavoidably affected by the cultural bias embedded in these LLMs, and may not collect data accurately enough for low-resource cultures. However, a main contribution of our work is the optimization of representativeness and distinctiveness for cultural alignment, which we believe could be easily extended to manually curated data.

(2) Due to constraints in computational and API resources, our experiments currently cover only five distinct cultures from various regions. Given the vast diversity and complexity of global cultures, we should consider the alignment of more cultures in the future.

(3) Emphasizing the representativeness of cultural data for alignment might overlook some long-tail or emerging practices. Cultures are dynamic and constantly evolving, and our current method may not fully capture these changes.

(4) Currently, we follow baselines to use supervised fine-tuning. But it is easy to collect dispreferred responses in the context of cultural alignment. Thus, we can explore more effective fine-tuning techniques.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, 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*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas

Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Maja Becker, Vivian L Vignoles, Ellinor Owe, Rupert Brown, Peter B Smith, Matt Easterbrook, Ginette Herman, Isabelle De Sauvage, David Bourguignon, Ana Torres, et al. 2012. Culture and the distinctiveness motive: constructing identity in individualistic and collectivistic contexts. *Journal of personality and social psychology*, 102(4):833.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrkke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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. *arXiv preprint arXiv:2303.17466*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during lm fine-tuning. *arXiv preprint arXiv:2405.12744*.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*.

689	Richard A Cosier and Gerald L Rose. 1977. Cognitive	W Penn Handwerker. 2002. The construct validity of	742
690	conflict and goal conflict effects on task performance.	cultures: Cultural diversity, culture theory, and a	743
691	<i>Organizational behavior and human performance</i> ,	method for ethnography. <i>American anthropologist</i> ,	744
692	19(2):378–391.	104(1):106–122.	745
693	Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan	G Hofstede and GJ Hofstede. 2005. Cultures and or-	746
694	Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu,	ganizations: Software of the mind. third millennium	747
695	Tianyu Liu, et al. 2022. A survey on in-context learn-	edition.	748
696	ing. <i>arXiv preprint arXiv:2301.00234</i> .		
697	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,	Julia Kharchenko, Tanya Roosta, Aman Chadha, and	749
698	Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,	Chirag Shah. 2024. How well do llms represent	750
699	Akhil Mathur, Alan Schelten, Amy Yang, Angela	values across cultures? empirical analysis of llm re-	751
700	Fan, et al. 2024. The llama 3 herd of models. <i>arXiv</i>	sponses based on hofstede cultural dimensions. <i>arXiv</i>	752
701	<i>preprint arXiv:2407.21783</i> .	<i>preprint arXiv:2406.14805</i> .	753
702	Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas	Hannah Rose Kirk, Alexander Whitefield, Paul Rottger,	754
703	Schiefer, Amanda Askeff, Anton Bakhtin, Carol	Andrew M Bean, Katerina Margatina, Rafael	755
704	Chen, Zac Hatfield-Dodds, Danny Hernandez,	Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina	756
705	Nicholas Joseph, et al. 2023. Towards measuring	Williams, He He, et al. 2025. The prism alignment	757
706	the representation of subjective global opinions in	dataset: What participatory, representative and in-	758
707	language models. <i>arXiv preprint arXiv:2306.16388</i> .	dividualised human feedback reveals about the subjec-	759
708	Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh	tive and multicultural alignment of large language	760
709	Modi. 2023. Elicor: Corpus for analyzing llms for	models. <i>Advances in Neural Information Processing</i>	761
710	etiquettes. <i>arXiv preprint arXiv:2310.18974</i> .	<i>Systems</i> , 37:105236–105344.	762
711	Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff,	Louis Kwok, Michal Bravansky, and Lewis D Griffin.	763
712	Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model	2024. Evaluating cultural adaptability of a large lan-	764
713	alignment as prospect theoretic optimization. <i>arXiv</i>	guage model via simulation of synthetic personas.	765
714	<i>preprint arXiv:2402.01306</i> .	<i>arXiv preprint arXiv:2408.06929</i> .	766
715	Susan T Tufts Fiske and Shelley E Taylor. 2020. Social	Harrison Lee, Samrat Phatale, Hassan Mansoor, Kel-	767
716	cognition: From brains to culture.	lie Ren Lu, Thomas Mesnard, Johan Ferret, Colton	768
717	Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and	Bishop, Ethan Hall, Victor Carbune, and Abhinav	769
718	Heng Ji. 2024. Massively multi-cultural knowledge	Rastogi. 2023. Rlaif: Scaling reinforcement learning	770
719	acquisition & lm benchmarking. <i>arXiv preprint</i>	from human feedback with ai feedback.	771
720	<i>arXiv:2402.09369</i> .		
721	Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen	Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana	772
722	Rambow, Smaranda Muresan, and Heng Ji. 2022.	Sitaram, and Xing Xie. 2024a. Culturellm: In corpo-	773
723	Normsage: Multi-lingual multi-cultural norm discov-	rating cultural differences into large language models.	774
724	ery from conversations on-the-fly. <i>arXiv preprint</i>	<i>arXiv preprint arXiv:2402.10946</i> .	775
725	<i>arXiv:2210.08604</i> .		
726	Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda	Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen,	776
727	Askeff, Yuntao Bai, Saurav Kadavath, Ben Mann,	Xing Xie, and Jindong Wang. 2024b. Culturepark:	777
728	Ethan Perez, Nicholas Schiefer, Kamal Ndousse,	Boosting cross-cultural understanding in large lan-	778
729	et al. 2022. Red teaming language models to re-	guage models. <i>arXiv preprint arXiv:2405.15145</i> .	779
730	duce harms: Methods, scaling behaviors, and lessons		
731	learned. <i>arXiv preprint arXiv:2209.07858</i> .	Margarita Limón. 2001. On the cognitive conflict as	780
732	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song,	an instructional strategy for conceptual change: A	781
733	Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma,	critical appraisal. <i>Learning and instruction</i> , 11(4-	782
734	Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: In-	5):357–380.	783
735	centivizing reasoning capability in llms via reinforc-	Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip	784
736	ement learning. <i>arXiv preprint arXiv:2501.12948</i> .	Treleaven, and Miguel Rodrigues. 2023. Cultural	785
737	Kshitij Gupta, Benjamin Thérien, Adam Ibrahim,	alignment in large language models: An explana-	786
738	Mats L Richter, Quentin Anthony, Eugene Belilovsky,	tory analysis based on hofstede’s cultural dimensions.	787
739	Irina Rish, and Timothée Lesort. 2023. Continual pre-	<i>arXiv preprint arXiv:2309.12342</i> .	788
740	training of large language models: How to (re) warm	Yuri Miyamoto, Jiah Yoo, Cynthia S Levine, Jiy-	789
741	your model? <i>arXiv preprint arXiv:2308.04014</i> .	oung Park, Jennifer Morozink Boylan, Tamara Sims,	790
		Hazel Rose Markus, Shinobu Kitayama, Norito	791
		Kawakami, Mayumi Karasawa, et al. 2018. Culture	792
		and social hierarchy: Self-and other-oriented corre-	793
		lates of socioeconomic status across cultures. <i>Jour-</i>	794
		<i>nal of personality and social psychology</i> , 115(3):427.	795

796	Anjishnu Mukherjee, Aylin Caliskan, Ziwei Zhu, and	850
797	Antonios Anastasopoulos. 2024. Global gallery: The	851
798	fine art of painting culture portraits through multilin-	852
799	gual instruction tuning. In <i>Proceedings of the 2024</i>	853
800	<i>Conference of the North American Chapter of the</i>	854
801	<i>Association for Computational Linguistics: Human</i>	
802	<i>Language Technologies (Volume 1: Long Papers)</i> ,	
803	pages 6398–6415.	
804	Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani	855
805	Aljunied, Qingyu Tan, Liying Cheng, Guanzheng	856
806	Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023.	857
807	Seallms—large language models for southeast asia.	858
808	<i>arXiv preprint arXiv:2312.00738</i> .	859
809	OpenAI. 2024. Gpt-4 technical report . <i>Preprint</i> ,	860
810	<i>arXiv:2303.08774</i> .	861
811	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	862
812	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	
813	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	
814	2022. Training language models to follow instruc-	
815	tions with human feedback. <i>Advances in neural in-</i>	
816	<i>formation processing systems</i> , 35:27730–27744.	
817	Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee	863
818	Manakul, Sittipong Sripaisarnmongkol, Ruangsak	864
819	Patomporn, Pathomporn Chokchainant, and Kasima	865
820	Tharnpipitchai. 2023. Typhoon: Thai large language	866
821	models. <i>arXiv preprint arXiv:2312.13951</i> .	
822	Ramon Pires, Hugo Abonizio, Thales Sales Almeida,	
823	and Rodrigo Nogueira. 2023. Sabiá: Portuguese	
824	large language models. In <i>Brazilian Conference on</i>	
825	<i>Intelligent Systems</i> , pages 226–240. Springer.	
826	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo-	867
827	pher D Manning, Stefano Ermon, and Chelsea Finn.	868
828	2024. Direct preference optimization: Your language	869
829	model is secretly a reward model. <i>Advances in Neu-</i>	870
830	<i>ral Information Processing Systems</i> , 36.	871
831	Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina	
832	Reinecke, and Maarten Sap. 2024. Normad: A bench-	
833	mark for measuring the cultural adaptability of large	
834	language models. <i>arXiv preprint arXiv:2404.12464</i> .	
835	Michael J Ryan, William Held, and Diyi Yang. 2024.	872
836	Unintended impacts of llm alignment on global rep-	873
837	resentation. <i>arXiv preprint arXiv:2402.15018</i> .	874
838	William Saunders, Catherine Yeh, Jeff Wu, Steven Bills,	875
839	Long Ouyang, Jonathan Ward, and Jan Leike. 2022.	876
840	Self-critiquing models for assisting human evaluators.	877
841	<i>arXiv preprint arXiv:2206.05802</i> .	
842	John Schulman, Filip Wolski, Prafulla Dhariwal,	
843	Alec Radford, and Oleg Klimov. 2017. Proxi-	
844	mal policy optimization algorithms. <i>arXiv preprint</i>	
845	<i>arXiv:1707.06347</i> .	
846	Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu,	878
847	Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu,	879
848	and Deyi Xiong. 2023. Large language model align-	880
849	ment: A survey. <i>arXiv preprint arXiv:2309.15025</i> .	881
	Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya	882
	Horesh, Rogério Abreu de Paula, Diyi Yang, et al.	
	2024. Culturebank: An online community-driven	
	knowledge base towards culturally aware language	
	technologies. <i>arXiv preprint arXiv:2404.15238</i> .	
	Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei	883
	Huang, Yongbin Li, and Houfeng Wang. 2024. Pref-	884
	erence ranking optimization for human alignment.	885
	In <i>Proceedings of the AAAI Conference on Artificial</i>	886
	<i>Intelligence</i> , volume 38, pages 18990–18998.	887
	Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizil-	
	cec. 2024. Cultural bias and cultural alignment of	
	large language models. <i>PNAS nexus</i> , 3(9):pgae346.	
	Harry C Triandis, Robert Bontempo, Kwok Leung, and	888
	C Harry Hui. 1990. A method for determining cul-	889
	tural, demographic, and personal constructs. <i>Journal</i>	890
	<i>of Cross-Cultural Psychology</i> , 21(3):302–318.	
	Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi	
	Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R	
	Lyu. 2023a. Not all countries celebrate thanksgiving:	
	On the cultural dominance in large language models.	
	<i>arXiv preprint arXiv:2310.12481</i> .	
	Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao,	891
	Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan	892
	Xu, Maosong Sun, and Xing Xie. 2024. On the	893
	essence and prospect: An investigation of align-	894
	ment approaches for big models. <i>arXiv preprint</i>	
	<i>arXiv:2403.04204</i> .	
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Al-	895
	isa Liu, Noah A Smith, Daniel Khashabi, and Han-	896
	naneh Hajishirzi. 2022. Self-instruct: Aligning lan-	897
	guage models with self-generated instructions. <i>arXiv</i>	898
	<i>preprint arXiv:2212.10560</i> .	899
	Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xing-	
	shan Zeng, Wenying Huang, Lifeng Shang, Xin	
	Jiang, and Qun Liu. 2023b. Aligning large lan-	
	guage models with human: A survey. <i>arXiv preprint</i>	
	<i>arXiv:2307.12966</i> .	
	Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong.	900
	2024. Self-pluralising culture alignment for large	901
	language models. <i>arXiv preprint arXiv:2410.12971</i> .	902
	Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang,	903
	and Xing Xie. 2023. From instructions to intrinsic	
	human values—a survey of alignment goals for big	
	models. <i>arXiv preprint arXiv:2308.12014</i> .	
	Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang,	
	Songfang Huang, and Fei Huang. 2024. Rrhf: Rank	
	responses to align language models with human feed-	
	back. <i>Advances in Neural Information Processing</i>	
	<i>Systems</i> , 36.	
	A Supplements for Methodology	
	A.1 The Cultural Framework	
	We construct a cultural framework through integrat-	
	ing diverse definitions of cultures from multiple	

disciplines such as language, ethics and value The framework contains a total of 38 topics across four levels of various granularities.

I. Cultural Values

- **Schwartz's Theory of Basic Values:** Self-direction, Stimulation, Hedonism, Achievement, Power, Security, Tradition, Conformity, Benevolence, and Universalism.
- **Hofstede Cultural Dimensions (Hofstede and Hofstede, 2005):** Power Distance Index, Individualism vs. Collectivism, Uncertainty Avoidance Index, Masculinity vs. Femininity, Long-Term Orientation, and Indulgence vs. Restraint.

Definition about these value dimensions can be referred to the corresponding theory.

II. Social Norms

- **Gender Roles:** Refers to cultural expectations and behaviors assigned to genders. Key elements include roles in the family, workplace, and society, as well as attitudes toward gender equality and stereotypes.
- **Respect Elders:** Explores how elders are treated and regarded in society. Key elements include deference, caregiving, decision-making authority, and intergenerational relationships.
- **Family Obligations:** Refers to the responsibilities and expectations individuals have toward their family, including financial support, caregiving, and prioritizing family over personal needs.
- **Justice and Fairness:** Encompasses cultural attitudes toward fairness, equality, and the application of justice. Key elements include perceptions of legal systems, social equality, and ethical decision-making.
- **Individual Rights:** Individual Rights [Ethics and Norms]: Focuses on the emphasis placed on personal freedoms, autonomy, and individual rights within society. Key elements include freedom of speech, privacy, and access to opportunities.
- **Social Norms:** Refers to unwritten rules and expectations governing appropriate behavior

in social settings. Key elements include dress codes, public behavior, and communication styles.

- **Moral Duties and Altruism:** Explores the cultural emphasis on moral obligations and selfless acts for the welfare of others. Key elements include charity, volunteerism, and moral responsibility.
- **Environmental Ethics:** Refers to cultural attitudes and practices toward nature and the environment. Key elements include sustainability, conservation, and ecological responsibility.

III. Behavioral Practices

- **Social Relationship:** Examines the relationships within different social groups, including family, friends, colleagues, acquaintances, and strangers. Key elements include hierarchy, trust, intimacy, and obligations.
- **Work Behaviors:** Focuses on behaviors, hierarchies, and expectations in professional and business environments. Key elements include authority, teamwork, and professional etiquette.
- **Economic Behaviors:** Explores cultural attitudes toward money, wealth, and economic activities. Key elements include saving habits, spending patterns, and attitudes toward entrepreneurship.
- **Education System and Relationship:** Explores the structure, relationships, and norms within educational institutions, such as schools. Key elements include authority, learning methods, and examination systems.
- **Religious and Ceremonial Behaviors:** Rituals, festivals, and traditions tied to religious or secular practices. Key elements include rites of passage, community celebrations, and individual practices.

IV. Specific Customs

- **Language and Communication:** Refers to linguistic styles, communication methods, and formal versus informal interactions. Key elements include linguistic diversity, formality, and nonverbal cues.

993
994
995
996

997
998
999
1000

1001
1002
1003
1004
1005

1006
1007
1008
1009
1010

1011
1012
1013
1014
1015

1016
1017
1018
1019
1020
1021

1022
1023
1024
1025

1026
1027
1028
1029
1030

1031

1032
1033
1034
1035
1036
1037

- **Food:** Explores cuisine, drinks, and the social significance of food in different cultures. Key elements include culinary traditions, dining etiquette, and food symbolism.
- **Dining Etiquette:** Focuses on table manners, dining customs, and social norms related to eating. Key elements include utensils, seating arrangements, and conversation rules.
- **Festival and Holidays:** Examines celebrations, rituals, and traditions tied to holidays and festivals. Key elements include cultural events, religious observances, and seasonal customs.
- **Entertainment:** Focuses on leisure and recreational activities, such as games, movies, sports, and performances. Key elements include entertainment preferences, social gatherings, and cultural events.
- **Professional Settings:** Explores workplace norms, business etiquette, and professional behavior in different cultural contexts. Key elements include dress codes, communication styles, and work ethics.
- **Daily Rituals and Courtesies:** Examines cultural practices related to greetings, gift-giving, personal space, time management, visiting, and punctuality. Key elements include social norms, etiquette, and interpersonal communication.
- **Clothing:** Focuses on cultural norms surrounding attire, jewelry, and materials. Key elements include dress codes, fashion trends, and symbolic meanings of clothing.
- **Travel/Transport:** Explores cultural preferences and norms regarding mobility and transportation. Key elements include transportation modes, travel etiquette, and attitudes toward public and private transport."

B Supplements for Experimental Settings

B.1 License of Datasets

GlobalOpinionQA (Durmus et al., 2023) is under cc-by-nc-sa-4.0 license. CulturalBench (Chiu et al., 2024) is under cc-by-4.0 license. CultureBank (Shi et al., 2024) is under MIT license. And Prism (Kirk et al., 2025) is under cc license.

B.2 More Details about Baselines

Culturally Fine-tuned LLMs: Recent studies about cultural alignment fall into this category, all of which depend on supervised fine-tuning but collect training data in different ways.

- CultureLLM (Li et al., 2024a) employs 50 questions from the World Value Survey (WVS) with answers of the corresponding culture as seed data and augment semantically equivalent samples for training using a powerful LLM.
- CulturePark (Li et al., 2024b) builds an LLM-powered multi-agent communication framework, where agents playing roles of different cultures discuss about the topics from World Value Surveys thus high-quality cultural data is collected.
- CultureSPA (Xu et al., 2024) uncovers representative data of specific cultures by activating the LLM’s internal culture knowledge. It first synthesizes survey questions across cultural topics and identify the data that are different with culture-unaware and culture-aware prompting.
- CultureBank (Shi et al., 2024) collects self-narratives of diverse culture-aware scenarios such as working, immigration and traveling from the online community TikTok. It merges samples across all cultures to train a common model and applies the model through prompt engineering.

B.3 Usage of AI

In this paper, our proposed framework applies powerful LLMs to synthesis cultural data. In addition, we employ AI assistants to help writing.

C Supplements for Results

C.1 Overall Performance

Here, we present the alignment performance for each culture across the four datasets in Table 3, Table 4, Table 5 and Table 6.

C.2 Case Study

We conduct case studies to reveal the effectiveness of our framework for cultural alignment. A detailed analysis has been presented in Sec. 5.3, and we present more examples in Table 7.

1038
1039
1040
1041
1042

1043
1044
1045
1046
1047
1048

1049
1050
1051
1052
1053
1054

1055
1056
1057
1058
1059
1060
1061

1062
1063
1064
1065
1066
1067
1068

1069
1070
1071
1072

1073
1074
1075
1076
1077

1078
1079
1080
1081
1082

Models	Average	United Kingdom	China	South Koarea	India	Singapore
LLaMA-3.1-8B-Instruct as Backbone Model						
LLaMA-3.1-8B-Instruct	78.40	78.12	79.22	77.37	78.51	78.79
Role-Playing	79.32	78.12	78.99	77.54	82.48	79.46
CultureLLM	79.20	78.30	79.99	79.05	82.57	83.99
CulturePark	78.93	77.25	82.28	78.21	77.08	79.58
CultureSPA	78.16	79.76	79.45	78.41	81.76	79.50
CultureBank	80.99	82.57	81.46	77.71	82.16	83.01
Ours	81.28	79.78	79.01	82.73	84.42	80.48
Qwen2.5-7B-Instruct as Backbone Model						
Qwen2.5-7B-Instruct	81.25	83.76	80.61	83.55	76.22	82.11
Role-Playing	82.20	85.14	81.42	84.09	77.56	82.84
CultureLLM	82.34	82.59	81.56	83.22	83.91	83.99
CulturePark	81.22	79.86	79.58	80.31	78.41	81.26
CultureSPA	83.89	84.58	79.89	83.77	83.41	84.16
CultureBank	76.72	78.54	75.54	72.16	72.98	77.58
Ours	84.84	86.78	84.57	85.81	82.58	84.46

Table 3: Cultural alignment performance across various cultures on the GlobalOpinionQA dataset.

Models	Average	United Kingdom	China	South Korea	India	Singapore
gpt-3.5-turbo	27.38	24.00	27.12	29.27	21.74	34.78
gpt-4o-mini	46.92	48.00	44.07	51.22	52.17	39.13
gpt-4-turbo	56.87	68.00	62.71	53.65	52.17	47.82
gpt-3.5-turbo + Role-Play	34.69	56.00	25.42	26.83	39.13	26.09
gpt-4o-mini + Role-Play	53.40	72.00	50.85	46.34	50.00	47.83
gpt-4-turbo + Role-Play	65.28	72.00	64.40	68.29	52.17	69.56
LLaMA-3.1-8B-Instruct as Backbone Model						
LLaMA-3.1-8B-Instruct	29.04	28.00	25.42	24.39	32.61	34.78
Role-Playing	33.42	52.00	27.12	29.27	26.09	29.13
CultureLLM	31.51	56.00	35.59	26.83	17.39	21.74
CulturePark	23.41	48.00	10.17	21.90	28.26	20.87
CultureSPA	32.61	56.00	23.72	34.78	21.73	34.78
CultureBank	9.18	56.00	1.69	4.35	13.04	4.35
Ours	<u>34.90</u>	56.00	25.40	36.60	34.80	21.70
Qwen2.5-7B-Instruct as Backbone Model						
Qwen2.5-7B-Instruct	35.81	28.00	44.07	24.39	34.78	47.83
Role-Playing	33.51	52.00	38.98	24.39	30.43	21.74
CultureLLM	43.35	60.00	58.49	34.21	42.31	21.74
CulturePark	29.20	52.00	49.06	13.16	23.08	31.58
CultureSPA	32.00	44.00	45.28	34.21	19.23	17.39
CultureBank	32.02	44.00	45.28	34.21	19.23	17.39
Ours	32.35	52.00	22.03	26.83	43.48	17.39

Table 4: Cultural alignment performance across various cultures on the CulturalBench-Hard dataset.

Models	Average	United Kingdom	China	South Korea	India	Singapore
gpt-3.5-turbo	4.533	4.503	4.296	4.685	4.364	4.818
gpt-4o-mini	4.741	4.740	4.407	4.833	4.909	4.818
gpt-4-turbo	4.813	4.786	4.370	4.907	5.000	5.000
gpt-3.5-turbo + Role-Play	4.710	4.763	4.407	4.833	4.909	4.636
gpt-4o-mini + Role-Play	4.830	4.824	4.481	4.926	5.000	4.919
gpt-4-turbo + Role-Play	4.816	4.840	4.407	4.926	5.000	4.909
Llama-3.1-8b-Instruct as backbone model						
Llama3.1-8b-Instruct	4.210	4.305	4.000	4.018	4.545	4.181
Role-Playing	4.131	4.380	3.960	3.680	4.545	4.091
CultureLLM	3.810	4.099	3.885	3.250	4.000	3.818
CulturePark	3.402	4.038	3.320	2.833	3.545	3.273
CultureSPA	3.773	4.153	3.077	3.636	4.000	3.636
CultureBank	3.764	4.168	3.577	3.167	3.909	4.091
Ours	<u>4.429</u>	4.427	4.259	4.278	4.818	4.364
Qwen2.5-7B-Instruct as backbone model						
Qwen2.5-7B-Instruct	4.257	4.207	4.037	4.222	4.545	4.273
Role-Playing	4.298	4.321	4.259	4.093	4.545	4.273
CultureLLM	4.294	4.321	3.963	3.857	4.545	4.455
CulturePark	4.280	4.191	3.722	4.167	4.545	4.273
CultureSPA	2.756	2.466	2.423	2.167	2.727	4.000
CultureBank	4.258	4.305	4.307	4.222	4.364	4.273
Ours	<u>4.358</u>	4.527	4.222	4.130	4.636	4.273

Table 5: Cultural alignment performance across various cultures on the CulturalBank dataset.

Models	Average	United Kingdom	China	South Korea	India	Singapore
gpt-3.5-turbo	2.197	2.570	2.010	1.828	2.212	2.367
gpt-4o-mini	2.271	2.720	2.040	2.150	2.364	2.414
gpt-4-turbo	2.192	2.707	2.000	1.980	2.393	2.353
gpt-3.5-turbo + Role-Play	3.836	3.810	3.600	3.740	3.930	4.136
gpt-4o-mini + Role-Play	4.046	4.030	3.828	4.050	4.020	4.300
gpt-4-turbo + Role-Play	3.915	4.020	3.818	3.880	3.747	4.110
LLaMA-3.1-8B-Instruct as Backbone Model						
LLaMA-3.1-8B-Instruct	2.099	2.510	1.798	1.785	2.161	2.242
Role-Playing	3.530	3.727	3.290	3.292	3.630	3.710
CultureLLM	3.519	3.880	3.350	3.340	3.566	3.460
CulturePark	3.525	3.890	3.404	3.350	3.570	3.470
CultureSPA	3.519	3.870	3.343	3.343	3.590	3.440
CultureBank	3.092	3.380	2.650	3.040	3.290	3.100
Ours	<u>4.120</u>	4.350	3.940	4.030	4.070	4.210
Qwen2.5-7B-Instruct as Backbone Model						
Qwen2.5-7B-Instruct	2.272	2.656	1.990	2.051	2.282	2.380
Role-Playing	3.448	3.570	3.450	3.220	3.469	3.530
CultureLLM	3.304	3.540	3.153	3.110	3.260	3.350
CulturePark	3.343	3.500	3.141	3.143	3.420	3.300
CultureSPA	3.262	3.404	3.110	3.143	3.354	3.380
CultureBank	3.153	3.310	3.050	3.080	3.230	3.100
Ours	<u>4.028</u>	4.220	3.910	3.910	4.080	4.082

Table 6: Cultural alignment performance across various cultures on the Prism dataset.

Question	Is it okay to cut family off?	
General Response (LLaMA)	Cutting ties with family is a difficult decision... Family is deeply valued in many cultures... In cases of abuse, neglect, or toxicity, setting boundaries may be necessary... Seeking support from friends or therapists is recommended... Ultimately, the decision depends on personal well-being...	
	Our Models	CulturePark
United Kingdom	Family relationships in the UK balance individualism and tradition... Maintaining family ties is important, especially during weddings, funerals, and holidays, but respecting personal boundaries is key... Younger generations prioritize autonomy and emotional well-being, sometimes distancing themselves... Open communication helps manage family dynamics, but there is less pressure for immediate conflict resolution...	While cultural expectations in the UK advocate for family unity, it is essential to prioritize your well-being and emotional health...
China	Family harmony and filial piety are foundational in Chinese culture... When disagreements occur, reconciliation is prioritized over cutting ties... Elders are revered for wisdom... The concept of (A harmonious family brings prosperity) reinforces family unity... In extreme cases, temporary separation for self-protection may happen, but families are expected to repair relationships through shared traditions...	Cutting family ties is viewed with regret and caution in Chinese culture... It is seen as a last resort due to family obligations...
Question	What are your thoughts or arguments about the death penalty?	
General Response (LLaMA)	The death penalty is a complex and highly debated issue... Some argue it serves as a deterrent and provides justice... Others contend it is state-sanctioned killing that disproportionately affects marginalized communities... Many countries have abolished it due to concerns about wrongful convictions...	
	Our Models	CulturePark
South Korea	The death penalty was suspended in 1997 due to international pressure and public opposition... Reflects a commitment to human rights and Confucian values, emphasizing life and harmony... The legal system focuses on rehabilitation over punishment, influenced by "jeong" (emotional connection and empathy)...	Many arguments against the death penalty in Korea are built around issues such as execution error, racism, and disproportionate application...
China	Justice in China is deeply rooted in Confucian values, emphasizing harmony and respect for authority... The death penalty is seen as a tool to restore social balance, applied in severe crimes... However, compassion and "restoring the social bond" also play a role...	The effectiveness of the death penalty is debated... Some argue it acts as a deterrent in preventing major crimes...

Table 7: Case studies to demonstrate the effectiveness of cultural alignment, comparing our method and baselines.

Question	I have had trouble with honesty since I was a child. I would like to be a more honest person but I struggle.	
General Response (LLaMA)	It takes courage to acknowledge and want to change... Honesty builds trust, respect, and peace of mind... Dishonesty often stems from fear, shame, or past experiences... Seek guidance from a therapist... Start with small, incremental changes and build confidence over time...	
	Our Models	CulturePark
United Kingdom	Honesty in the UK is a cornerstone of relationships... The idiom "straight down the line" reflects directness with respect... British humor helps navigate difficult truths... The concept of "keeping it British" balances honesty with politeness and empathy, ensuring truth is delivered without offense...	Honesty is highly valued in British culture, reflecting integrity and respect for others.
China	Honesty is deeply rooted in Confucian principles, emphasizing filial piety, loyalty, and integrity... The concept of "ren" (kindness and benevolence) underscores truthfulness... However, "social harmony" (héxié) influences communication, prioritizing relationship preservation over blunt truth... Honesty is approached mindfully, balancing individual values with community norms...	Honesty is highly valued in Chinese culture, associated with integrity and loyalty...
Question	I find it difficult to get on with my boss. Suggestions?	
General Response (LLaMA)	Challenges with a supervisor are common... Workplace dynamics depend on cultural backgrounds, personal styles, and organizational structures... Reflect on communication style, seek common ground, and address issues calmly... Empathy and adaptability are key to building strong workplace relationships...	
	Our Models	CulturePark
South Korea	Work relationships in Korea are shaped by "jeong" (emotional bond and care)... Informal gatherings, such as "ddeokbokki nights", help strengthen connections... Honorifics and polite language are essential... Active listening shows commitment and respect, fostering a positive dynamic while maintaining hierarchical harmony...	Respecting elders and authority figures is crucial in South Korean culture...
China	Chinese workplace culture values harmony and respect for authority... Using honorifics like teacher reflects humility... Indirect communication preserves "face" ... "Guanxi" (personal connections) fosters trust... Emphasizing collective success over personal ambition enhances workplace relationships...	In Chinese workplaces, respect and harmony are paramount...

Table 8: Model case studies.