CAREDIO: ENHANCING CULTURAL ALIGNMENT OF LLM VIA REPRESENTATIVENESS AND DISTINCTIVENESS GUIDED DATA OPTIMIZATION

Anonymous authors
Paper under double-blind review

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 engagement and mitigating cultural conflicts. For this purpose, recently, different culture-specific corpora have been carefully curated, either synthesized or manually annotated. 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, causing insufficient cultural coverage with redundancy; (2) *Distinctiveness*: They struggle to distinguish the unique nuances of a given culture from shared patterns across other relevant ones, hindering precise cultural modelling. To handle these challenges, we introduce **CAReDiO**, a novel data optimization framework, which alternatively refines culture-sensitive questions and responses according to information-theoretic objectives in an in-context optimization manner, enhancing the cultural informativeness and distinguishability of constructed data. Extensive experiments on 15 distinct cultures demonstrate that CAReDiO can create high-quality data with richer cultural information and enable efficient alignment of small open-source or large proprietary LLMs with as few as 200 training samples, consistently outperforming previous datasets in both multi-choice and open-ended cultural benchmarks.

1 INTRODUCTION

As Large Language Models (LLMs) are widely integrated into human life (Bubeck et al., 2023; OpenAI, 2024; Dubey et al., 2024; Guo et al., 2025), aligning them with human values are imperative to mitigate safety risks and further improve user experience (Ouyang et al., 2022; Wang et al., 2024b). Focusing on universal values, *e.g*, the HHH principle (Askell et al., 2021; Bai et al., 2022), most prior studies overlook the *cultural diversity* rooted in human values. As a result, LLMs predominantly trained on English corpus are often biased towards Western cultures (Cao et al., 2023; Durmus et al., 2023), dissatisfying underrepresented cultural communities and raising unintended social tensions (Ryan et al., 2024). Therefore, *aligning LLMs with diverse and nuanced cultural values has become both an ethical and practical necessity* (AlKhamissi et al., 2024; Tao et al., 2024).

Early efforts align LLMs with the target culture in an In-Context Learning (ICL) way, through role-play instructions, native-language prompts, or few-shot examples (Durmus et al., 2023; Cao et al., 2023; Kwok et al., 2024), which suffers from inconsistent performance across tasks, especially for small models, as well as expensive inference cost and privacy concerns (Saunders et al., 2022). More recently, fine-tuning culture-aware LLMs has proven a practical alternative (Li et al., 2024b). Large-scale local-language corpora are used to produce regional LLMs (Gupta et al., 2023; Nguyen et al., 2023b; Pipatanakul et al., 2023), yet language alone does not sufficiently capture cultural values (Choenni et al., 2024; Mukherjee et al., 2024; Rystrøm et al., 2025). A more precise avenue is to build dedicated, culture-specific alignment datasets (Fung et al., 2024; Shi et al., 2024; Li et al., 2024a;b), while this demands massive manual annotation cost and is hard to scale.

Following this line, we ask *can we achieve cultural alignment at minimal cost by using fewer but more effective data?* To answer this question, we investigate culture theories such as the emic-etic theory (Triandis et al., 1990; Hofstede & Hofstede, 2005; Miyamoto et al., 2018; Fiske & Taylor,

2020; Mostowlansky & Rota, 2020), which argues that fully understanding a culture requires two complementary perspectives: *an internal (emic) view*, capturing the shared beliefs and practices that bind its member, and *an external (etic) view*, highlighting the traits that differentiate it from others. However, current datasets encounter two key challenges of reflecting both views: **Challenge 1. Representativeness**: The dataset should accurately capture the prominent and memorable constructs of the target culture without irrelevant and less important noise or redundancy (emic); **Challenge 2. Distinctiveness**: They need to highlight the unique nuances of a given culture from shared patterns across multiple relevant cultures (etic, *e.g.*, Korea and Japan), as shown in Fig. 1. Failing to handle these challenges hinders the precise and efficient modelling of specific culture stimuli and preferences, hence hurting cultural alignment efficiency.

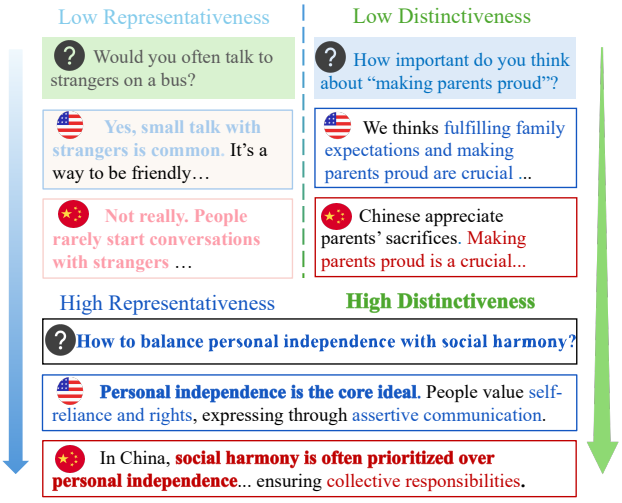


Figure 1: Current cultural alignment data fails to fully cover core features or capture subtle cultural differences.

This work proposes CAReDiO[1], a novel LLM-empowered in-context data optimization framework for automatic cultural data construction. CAReDiO alternately generates and refines cultural questions and responses to fulfill two information-theoretic objectives: i) an *information gain objective*, inspired by the Cultural Consensus Theory (Weller, 2007), to identify data samples that better reduce the LLM's cultural uncertainty and elicit more consensus, *improving representativeness*; ii) a *culture divergence objective*, grounded in the Cognitive Conflict Theory (Limón, 2001), to enhance samples' distinguishability from non-target cultures, which theoretically performs a point-wise optimization of different cultures' JS divergence, *achieving distinctiveness*. In this way, CAReDiO can utilize any given LLMs, either the smaller open-sourced one to be aligned, *e.g.*, Llama-3.1-8B-Instruct, or a separate larger one like GPT-4o, to automatically produce more informative and distinctive data for any specific culture (shown in Fig, 1). Such data could better capture culture boundaries and enable effective alignment across diverse cultures.

Our contributions are three-fold: (1) We are the first to investigate the *representativeness* and *distinctiveness* challenges in cultural alignment data motivated by culture theories. (2) We propose CAReDiO, an effective data optimization framework with two novel information-theoretic objectives, to tackle these challenges. (3) Using CAReDiO, we create the CARDSet, covering 15 cultures, and manifest our method can achieve better alignment across backbone LLMs and under both multi-choice and open-ended benchmarks, showing superiority to larger and even manually-curated datasets.

2 Related Work

Cultural alignment refers to adapting LLMs so that they better align with the nuanced values of diverse cultural communities. Existing research has focused on (1) evaluating culture awareness of LLMs and (2) developing methods for enhancing cultural alignment.

Evaluation of Culture Awareness Culture, as defined in (Adilazuarda et al., 2024), encompasses values, social norms, interpersonal behaviors and customs, etc, around which benchmarks are constructed. Many studies adopt well-established questionnaires from social sciences to analyze cultural values, such as the World Value Survey (WVS) (AlKhamissi et al., 2024), Hofstede framework (Cao et al., 2023; Masoud et al., 2023; Kharchenko et al., 2024; Sukiennik et al., 2025; Wang et al., 2023b), European Value Survey (EVS) (Tao et al., 2024) and GlobalOpinionQA (Durmus et al., 2023). More recent benchmarks construct open-ended QA data around these frameworks to better mirror real-world

[1]**C**ultural **A**lignment via **Re**presentativeness and **Di**stinctivenss **O**ptimization.

LLM use cases (Karinshak et al., 2024; Banerjee et al., 2024). Besides, other cultural dimensions have also been investigated: NORMSAGE (Fung et al., 2022) and NormAd (Rao et al., 2024) for social norms, and EtiCor (Dwivedi et al., 2023) for social etiquette. CulturalBench (Chiu et al., 2024) is a multiple-choice benchmark across comprehensive domains, curated and verified by humans. In most evaluations, *even advanced LLMs exhibit biases towards Western-centric values (Wang et al., 2023a), underscoring the urgency of promoting cultural alignment.*

Approaches to Cultural Alignment Early efforts focus on In-Context Learning (ICL) (Dong et al., 2022), including prompting LLMs to consider from cultural perspectives (Durmus et al., 2023), role-playing with demographic attributes (Kwok et al., 2024; Kharchenko et al., 2024) or enriching prompts with cultural value descriptions (Choenni & Shutova, 2024). Native language prompts sometimes also improve alignment (Durmus et al., 2023; Cao et al., 2023). However, these methods depend on strong ICL capabilities and pre-existing cultural knowledge, making them less effective for smaller or weaker LLMs (Saunders et al., 2022). A more scalable solution involves fine-tuning LLMs with culturally grounded datasets (Li et al., 2024a;b) and cultural learning-inspired strategies (Liu et al., 2025; Yuan et al., 2024), highlighting the need for high-quality cultural datasets.

Datasets for Cultural Alignment Current studies have explored four main categories of cultural datasets. The first is *large-scale local corpora*, which are used to pre-train regional LLMs from English-centric models (Pires et al., 2023; Nguyen et al., 2023b; Pipatanakul et al., 2023; Abbasi et al., 2023). Nonetheless, this approach requires prohibitive computational cost and language data alone provides limited cultural specificity. The second is *culture-related data filtered from large corpora or websites*. CultureInstruct (Pham et al., 2025) and CRAFT (Wang et al., 2024a) automatically identify culturally rich samples from large-scale collections. CultureBank (Shi et al., 2024) and CultureAtlas (Fung et al., 2024) collect cultural expressions from Tiktok/Reddit and Wikipedia respectively. To ensure higher quality, *manually curated datasets* are also incorporated, including NORMSAGE (Fung et al., 2022) and NORMBANK (Ziems et al., 2023) for social norms (Feng et al., 2025), CLIcK (Kim et al., 2024) and BLEnD (Myung et al., 2024) for cultural commensense (Nguyen et al., 2023a), and WVS survey for values (AlKhamissi et al., 2024). Finally, *cultural datasets augmented or synthesized by LLMs* with abundant cultural knowledge is an emerging category (Yuan et al., 2024). For example, CultureLLM (Li et al., 2024a) and CulturePark (Li et al., 2024b) augment WVS results with model-generated opinions. CultureSPA (Xu et al., 2024a) synthesizes questions with shifted answers under culture-unaware and -aware settings. Though these datasets are beneficial for cultural alignment, they could still encounter the challenges of representativeness and distinctiveness to achieve efficient and effective alignment, which are mainly optimized in this paper.

3 METHODOLOGY

3.1 FORMALIZATION AND OVERVIEW

Define $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ as an LLM parameterized by $\boldsymbol{\theta}$, which generates a response $\mathbf{y}$ to a given question $\mathbf{x}$; $p_{\mathbf{c}_1}(\mathbf{x},\mathbf{y}),\ldots,p_{\mathbf{c}_{K+1}}(\mathbf{x},\mathbf{y})$ as the true distributions of $K+1$ different cultures. Our goal is to find a set of cultural question-response pairs $q^*_{\mathbf{c}}(\mathbf{x}^*,\mathbf{y}^*)=\{(\boldsymbol{x}^*,\boldsymbol{y}^*)\}$ for the target culture $\mathbf{c}$ with satisfactory *representativeness* and *distinctiveness*, to achieve effective and sample-efficient cultural alignment of $p_{\boldsymbol{\theta}}$. For this purpose, we must solve the objective below:

$$q^*_{\mathbf{c}} = \underset{(\boldsymbol{x},\boldsymbol{y})}{\text{argtopN}} \{p_{\mathbf{c}}(\boldsymbol{x},\boldsymbol{y}) - \gamma * \frac{1}{K}\sum_{\boldsymbol{c}_k\neq\mathbf{c}} p_{\boldsymbol{c}_k}(\boldsymbol{x},\boldsymbol{y})\}, \tag{1}$$

where γ is a hyperparameter. This objective helps identify the data samples $(\boldsymbol{x},\boldsymbol{y})$ that are salient for the target culture $\mathbf{c}$ yet not shared by the non-target ones.

Nevertheless, each *true* $p_{\mathbf{c}_k}, k=1,\ldots,K+1$, is unavailable. Therefore, inspired by culture theories, we propose CAReDiO, an in-context framework to approximate and optimize Eq.(1). As shown in Fig. 2, CAReDiO consists of three core components: i) an *information gain objective* to create samples that can contribute more information and further reduce $p_{\boldsymbol{\theta}}$'s cultural uncertainty, improving representativeness; ii) a *culture divergence objective* to enhance data distinguishability from non-target cultures, improving distinctiveness, and iii) an iterative schema which alternatively refines the cultural question $\boldsymbol{x}$ and corresponding response $\boldsymbol{y}$ until convergence, leading to coherent, clean and informative alignment data. We elaborate on each part in the following subsections.

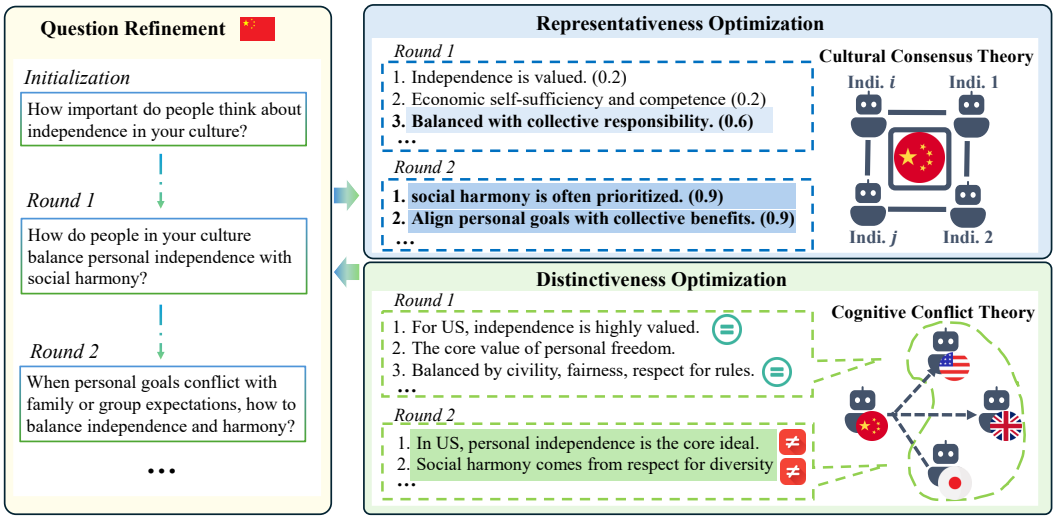


Figure 2: The CAReDiO framework, including modules for optimizing representativeness and distinctiveness, as well as an iterative schema to alternately refine questions and responses.

3.2 The CAREDIO Framework

We introduce our CAReDiO framework to optimize questions and answers for the target culture $\mathbf{c}$, handling the *C1: Representativeness* and *C2: Distinctiveness* challenges discussed in Sec. 1.

Representativeness Optimization via Information Gain The major challenge of solving Eq.(1) lies in that $p_{\mathbf{c}}(\mathbf{y}|\mathbf{x})$[2] is unavailable, and thus we can neither sample $\boldsymbol{y}$ from it nor obtain the density of $\boldsymbol{y}$. Fortunately, *Cultural Consensus Theory* (CulCT) (Weller, 2007) from cognitive anthropology and cultural psychology indicates that for a culture $\mathbf{c}$, its salient elements are shaped by shared cognition of people with cultural competence. Building upon this theory, we approximate representativeness optimization as a *consensus elicitation* problem, and quantify how well a response $\boldsymbol{y}$ reflects a given LLM $p_{\boldsymbol{\omega}}$'s cognition of culture $\boldsymbol{c}$ by Mutual Information (MI) $I_{\boldsymbol{\omega}}(\mathbf{c};\mathbf{y}|\mathbf{x}=\boldsymbol{x}) = \mathbb{E}_{p_{\boldsymbol{\omega}}(\mathbf{y}|\boldsymbol{x})}\mathbb{E}_{p_{\boldsymbol{\omega}}(\mathbf{c}|\mathbf{y},\boldsymbol{x})}\left[\log p_{\boldsymbol{\omega}}(\mathbf{c}|\boldsymbol{x},\mathbf{y}) - \log p_{\boldsymbol{\omega}}(\mathbf{c}|\boldsymbol{x})\right]$. Concretely, for a sampled $\boldsymbol{y}$ and a fixed target culture $\boldsymbol{c}$, we use point-wise MI to guide the data optimization process:

$$\hat{I}_{\boldsymbol{\omega}}(\boldsymbol{c};\boldsymbol{y}|\boldsymbol{x}) = \log p_{\boldsymbol{\omega}}(\boldsymbol{c}|\boldsymbol{x},\boldsymbol{y}) - \log p_{\boldsymbol{\omega}}(\boldsymbol{c}|\boldsymbol{x}) = \Delta^{\boldsymbol{\omega}}_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}). \tag{2}$$

Eq.(2) can be regarded as Information-Directed Sampling (IDS) (Hao et al., 2022), which identifies responses $\boldsymbol{y}$ that reinforce $p_{\boldsymbol{\omega}}$' understanding of culture $\boldsymbol{c}$. $p_{\boldsymbol{\omega}}$ could be either $p_{\boldsymbol{\theta}}$ (the target LLM to be aligned), where Eq.(2) performs a kind of Eliciting Latent Knowledge (ELK) (Mallen et al., 2023), or a stronger one (*e.g.*, GPT-5), where it degenerates into knowledge distillation (Xu et al., 2024b).

In practice, we use multiple LLMs $p_{\boldsymbol{\omega}_1}, \ldots, p_{\boldsymbol{\omega}_M}$ to mimic the rating and consensus forming process among a group of individuals, and use the following score to find the best $\boldsymbol{y}$:

$$\Delta^{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_M}_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{M}\sum_{i}^{M}\Delta^{\boldsymbol{\omega}_i}_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}), \tag{3}$$

where each $\boldsymbol{\omega}_i$ could be either a heterogeneous LLM or the same one with different demographic role-plays. In this way, for a given question, we can create representative responses that better reflect the *shared cognition* of a target culture $\boldsymbol{c}$. We then give the following conclusion:

Proposition 1 *If we use the data $\boldsymbol{y}$ optimized by Eq.(3) to fine-tune the target model $p_{\boldsymbol{\theta}}$, under mild conditions, its cultural learning converges faster with large gradients.*

Proof. See Appendix. A.

[2]For brevity, we describe the optimization of $\mathbf{y}$, assuming $\mathbf{x}$ is obtained in all equations. In practice, we also conduct a dual refinement process for $\mathbf{x}$ to optimize $p_{\mathbf{c}}(\mathbf{x},\mathbf{y})$.

Distinctiveness Optimization by Culture Divergence As demonstrated in Fig. 1, highly representative samples $\boldsymbol{y}$ for $\boldsymbol{c}$ won't naturally be distinguishable from other non-target ones, as closely connected cultures, *e.g.*, China, Japan and Korea, often share similar values, norms and behaviors. To achieve precise cultural modelling, we must better capture culture boundaries and construct distinguishable $\boldsymbol{y}$. Here, we resort to Cognitive Conflict Theory (CogCT) (Cosier & Rose, 1977), which indicates cognitive conflicts among cultures can provoke self-reflection on their own culture.

Technically, we set the target culture $\boldsymbol{c}=\boldsymbol{c}_{K+1}$, and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$ as the non-target ones for convenience, and use the generalized JS divergence (Englesson & Azizpour, 2021) between the empirical data distribution $q_{\boldsymbol{c}}$ and $p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}$, *i.e.*, $\mathrm{GJS}_{\alpha,\mathbf{w}}\left[p(\mathbf{y}|\boldsymbol{x}), p_{\boldsymbol{c}_1}(\mathbf{y}|\boldsymbol{x}), \ldots, p_{\boldsymbol{c}_K}(\mathbf{y}|\boldsymbol{x})\right]$, where $\alpha, \boldsymbol{w} = (w_1, \ldots, w_K) > 0$ are weights (hyperparameters) for each distribution with $\alpha + \sum_{i=1}^{K} w_i = 1$, to measure distinctiveness. Nevertheless, the difficulty still lies in that each true culture distribution $p_{\boldsymbol{c}_k}$ is unattainable. Therefore, we use the following objective instead:

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) \left[\log \frac{\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})}{1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})} + \log \frac{1-\alpha}{2\alpha}\right] + \log(1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})) = \Gamma_{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K}(\boldsymbol{y}|\boldsymbol{x}), \tag{4}$$

where $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$[3] is a (fine-tuned or approximated) classifier to give the probability that the response $\boldsymbol{y}$ (together with $\boldsymbol{x}$) does NOT come from any of the K non-target cultures. We provide a conclusion:

Proposition 2 *For any give $\boldsymbol{x}$ and $\boldsymbol{y}$, if the classifier error $|\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) - p^*(\boldsymbol{y} \notin p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}|\boldsymbol{y}, \boldsymbol{x})| < \epsilon$, and the classifier is not over-confident, i.e., $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) < \eta$, then maximizing Eq.(4) is an approximated point-wise maximization of the lower bound of $\mathrm{GJS}_{\alpha,\mathbf{w}}\left[q(\mathbf{y}|\boldsymbol{x}), p_{\boldsymbol{c}_1}(\mathbf{y}|\boldsymbol{x}), \ldots, p_{\boldsymbol{c}_K}(\mathbf{y}|\boldsymbol{x})\right]$, and the approximation error $\mathcal{E}$ is bounded by $\mathcal{E} < \epsilon |\log \frac{\eta(1-\alpha)}{2(1-\eta)\alpha}|$.*

Proof. See Appendix. A.

Prop. 2 implies we can directly use a reliable classifier to score each created $\boldsymbol{y}$ according to Eq.(4) and select the top-N ones to form the dataset $q_{\boldsymbol{c}}$. This process actually maximizes a lower bound of the true distinctiveness and elicits cultural differences from conflicts, even if we cannot access the real culture distributions. Eventually, we use the following score for data optimization:

$$\mathcal{S}_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) = \lambda_1 \cdot \Delta_{\boldsymbol{c}}^{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N}(\boldsymbol{y}|\boldsymbol{x}) + \lambda_2 \cdot \Gamma_{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K}(\boldsymbol{y}|\boldsymbol{x}) + \lambda_3 \cdot \mathbb{E}_{(\boldsymbol{x}', \boldsymbol{y}') \sim q_{\boldsymbol{c}}}\left[\| \left((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\right)\right], \tag{5}$$

where λ_1, λ_2, λ_3 are hyperparameters, and $\|$ represents sematic distance to enhance data diversity.

Iterative Data Optimization Based on the information-theoretic objectives, we iteratively and alternately create and refine the questions $\boldsymbol{x}$ and responses $\boldsymbol{y}$ for a specified target culture $\boldsymbol{c}$. Concretely, utilizing an LLM $p_{\boldsymbol{\omega}}$ and classifier $p_{\boldsymbol{\phi}}$ (in all our experiments, $\boldsymbol{\theta} = \boldsymbol{\omega}$, and $\boldsymbol{\phi}$ is a clustering based distance measurement), at the t-th iteration, we fix the $\boldsymbol{x}^{t-1}$ from the last iteration, and instruct the LLM to generate $\boldsymbol{y}^t$ that maximizes Eq.(5). Once we obtain the optimal $\boldsymbol{y}^t$, we further refine $\boldsymbol{x}^t$ to improve the score $\mathcal{S}_{\boldsymbol{c}}(\boldsymbol{x}|\boldsymbol{y})$. Concretely, we keep all generated $\boldsymbol{y}$ in the t-th iteration, and refine $\boldsymbol{x}^t$ to increase the generation probability of representative and distinctive responses and suppress that with lower scores. This optimization process is performed through LLM ICL without any training, until convergence or reaching early stopping criteria. The complete algorithm is summarized in Algo. 1.

3.3 CARDSET CONSTRUCTION AND ALIGNMENT

CARDSet Creation To validate the effectiveness of CAReDiO , we create a cultural alignment dataset CARDSet using our method. First, we prompt the LLM $p_{\boldsymbol{\omega}}$ to initiate questions $\boldsymbol{x}^0$ around diverse topics from core culture aspects, including cultural values, *e.g.*, Hofstede Cultural Dimensions (Hofstede & Hofstede, 2005), norms and behavioral practices (See Appendix. B.1 for topic details). Specifically, we employ the Self-Instruct approach (Wang et al., 2022) to synthesize N distinct questions for each topic, following four common question formats to align with the practical usage of LLMs: scenarios-based, value-oriented, open-ended and multiple-choice questions.

With questions $\boldsymbol{x}^0$, we perform the iterative data optimization process above as described in Algo. 1. For distinctiveness optimization, we prompt $p_{\boldsymbol{\omega}}$ with the question $\boldsymbol{x}$ and answers generated for other non-target cultures $\boldsymbol{y}_{\boldsymbol{c}_1}, \boldsymbol{y}_{\boldsymbol{c}_2}, \ldots, \boldsymbol{y}_{\boldsymbol{c}_K}$, then require it to generate $\boldsymbol{y}$ that are distinctive from the other answers to maximize Eq.(4). Concretely, $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$ in Eq.(4) is implemented as the softmax score of the embedding similarity between $\boldsymbol{y}$ and $\boldsymbol{y}_{\boldsymbol{c}}, \boldsymbol{y}_{\boldsymbol{c}_1}, \boldsymbol{y}_{\boldsymbol{c}_2}, \ldots, \boldsymbol{y}_{\boldsymbol{c}_K}$.

[3]We abbreviate $p_{\boldsymbol{\phi}}(\boldsymbol{y} \notin p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}|\boldsymbol{y}, \boldsymbol{x})$ or $p_{\boldsymbol{\phi}}(\boldsymbol{x} \notin p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}|\boldsymbol{y}, \boldsymbol{x})$ as $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$.

For representativeness optimization, we sample diverse responses $\boldsymbol{y}_{\{1...n\}}$ from $p_{\boldsymbol{\omega}}$ and summarize them into separated cultural expressions $\{\boldsymbol{e}_1^{\boldsymbol{y}}, \boldsymbol{e}_2^{\boldsymbol{y}}, \ldots, \boldsymbol{e}_m^{\boldsymbol{y}}\}$ (usually $m > n$). Then, we calculate their representativeness scores using Eq.(3) and produce the best $\boldsymbol{y}$ by aggregating all high-scoring expressions. Concretely, we implement Eq.(3) by prompting $p_{\boldsymbol{\omega}}$ to role-play a group of individuals from the target culture $\boldsymbol{c}$. To introduce comprehensive knowledge and enhance reliability, multiple individuals are set in three types: (i) *general people with various demographics* sampled from the WVS data of $\boldsymbol{c}$; (ii) *cultural experts with different backgrounds*, such as sociologist; and (iii) *cross-cultural researchers*. Maintaining the responses with a large score of Eq.(5), refine $\boldsymbol{x}^t$ to maximize the score $\mathcal{S}_{\boldsymbol{c}}(\boldsymbol{x}|\boldsymbol{y})$ following a process similar to above, replacing the generation of $\boldsymbol{y}$ to $\boldsymbol{x}$.

Algorithm 1: The CAREDiO Framework

Input: Maximum number iteration T, the LLM $P_{\boldsymbol{\theta}}$, target culture $\boldsymbol{c}$
Output: Optimized data $q_{\boldsymbol{c}} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=0}^{N}$
Initialize cultural questions $\{\boldsymbol{x}_i^0\}$
for $t = 1, 2, \ldots, T$ **do**
  Create responses $\{\boldsymbol{y}_i^t\}$ from the LLM;
  Calculate score $\mathcal{S}_{\boldsymbol{c}}(\boldsymbol{y}_i^t|\boldsymbol{x}^{t-1})$ by Eq.(5);
  Instruct the LLM to refine each $\boldsymbol{x}_i^{t-1}$ to $\hat{\boldsymbol{x}}_i^t$ and calculate $\mathcal{S}_{\boldsymbol{c}}(\hat{\boldsymbol{x}}_i^t|\boldsymbol{y}^t)$;
  Select high-score ones as $\{\boldsymbol{x}_i^t\}$

Alignment Fine-tuning Through the above optimization process, we can obtain a great deal of cultural data $q_{\boldsymbol{c}}^* = \{(x, y)\}$ reinforcing culture *representativeness* and *distinctiveness*, where each sample has a score $\mathcal{S}_{\boldsymbol{c}}(\boldsymbol{x}, \boldsymbol{y})$. However, training on the entire dataset incurs high computational costs and some data might be redundant. Therefore, we rank all these samples based on their score $\mathcal{S}_{\boldsymbol{c}}$ and sequentially select those according to the pre-defined computational budget. To ensure data diversity, we compute the similarity between the subsequent candidate and the selected ones, omitting those with a similarity score higher than τ. With responses generated for other cultures in the distinctiveness optimization step as dispreferred ones, we can fine-tune cultural LLMs via their SFT or DPO. To ensure a fair comparison, we follow most baselines to use SFT approaches in all experiments.

4 Experiments

4.1 Experimental Settings

Evaluation Datasets and Metrics We measure cultural alignment with four benchmarks targeting distinct aspects. Statistical information is listed in Table 1.

Table 1: Statistics of evaluation benchmarks.

Dataset	Types	#Samples	Metrics
CulturalBench	Multiple-Choice	1,227	Accuracy
Prism	Open-Ended QA	468	Quality Rating
GlobalOpinionQA	Questionnaire	2,556	Accuracy
WVS	Questionnaire	260	Consistency

(1) **CulturalBench** (Chiu et al., 2024): A manually curated benchmark with 1,227 four-choice questions for cultural knowledge, spanning 45 regions and 17 topics. There are two variants *Easy* and *Hard*, where *Hard* transforms each item into four binary questions and the LLM should judge all options correctly. *Accuracy* is calculated on the ground truth.

(2) **Prism** (Kirk et al., 2025): It contains conversations between 1,500 participants across 75 countries and 21 LLMs. We filter value-related questions raised by people from difference countries with two criteria: i) the question involves cultural topics such as relationship management and abortion; and ii) responses could vary meaningfully across cultures. We introduce both LLM-as-judge (Gemini-2.5-Pro) and native annotators to rate the *quality* of the responses on a 1-5 scale.

(3) **GlobalOpinionQA** (Durmus et al., 2023): It compiles items from Global Attitudes surveys (GAS) and World Value Survey (WVS). To avoid overlap with the next dataset, we retain only the GAS subset. Each item consists of a question, multiple choices and the choice distributions across various countries. We report *accuracy* as whether the model's prediction matches the top-1 human choice.

(4) **World Value Suryve (WVS)** (Xu et al., 2024a): A questionnaire surveying people's values across 13 topics. It collects real responses from people across countries. We compute the *consistency* between the predictions of an LLM $\boldsymbol{p}_\theta$ and the real answers from the culture $\boldsymbol{c}$ following Xu et al. (2024a). More details about these benchmarks and human evaluation are provided in Sec. C.1.

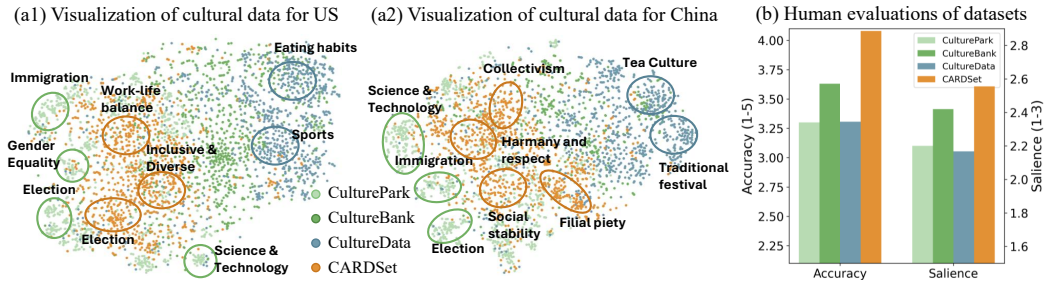


Figure 3: TSNE visualization and human evaluation of cultural datasets.

Baselines To comprehensively assess our approach, we conduct cultural alignment on LLMs from different families and scales, including proprietary GPT-4.1 and GPT-5, open-source models including LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Team, 2024) and Gemma-3-27B-IT (Team et al., 2025). Across all backbones, we adopt the **Role-Play** baseline that applies system prompts to simulate individuals from specific cultural backgrounds. Besides, we finetune cultural LLMs with our CARDSet and multiple datasets derived from different sources.

As summarized in Tab. 2, six cultural datasets are compared: **CultureLLM** (Li et al., 2024a) and **CulturePark** (Li et al., 2024b) are augmented by GPT-4-Turbo based on the real WVS data. **CultureSPA** (Xu et al., 2024a) is WVS-style opinion data synthesized by LLMs. **CultureBank** (Shi et al., 2024) and **CultureInstruct** (Pham et al., 2025) are culture-relevant text filtered from Tiktok/Reddit and the DOLMA corpus (Soldaini et al., 2024) respectively. **CultureData** is constructed by ourselves through merging all public manually created cultural datasets, *e.g.*, NORMBANK (Ziems et al., 2023) and CultureAtlas (Fung et al., 2024). For fair comparisons, we employ 1,000 samples for each culture except for CultureInstruct which is a mixed cultural dataset lacking explicit cultural labels.

More details about baselines and implementations can be found in Appendix C.4, C.5.

Table 2: Statistics of cultural datasets. Cult. Points: Culture points representing distinctive cultural aspect extracted from the dataset by GPT-4.1; Sim and SB are cosine similarity and Self-BLEU within each set; Cult. Sim is cosine similarity across cultural subsets.

Datasets	Source	#sample	Avg.L ↑	#Cult. Points ↑	Sim ↓	SB ↓	Cult. Sim ↓
CultureLLM	WVS augmentation	1,000 each	48.6	245.2	0.246	0.616	0.246
CulturePark	WVS augmentation	1,000 each	68.6	494.6	0.235	0.406	0.223
CultureSPA	LLM-synthetic	1,000 each	46.7	517.6	0.261	0.410	0.264
CultureBank	web platform	18,396 total	87.4	442.0	0.229	**0.167**	0.187
CultureInstruct	pretrain corpus	46,878 total	191.1	-	-	-	-
CultureData	manual annotation	1,000 each	16.8	1521.0	**0.199**	0.330	**0.127**
CARDSet	LLM-synthetic	1,000 each	**200.4**	**2027.0**	0.251	0.324	0.202

4.2 Cultural Dataset Analysis

Before delving into the performance of cultural alignment, we first compare the quality of CARDSet generated by our CAReDiO framework with existing cultural datasets introduced in Sec. 4.1.

Quantitative Analysis As shown in Tab. 2, textual samples in CARDSet are generally longer and contain richer cultural information (with more cultural points in the raw text extracted by GPT-4.1). This suggests that CARDSet is more informative and better captures core cultural factors. Moreover, CARDSet exhibits lower intra- and inter-cultural similarity, indicating that the dataset encodes more diverse and unique cultural knowledge compared to prior baselines.

Visualization Analysis We further apply t-SNE to visualize the embedding space of cultural texts in Fig. 3. For CulturePark, it covers important but generic value topics for different countries, such as '*Gender Equality*'. It might be hard to distinguish subtle cultural variations. Regarding CultureData, it captures unique aspects of different cultures, like the 'Tea Culture' for China, while often at a superficial level and loosely connected to deeper cultural norms and values. By contrast,

CARDSet perfectly overcomes both challenges, highlighting representative and distinctive aspects that tie directly to cultural cores. For example, CARDSet captures a defining value *Inclusive & Diverse* for the US and the *Filial Piety* for China. At the distributional level, CARDSet also locates at a joint region of other datasets, suggesting it provides both broader coverage and core factors.

Human Evaluation on Data Quality To complement automatic evaluation, we recruit native annotators from the corresponding cultures to assess the data quality along two dimensions: 1) Accuracy (1-5), which means the consensus level of this data to the culture; 2) Saliance (1-3), representativeness and importance of the cultural aspect. As shown in Fig. 3 (b), CARDSet exhibits significant superiority to other datasets across both dimensions, highlight the effectiveness of our method. More details about annotator recruitment and guidance are provided in Appendix C.3.

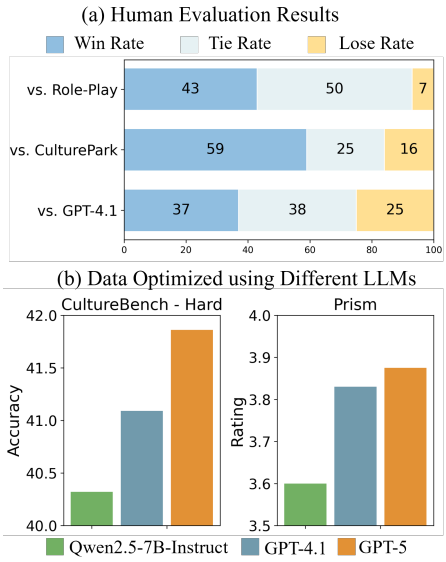


Figure 4: Results of cultural alignment.

4.3 Cultural Alignment Performance

Tab. 3 and Fig. 4 compares the cultural alignment performance of CAReDiO against various baselines. We conduct alignment across 15 cultures and report the average score here. Detailed per-culture results and results of Llama-3.1-8B-Instruct are in Appendix E.1.

First, *CAReDiO consistently enhances cultural alignment across LLM families and scales, including strong proprietary ones*. Compared to the raw models (Qwen2.5-7B, Gemma-3-27B, GPT-4.1) and the role-playing variants, CAReDiO achieves significant gains on all benchmarks. Interestingly, Fig. 4 (b) shows that using more capable LLM backbones for synthesis leads to further improvement, while data generated by the raw model also yields clear gains. This fully demonstrates that the improvement of cultural alignment by our framework does not solely derive from knowledge distillation but also designed objectives for representativeness and distinctiveness.

Second, *CAReDiO outperforms baselines in most benchmarks, especially on CultureBench and Prism.* CultureBench is a manually curated benchmark about extensive cultural aspects, while questions in Prism come from real-world interactions. Superior performance on these data highlights the practical robustness and adaptability of our method. On GlobalOpinionQA and WVS, CAReDiO lags slightly behind CultureLLM, we guess it is due to that CultureLLM is directly augmented from actual WVS survey data and thus have an advantage in similar evaluations. We also consider synthesizing more data around this perspective to enhance the results in the future. Notably, for the advanced GPT-5, CAReDiO achieves comparable or better results on Prism. Our framework can be extended to using GPT-5 for data synthesize to improve itself once GPT-5 is available for fine-tuning in the future.

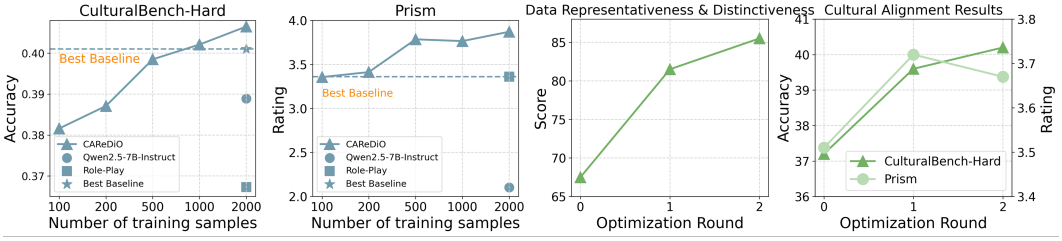
Human Evaluation on Alignment We also conduct human evaluation with *native annotators* from the US, China, Japan and Poland (three per culture). Showing questions from Prism and responses generated by different methods, they evaluate the quality (consensus level) from 1 (conflict the culture) to 5 (highly aligned with the culture). For each culture, we label 50-100 samples and report the average in Fig. 4 (a). *Human annotators consistently rate responses by CAReDiO higher than those from baselines.* This demonstrates that our approach not only improves benchmark scores but produces outputs perceived as more culturally aligned by humans from target cultures.

4.4 Hyperparameters Analysis

Number of Training Samples We conduct experiments by continuously increasing the training samples from 100 to 2,000, selecting higher-scoring samples first. As shown in Fig. 5, performance improves as more samples are introduced. Importantly, the earlier selected samples contribute more significant performance gains. Such observations indicate that our dataset is diverse enough to continuously provide learning benefits, while prioritized samples with higher representativeness and

Table 3: Evaluation results of cultural alignment across four benchmarks. 'CB' denotes CulturalBench; Average represents the mean score over all benchmarks. * marks the best results across all backbones. For each LLM, the best and second-best results are highlighted in bold and underlined.

Family	Method	CB-Easy	CB-Hard	Prism	GlobalOpinionQA	WVS	Average
Proprietaty LLMs	GPT-5	88.79	59.54	2.187	46.27	62.38	60.14
	GPT-5 + Role-Play	89.55	59.99	4.519	60.08*	70.44*	74.09*
Qwen2.5-7B-Instruct	Raw Model	72.01	38.90	2.103	53.28	59.68	53.18
	Role-Play	72.38	36.73	3.364	55.83	64.96	59.44
	CultureLLM	71.79	34.79	3.121	57.11	**65.49**	58.32
	CulturePark	71.99	34.41	3.107	56.47	57.83	56.57
	CultureSPA	70.92	36.25	3.108	52.83	62.00	56.83
	CultureBank	72.28	27.34	3.193	56.43	62.47	56.48
	CultureInstruction	72.77	27.75	3.346	**57.78**	63.25	57.69
	CultureData	72.83	40.11	3.354	57.44	64.69	60.43
	CAReDiO	**73.48**	**40.20**	**3.871**	56.23	65.26	**62.51**
Gemma-3-27B-IT	Raw Model	82.11	46.59	2.174	51.83	64.77	57.76
	Role-Play	81.33	48.28	4.571	54.84	67.22	68.62
	CultureLLM	80.46	46.31	4.441	58.15	66.99	68.14
	CulturePark	81.85	46.34	4.474	**59.74**	65.95	68.67
	CultureSPA	81.40	48.00	4.431	56.59	67.76	68.48
	CultureBank	81.82	41.88	4.323	55.89	67.11	66.63
	CultureInstruction	76.18	18.19	3.525	59.22	61.42	57.10
	CultureData	81.83	44.28	4.032	58.33	**68.02**	66.62
	CAReDiO	**82.56**	**48.88**	**4.627***	58.25	67.96	**70.04**
GPT-4.1	Raw Model	89.82	59.45	2.131	52.69	60.91	61.10
	Role-Play	89.29	63.47	4.270	53.76	**69.85**	72.35
	CultureBank	**90.80***	60.00	4.226	**56.76**	68.11	72.04
	CAReDiO	90.32	**63.54***	**4.336**	56.64	69.66	**73.37**



(a) Results with different numbers of training samples (b) Results with data across optimization rounds

Figure 5: Results of two hyper-parameters: number of training samples and optimization round.

distinctiveness are more effective for cultural alignment, supporting the efficiency of our approach. 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.

Optimization Rounds. We also compare the cultural datasets generated in different optimized rounds. As shown in Fig. 5 (b), the representativeness and distinctiveness score of data continuously increases along the iterative process, especially in the first round. Correspondingly, the alignment performance also improves. Besides, we found the performance gain is mostly achieved in the first round, thus our framework is also efficient in data synthetic to minimize the generation cost.

5 Conclusion

This paper addresses the challenges of representativeness and distinctiveness in cultural alignment datasets by introducing CAReDiO, an LLM-empowered data optimization framework for automatic cultural data construction. It involves an iterative process to generates and refines questions and responses for two information-theoretic objectives, thus enhancing representativeness and distinctiveness. Using the constructed dataset CARDSet covering 15 cultures, we demonstrate the superiority of CAReDiO over several recent datasets. Limitations and future directions are discussed in Appendix F.

Ethics Statement

This paper introduces CAReDiO, a novel framework to enhance cultural alignment of LLMs. We are aware of the potential ethical implications and societal impact of this line of work, and we emphasize the importance of responsible development. For transparency and reproducibility, we provide implementation details in the Appendix and commit to releasing the necessary code and data upon acceptance. Given the cultural biases that persist in current LLMs and the associated risks, our framework is specifically designed to improve cultural alignment and is not intended for malicious use. While our experiments focus on 15 cultures, the framework is generalizable to a wide range of cultures, which we believe contributes to greater fairness and inclusivity. Furthermore, by improving the efficiency of alignment, our approach makes it more feasible to support underrepresented cultures with limited resources.

Reproducibility statement

Due to space limitations, many technical details such as derivations, implementations, and some experimental settings could not be included in the main body and have instead been provided in the Appendix. Specifically, the Appendix contains: (1) derivation of CAReDiO algorithm in Appendix A, (2) details for dataset creation such as the topic definition and prompts in Appendix B, (3) more detailed experimental settings in Appendix C and (4) detailed experimental results in Appendix E. We submit the core code of our method as the supplementary materials for clarity and commit to release the necessary code and data upon acceptance to support reproducibility.

References

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

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

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

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. *arXiv preprint arXiv:2410.12880*, 2024.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

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

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*, 2023.

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. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*, 2024.

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

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

Richard A Cosier and Gerald L Rose. Cognitive conflict and goal conflict effects on task performance. *Organizational behavior and human performance*, 19(2):378–391, 1977.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. Eticor: Corpus for analyzing llms for etiquettes. *arXiv preprint arXiv:2310.18974*, 2023.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. *arXiv preprint arXiv:2505.19484*, 2025.

Susan T Tufts Fiske and Shelley E Taylor. Social cognition: From brains to culture. 2020.

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

Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *arXiv preprint arXiv:2210.08604*, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*, 2023.

Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pp. 8446–8464. PMLR, 2022.

G Hofstede and GJ Hofstede. Cultures and organizations: Software of the mind. third millennium edition, 2005.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*, 2024.

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

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2025.

Louis Kwok, Michal Bravansky, and Lewis D Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*, 2024.

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

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*, 2024b.

Margarita Limón. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and instruction*, 11(4-5):357–380, 2001.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. Cultural learning-based culture adaptation of language models. *arXiv preprint arXiv:2504.02953*, 2025.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2023.

Yuri Miyamoto, Jiah Yoo, Cynthia S Levine, Jiyoung Park, Jennifer Morozink Boylan, Tamara Sims, Hazel Rose Markus, Shinobu Kitayama, Norito Kawakami, Mayumi Karasawa, et al. Culture and social hierarchy: Self-and other-oriented correlates of socioeconomic status across cultures. *Journal of personality and social psychology*, 115(3):427, 2018.

Till Mostowlansky and Andrea Rota. Emic and etic. 2020.

Anjishnu Mukherjee, Aylin Caliskan, Ziwei Zhu, and Antonios Anastasopoulos. Global gallery: The fine art of painting culture portraits through multilingual instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6398–6415, 2024.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pp. 1907–1917, 2023a.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*, 2023b.

OpenAI. Gpt-4 technical report, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. Cultureinstruct: Curating multicultural instructions at scale. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9207–9228, 2025.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*, 2023.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pp. 226–240. Springer, 2023.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*, 2024.

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

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*, 2025.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*, 2024.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*, 2025.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2, 2024.

Harry C Triandis, Robert Bontempo, Kwok Leung, and C Harry Hui. A method for determining cultural, demographic, and personal constructs. *Journal of Cross-Cultural Psychology*, 21(3): 302–318, 1990.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F Chen. Craft: Extracting and tuning cultural instructions from the wild. *arXiv preprint arXiv:2405.03138*, 2024a.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*, 2023a.

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan Xu, Maosong Sun, and Xing Xie. On the essence and prospect: An investigation of alignment approaches for big models. *arXiv preprint arXiv:2403.04204*, 2024b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *arXiv preprint arXiv:2311.16421*, 2023b.

Susan C Weller. Cultural consensus theory: Applications and frequently asked questions. *Field methods*, 19(4):339–368, 2007.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. *arXiv preprint arXiv:2410.12971*, 2024a.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024b.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*, 2024.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*, 2023.

A SUPPLEMENTS FOR DERIVATION

Define $p_\theta(\mathbf{y}|\mathbf{x})$ as an LLM which generates a response $\mathbf{y}$ to a given question $\mathbf{x}$; $\mathbf{c}_1, \ldots, \mathbf{c}_K$ as K different cultures. Our goal is to find the best question-response pair $(\mathbf{x}^*, \mathbf{y}^*)$ which satisfies: i) $(\mathbf{x}^*, \mathbf{y}^*) = \text{argmax}_{(\mathbf{x},\mathbf{y})} p_{\mathbf{c}_i}(\mathbf{x}, \mathbf{y})$ where $\mathbf{c}_i$ is the target culture; and ii) $(\mathbf{x}^*, \mathbf{y}^*) = \text{argmin}_{(\mathbf{x},\mathbf{y})} \frac{1}{K-1} \sum_{k,k\neq i} p_{\mathbf{c}_k}(\mathbf{x}, \mathbf{y})$. Requirement i) follows our *Representativeness* rule and Requirement is in line with *Distinctiveness*. We should how it can be approximated and solved.

Representativeness Optimization The major challenge of $(\mathbf{x}^*, \mathbf{y}^*) = \text{argmax}_{(\mathbf{x},\mathbf{y})} p_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$ lies in that the true distribution of the target culture $p_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$ is unavailable, and thus we cannot either sample from it or get the density. We resort to the *Cultural Consensus Theory*, which shows the "culturally correct" answer is determined by the shared beliefs of people with cultural competence. Based on this theory, we assume each large enough LLM $p_\theta(\mathbf{y}|\mathbf{x})$ possesses sufficient competence but it is usually unelicited. Therefore, we approximate Representativeness as a *consensus elicitation* problem, and find the best $\mathbf{y}$[4] that maximizes $I_{\boldsymbol{\theta}}(\mathbf{c}; \mathbf{y}|\mathbf{x} = \boldsymbol{x})$:

$$\begin{aligned} I_{\boldsymbol{\theta}}(\mathbf{c}; \mathbf{y}|\mathbf{x} = \boldsymbol{x}) &= \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{x})} \int p_\theta(\mathbf{c}|\mathbf{y}, \boldsymbol{x}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{y}, \boldsymbol{x}) p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\mathbf{c}|\boldsymbol{x}) p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{x})} d\mathbf{c} \\ &= \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{x})} \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{y},\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\boldsymbol{x}, \mathbf{y}) - \log p_{\boldsymbol{\theta}}(\mathbf{c}|\boldsymbol{x})\right]. \end{aligned} \tag{6}$$

For a sampled $\boldsymbol{y}$, we then use point-wise mutual information for Eq.(6) and use the following information score to guide the data optimization process:

$$\hat{I}_{\boldsymbol{\theta}}(\boldsymbol{c}; \boldsymbol{y}|\boldsymbol{x}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{x}, \mathbf{y}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{x}) = \Delta^{\boldsymbol{\theta}}(\boldsymbol{y}). \tag{7}$$

Eq.(7) represents a form of Information-Directed Sampling (IDS) (Hao et al., 2022), which helps find the response $\boldsymbol{y}$ that reinforces the LLM $p_{\boldsymbol{\theta}}$' understanding of culture $\boldsymbol{c}$. When $p_{\boldsymbol{\theta}}$ is the target LLM (to be aligned) itself, the optimization process can be regarded as a kind of Eliciting Latent Knowledge (ELK) (Mallen et al., 2023); when $p_{\boldsymbol{\theta}}$ is a larger LLM (potentially with better culture competence), *e.g.*, GPT-5, we conduct typical knowledge distillation. In practice, we use multiple LLMs $\Delta^{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N}(\boldsymbol{y}) = \sum_i^N \Delta^{\boldsymbol{\theta}_i}(\boldsymbol{y})$ (each $\boldsymbol{\theta}_i$ could be either a new heterogeneous LLM or the same one with different settings) to perform a better consensus elicitation.

Theorem 1. If we use the vy optimized by Eq.(7) to fine-tune the target model $p_{\boldsymbol{\theta}}$, the its cultural learning converges faster with large gradients.

Proof. Assume we have true samples from the target culture $\boldsymbol{c}$, which forms the empirical distribution $p_{\boldsymbol{c}}(\mathbf{x}, \mathbf{y})$, and we use these samples to train the target LLM $p_{\boldsymbol{\theta}}$ with the following loss:

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{\boldsymbol{c}}} \left[\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\right], \tag{8}$$

and then, we have the gradient:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{\boldsymbol{c}}} \left[\nabla_{\boldsymbol{\theta}} l_\theta(\boldsymbol{y}, \boldsymbol{x})\right], \; l_\theta(\boldsymbol{y}, \boldsymbol{x}) = -\log p_{\boldsymbol{\theta}}(y|x). \tag{9}$$

We then consider $||\nabla_{\boldsymbol{\theta}} l_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x})||$ and demonstrate $||\nabla_{\boldsymbol{\theta}} l_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x})|| \propto [\log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) - p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})]$. Suppose the model conducts softmax to compute the probability of output in the last layer. For brevity, we consider $\boldsymbol{y}$ as one single token (which is typical in cultural questionnaire), and thus have:

$$\begin{aligned} \boldsymbol{z}_{\boldsymbol{y}}(\boldsymbol{x}) &= \boldsymbol{w}_{\boldsymbol{y}}^T \cdot h(\boldsymbol{x}), \\ p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) &= \frac{\exp(z_{\boldsymbol{y}}(\boldsymbol{x}))}{\sum_i \exp(z_i(x))}, \\ \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) &= \boldsymbol{z}_{\boldsymbol{y}}(\boldsymbol{x}) - \log \sum_i \exp(\boldsymbol{z}_i(\boldsymbol{x})) \end{aligned} \tag{10}$$

where $h(\boldsymbol{x})$ is the input of the last layer. Computing the gradient for $\boldsymbol{w}_i$, we have:

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})}{\partial w_i} = (\mathbb{I}[i = y] - p_{\boldsymbol{\theta}}(i|\boldsymbol{x})) \cdot h(\boldsymbol{x}). \tag{11}$$

[4]We do an iterative optimization of $\mathbf{y}$ and $\mathbf{x}$ separately. For brevity, we fix $\mathbf{x} = \boldsymbol{x}$ and optimize $\mathbf{y}$.

Then we have $||\nabla_{\boldsymbol{\theta}} l_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x})|| = (\mathbb{I}[i = y] - p_{\boldsymbol{\theta}}(i|\boldsymbol{x})) \cdot h(\boldsymbol{x})$. For simplicity, we consider the case $i = y$, and have $||1 - p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})|| \cdot ||h(x)|| \propto 1 - e^{\log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) - \delta(\boldsymbol{x}, \boldsymbol{y})}$ where $\delta(\boldsymbol{x}, \boldsymbol{y}) = \log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$. This indicate that using a $(\boldsymbol{y}, \boldsymbol{x})$ with large $\delta(\boldsymbol{x}, \boldsymbol{y})$ to fine-tune the LLM $p_{\boldsymbol{\theta}}$ can leads to larger gradients which accelerates the cultural training/learning.

Since the concrete value of $\log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x})$ is unavailable, again, we use $p_{\boldsymbol{\theta}}$'s self-judgement to approximate it, that is, $\log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) \approx \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c})$. Then, we have

$$\begin{aligned}
\delta(\boldsymbol{x}, \boldsymbol{y}) &= \log p_{\boldsymbol{c}}(\boldsymbol{y}|\boldsymbol{x}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \\
&\approx \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \\
&= \log p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{y}, \boldsymbol{x}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{x}) \\
&= \Delta^{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}),
\end{aligned} \tag{12}$$

which indicates the $\boldsymbol{y}$ optimized using Eq.(7) can approximately accelerate $p_{\boldsymbol{\theta}}$'s cultural learning.

Distinctiveness Optimization To optimize $(\mathbf{x}^*, \mathbf{y}^*) = \text{argmin}_{(\mathbf{x},\mathbf{y})} \frac{1}{K} \sum_{k,k\neq i} p_{\mathbf{c}_k}(\mathbf{x}, \mathbf{y})$, we refer to the *Cognitive Conflict Theory*, and elicit cultural differences from conflicts. For a given question $\boldsymbol{x}$, assume we have collected a set of $\boldsymbol{y}$ by Eq.(7), which forms an empirical distribution $q(\mathbf{y}|\boldsymbol{x})$ for the target culture, *e.g.*, Japan, we aim to find the best q with minimal overlap with non-target cultures, *e.g.*, Korea, Singapore and UK, $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$.

Concretely, we use the following objective:

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) \left[\log \frac{\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})}{1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})} + \log \frac{1 - \alpha}{2\alpha} \right] + \log(1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})) = \Gamma(\boldsymbol{y}|\boldsymbol{x}), \tag{13}$$

$$q^*(\boldsymbol{y}|\boldsymbol{x}) = \underset{\boldsymbol{y}}{\text{argtop}}\, \Gamma(\boldsymbol{y}|\boldsymbol{x}) \tag{14}$$

where $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$ is a classifier to give the probability that the response $\boldsymbol{y}$ comes from any of the K non-target culture, and $\alpha \in [0, 1]$ is a hyperparamter (the weight of the target culture).

Theorem 2. For any give $\boldsymbol{x}$ and $\boldsymbol{y}$, if the classifier error $|\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) - p^*(\boldsymbol{y} \notin p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}|\boldsymbol{y}, \boldsymbol{x})| < \epsilon$, and the classifier is not over-confident, i.e., $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) < \eta$, then maximizing Eq.(13) is an approximated point-wise maximization of the lower bound of GJS$_{\alpha,\mathbf{w}}$ $[q(\mathbf{y}|\boldsymbol{x}), p_{\boldsymbol{c}_1}(\mathbf{y}|\boldsymbol{x}), \ldots, p_{\boldsymbol{c}_K}(\mathbf{y}|\boldsymbol{x})]$, and the approximation error $\mathcal{E}$ is bounded by $\mathcal{E} < \epsilon |\log \frac{\eta(1-\alpha)}{2(1-\eta)\alpha}|$.

Proof. For a given question $\boldsymbol{x}$, assume we have collected a set of $\boldsymbol{y}$ by Eq.(7), which forms an empirical distribution $q(\mathbf{y}|\boldsymbol{x})$ for the target culture, *e.g.*, Japan, we aim to find the best q with minimal overlap with non-target cultures, *e.g.*, Korea, Singapore and UK, $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$. We optimize:

$$q^*(\mathbf{y}|\boldsymbol{x}) = \underset{\boldsymbol{q}}{\text{argmax}}\, \text{GJS}_{\alpha,\mathbf{w}} \left[q(\mathbf{y}|\boldsymbol{x}), p_{\boldsymbol{c}_1}(\mathbf{y}|\boldsymbol{x}), \ldots, p_{\boldsymbol{c}_K}(\mathbf{y}|\boldsymbol{x}) \right], \tag{15}$$

where GJS is Generalized Jensen divergence, and α, $\boldsymbol{w} = (w_1, \ldots, w_K) > 0$ are weights for each distribution with $\alpha + \sum_{i=1}^K w_i = 1$.

For brevity, we omit $\boldsymbol{x}$. Since $1 - \alpha = \sum_{i=1}^K w_k$, define $\beta_i = \frac{w_i}{1-\alpha}$, and the average non-target culture distribution as $\hat{p} = \frac{\sum_{i=1}^K \boldsymbol{w}_i p_{\boldsymbol{c}_i}}{1-\alpha}$, we have $m = \alpha p + \sum_{i=1}^K w_i p_{\boldsymbol{c}_i}$, and thus $\text{GJS}_{\alpha,\mathbf{w}} [q, p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}] = \alpha \text{KL}[q||m] + (1 - \alpha) \sum_{i=1}^K \beta_i \text{KL}[p_{\boldsymbol{c}_i}||m] = \alpha q + (1 - \alpha)\hat{p}$. Then,

we further have:

$$
\begin{aligned}
&\text{GJS}_{\alpha,\mathbf{w}}\left[q, p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}\right] \\
&= \alpha \mathbb{E}_q[\log q - \log m] + (1-\alpha) \sum_{i=1}^{K} \beta_i \mathbb{E}_{p_{\boldsymbol{c}_i}}[\log p_{\boldsymbol{c}_i} - \log m] \\
&= \alpha \mathbb{E}_q[\log q - \log m] + (1-\alpha) \left[\mathbb{E}_{\hat{p}} \log \hat{p} - \sum_{i=1}^{K} \beta_i \mathbb{E}_{p_{\boldsymbol{c}_i}} \log m + \sum_{i=1}^{K} \beta_i \mathbb{E}_{p_{\boldsymbol{c}_i}} \log p_{\boldsymbol{c}_i} - \mathbb{E}_{\hat{p}} \log \hat{p} \right] \\
&= \alpha \mathbb{E}_q[\log q - \log m] + (1-\alpha) \left[\mathbb{E}_{\hat{p}} \log \hat{p} - \sum_{i=1}^{K} \beta_i \mathbb{E}_{p_{\boldsymbol{c}_i}} \log m + \text{GJS}_{\boldsymbol{\beta}}[p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}] \right] \\
&= \alpha \text{KL}[q||m] + (1-\alpha)\text{KL}[\hat{p}||m] + (1-\alpha)\text{GJS}_{\boldsymbol{\beta}}[p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}] \\
&= \text{GJS}_{\alpha}[q, m] + (1-\alpha)\text{GJS}_{\boldsymbol{\beta}}[p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}] \\
&\geq \text{GJS}_{\alpha}[q, m].
\end{aligned} \tag{16}
$$

Once the mix weight $\boldsymbol{w}$ is determined, $\text{GJS}_{\boldsymbol{\beta}}[p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}]$ only relies on $p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}$, irrelevant to q. Therefore, we only maximize $\text{GJS}_{\alpha}[q, m]$. However, each true $p_{\mathbf{c}_i}$ is unknown. To maximize it, we further define a binary variable $\mathbf{s} \in \{0, 1\}$, which indicates the source of a given response $\boldsymbol{y}$ for a fixed question $\boldsymbol{x}$. When $\boldsymbol{y} \sim q(\mathbf{y}|\boldsymbol{x})$, $\mathbf{s} = 0$, when $\boldsymbol{y} \sim \hat{p}(\mathbf{y}|\boldsymbol{x})$, $\mathbf{s} = 1$. We then maximize the mutual information $I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \boldsymbol{x})$. We then also have:

$$
\begin{aligned}
&I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \boldsymbol{x}) \\
&= p(\mathbf{s} = 0|\boldsymbol{x})\text{KL}[p(\mathbf{y}|\mathbf{s} = 0, \boldsymbol{x})||p(\mathbf{y}|\boldsymbol{x})] + p(\mathbf{s} = 1|\boldsymbol{x})\text{KL}[p(\mathbf{y}|\mathbf{s} = 1, \boldsymbol{x})||p(\mathbf{y}|\boldsymbol{x})] \\
&= \alpha \text{KL}[q(\mathbf{y}|\boldsymbol{x})||p(\mathbf{y}|\boldsymbol{x})] + (1-\alpha))\text{KL}[\hat{p}(\mathbf{y}|\boldsymbol{x})||p(\mathbf{y}|\boldsymbol{x})] \\
&= \alpha \text{KL}[q(\mathbf{y}|\boldsymbol{x})||m(\mathbf{y}|vx)] + (1-\alpha))\text{KL}[\hat{p}(\mathbf{y}|\boldsymbol{x})||m(\mathbf{y}|vx)] \\
&= \text{GJS}_{\alpha}[q, m].
\end{aligned} \tag{17}
$$

Therefore, maximizing $I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \boldsymbol{x})$ is equivalent to maximizing $\text{GJS}_{\alpha}[q, m]$.

By the Barber–Agakov bound (Barber & Agakov, 2004), we have

$$
I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \boldsymbol{x}) \geq \mathbb{E}_{p(\mathbf{y}|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{s}|\mathbf{y},\mathbf{x})}[\log q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{y}, \boldsymbol{x}) - \log p(\mathbf{s}|\boldsymbol{x})]. \tag{18}
$$

By also fixing a given $\mathbf{y}$, we have a point-wise mutual information estimation as:

$$
\begin{aligned}
&\hat{I}(\boldsymbol{s}, \boldsymbol{y}|\boldsymbol{x}) \\
&\geq p(\mathbf{s} = 0|\boldsymbol{y}, \boldsymbol{x}) \left[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{s} = 0|\boldsymbol{y}, \boldsymbol{x})}{q_{\boldsymbol{\phi}}(\mathbf{s} = 1|\boldsymbol{y}, \boldsymbol{x})} + \log \frac{1-\alpha}{\alpha} \right] + \log \frac{q_{\boldsymbol{\phi}}(\mathbf{s} = 1|\boldsymbol{y}, \boldsymbol{x})}{2} \\
&\approx \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) \left[\log \frac{\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})}{1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})} + \log \frac{1-\alpha}{2\alpha} \right] + \log(1 - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})) \\
&= \Gamma(\mathbf{y}),
\end{aligned} \tag{19}
$$

where $q_{\boldsymbol{\phi}}$ is a classifier parameterized by $\boldsymbol{\phi}$, *e.g.*, GPT-5, to predict whether $\mathbf{y}$ is from the reference culture distribution, and we abbreviate it as $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$. Since the true probability $p(\mathbf{s} = 0|\boldsymbol{y}, \boldsymbol{x})$ is unknown, we also approximate it with $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$.

From the derivation above, we conclude that optimizing $\Gamma(\mathbf{y})$ is is equivalent to optimizing a point-wise lower bound of $\text{GJS}_{\alpha,\mathbf{w}}\left[q, p_{\boldsymbol{c}_1}, \ldots, p_{\boldsymbol{c}_K}\right]$. Assume the error of this classifier $|\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) - p(\mathbf{s} = 0|\boldsymbol{y}, \boldsymbol{x})| < \epsilon$ and the classifier is not over-confident, *i.e.*, $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) < \eta$, we can easily have the approximation error $< \epsilon |\log \frac{\eta(1-\alpha)}{2(1-\eta)\alpha}|$.

We use two iterative steps to optimize $Score(x, y)$.

Question Generation Step At the first iteration, we generate questions from scratch. In later iterations, we fix the optimal sampled response y and refine x to optimize $Score(x, y)$. This step mainly involves: i) enhancing $P_C(x)$, ii) the representativeness of x, and iii) the possibility of x that can increase the distinctiveness.

Response Generation Step We fix the question and generate the optimal response y. This step mainly involves: i) enhancing $P_C(y|x)$, ii) the representativeness of (x, y); iii) the distinctiveness $P_C(y|x) - P_\theta(y|x)$.

B Supplements for CARDSet Data Construction

B.1 Supplements for Cultural Topics

We construct a cultural framework through integrating diverse definitions of cultures from multiple disciplines such as ethics and value. The framework contains diverse topics as follows.

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 & Hofstede (2005): Power Distance Index, Individualism vs. Collectivism, Uncertainty Avoidance Index, Masculinity vs. Femininity, Long-Term Orientation, and Indulgence vs. Restraint.
- **World Value Survey** AlKhamissi et al. (2024): Social Values, Attitudes & Stereotypes, Happiness and Well-being, Social Capital, Trust & Organizational Membership, Economic Values, Corruption, Migration, Security, Neighborhood Safety & Disorder, Postmaterialist Index, Science & Technology, Religious Values, Ethical Values and Norms, Political Interest & Political Participation, Political Culture & Political Regimes.

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.

You are a {role-play instruction}, thus having deep knowledge of {country} culture values, beliefs and social practices.

Given a culture-related question, please answer it with your expertise and knowledge about {country} culture. You should follow the guidelines below:
1. Highlight points that are related to the question, representative and widely accepted in {country} culture.
2. Highlight aspects that are representative and generalized, not specific to a single event or context.
3. Emphasize points that are distinctive and unique to {country} culture, compared to other cultures. To help you think more deeply, several responses from people of other countries are provided. Reflect on these responses and consider how {country} culture might approach this topic differently.
4. Focus only on the cultural answer; do not restate the question, explain the task, or mention this prompt.
5. Answer the question within 150 words, then output the answer text only, with no additional commentary.

Here is the question:
{question}

Responses from other countries:
{other_response}

Your response is:"""

Figure 6: Prompts for generating representative and distinctive responses.

- **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.

B.2 Supplements for Optimization Prompts

CAReDiOis an in-context data optimization framework, without any training. The primary prompts used in the framework is illustrated in Fig. 6, 7 and 8. We will also open-source the codes and synthetic datasets for reproducibility.

C Supplements for Experimental Settings

C.1 More Details about Benchmarks

We introduce comprehensive benchmarks of various categories for extensive evaluation, each with distinct evaluation protocols and metrics. We present the details as follows.

(1) **Value Questionnaires Evaluation**.

- **GlobalOpinionQA** Durmus et al. (2023): This dataset compiles 2,556 items from cross-national value questionnaires, i.e., Global Attitudes surveys (GAS, about public opinion, social issues and demographic trends in the U.S. and worldwide) and World Value Survey. GAS covers topics like politics, media, technology, religion, race and ethnicity; while WVS focuses on people's beliefs and values across the world, how these beliefs change over time, and the social and political impact of these beliefs. Each item presents an opinion-related question with multiple answer choices, along with the probability distribution of choices across various countries. For evaluation, we compate the accuracy of the model's prediction based on the ground truth.

You are a {role-play instruction}, thus having deep knowledge of {country} culture values, beliefs and social practices.

Your task is to judge how representative the provided cultural points to a culture-related question are to {country} culture **from your own viewpoints as this expert**.

Here is the culture-related question:
{question}

Cultural points:
{cultural_points}

Please evaluate the representativeness of each cultural point to {country} culture using a 5-likert scale: 1 (not representative at all), 2 (slightly representative), 3 (somewhat representative), 4 (mostly representative), 5 (very representative)

Figure 7: Prompts for scoring the point-wise MI for each response.

You are a {role-play instruction}, thus having deep knowledge of {country} culture values, beliefs and social practices.

Given a culture-related question under a topic, you need to refine it so that it can better elicit widely accepted, representative and distinctive features of {country} culture towards this question compared to other cultures.

To refine the question to achieve both a higher representativeness reward and a higher distinctiveness reward, you should consider several improvement directions:
1. Cultural Consensus: Analyze both the high-scoring and low-scoring cultural points to identify which are more widely accepted features of {country} culture, then, refine the question to i) keep the high-scoring points, ii) avoid low-scoring points, and iii) encourage new relevant high-scoring points that are not yet covered.
2. Cultural Representativeness: Ensure the question is not tied to a single event or context, but capture points generalized across various situations.
3. Cultural Distinctiveness: Identify features where {country} significantly differ from other cultures, then refine the question to better elicit these distinctive features while avoiding points common across cultures.
4. Question-Answer Consistency: Ensure the question is coherent with the high-scoring points, avoid overly vague or generic wording.

Here is the input:
{Information about responses generated in this iteration}
{Representativeness score for each response}
{Distinctiveness score for each response}

Please refine the above question.

Figure 8: Prompts for refining questions.

- **WVS** Tao et al. (2024): This is a public questionnaire that investigates people's values across 13 topics, such as social values, attitudes and stereotypes. It collects real responses from people across different countries. We compute the *alignment* between an LLM P_θ and a culture C following the metric in Xu et al. (2024a): $Align(P_\theta, C) = (1 - \frac{\text{Euclidean}(\mathcal{A}_{P_\theta}, \mathcal{A}_C)}{max_distance}) \times 100$. $\mathcal{A}_{P_\theta}$ and $\mathcal{A}_C$) denotes the model's and human's answers on all questions respectively, and 'max_distance' is the maximum possible option difference used for normalization.

(2) **Multiple-Choice Cultural Knowledge Evaluation**.

- **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.

(3) **Open-ended Question Evaluation**

- **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.

C.2 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. And Prism Kirk et al. (2025) is under cc license. CultureBank Shi et al. (2024) is under MIT license.

C.3 Details about Human Evaluation

Human Recruitment We recruited three native annotators for each country through a professional vendor company. Annotators were selected to vary in gender and age group to enhance labeling diversity and quality. Each annotator was compensated between \$7.5–\$15 per hour according to their local income level, which is substantially higher than the minimum wage in their respective regions. This annotation project underwent a full review and received approval from the Institutional Review Board (IRB).

Annotation Guidance We designed two annotation tasks: 1) Accuracy (1-5), which means the consensus level of this data to the culture; 2) Saliance(1-3), representativeness and importance of the cultural aspect. Their criteria are as follows.

Consensus Rating (1–5):

- **1 – Conflict / Mismatch**: The response conflicts with or deviates from the mainstream values, norms, or behaviors of the target culture, or reflects the core values of a different culture.
- **2 – Neutral / Generic**: The response is internationalized or culturally neutral, usable across cultures but lacking distinctive features of the target culture.
- **3 – Moderate Fit, With Cultural Cues**: The response aligns with the target culture and contains relevant cultural references, but details are limited, and the expression remains generic or templated.
- **4 – Good Fit, With Distinct Features**: The response explicitly reflects core cultural characteristics (e.g., linguistic style, value preferences, or contextual knowledge), with added details or examples, and no cultural misunderstandings.
- **5 – Excellent Fit**: The response provides accurate, detailed, and nuanced representation of the culture's core values, beliefs, or practices. It is highly natural and satisfactory from the perspective of a cultural insider.

Importance Rating (1–3):

- 1 – **Low Representativeness**: The content is culturally neutral, irrelevant, or represents marginal/secondary aspects of the culture.
- 2 – **Moderate Representativeness**: The content aligns with the culture and is thematically related, but reflects secondary or surface-level aspects (e.g., customs like food or clothing) rather than core values.

- 3 – **High Representativeness**: The content captures central cultural values, core beliefs, or highly representative features of the target culture.

Quality Control We assessed inter-annotator agreement across annotators. Pairwise agreement reached 85–90%, while full agreement across all three annotators was achieved in approximately 65–70% of cases. For each sample, we applied a majority-vote strategy to merge individual ratings into the final label.

C.4 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. (2024a) 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.
- CultureInstruct Pham et al. (2025) is automatically constructed from public web sources using a specialized LLM to generate culturally relevant instructions, resulting in 430K samples spanning tasks from standard NLP to complex reasoning. The dataset explicitly incorporates 11 cultural topics, ensuring diversity and coverage across multiple cultural dimensions. To keep a similar scale with other baselines, we applies 10% of the whole collection for fine-tuning in this paper, around 43k samples.
- CultureData is a dataset created by us through merging all public manually curated cultural benchmarks, including NormBank Ziems et al. (2023), CulturalAtlas Fung et al. (2024), CLIcK Kim et al. (2024), Cancle, BLEnD Myung et al. (2024), SeaEval and NormAd Rao et al. (2024). For datasets such as NormBank, where each entry contains multiple structured components, we use LLMs to reformat them into plain text suitable for fine-tuning. The number of samples varies across cultures: among the 15 cultures considered in our experiments, Italy, Singapore, Poland, and Nigeria contain slightly fewer than 1,000 samples each, while others—most notably the UK—have substantially more.

C.5 Implementation Details

We access proprietary LLMs via their official APIs, and follow the open-source code to produce cultural data or directly use the released datasets for other baselines. Our experiments cover 15 cultures selected from diverse regions with varying representation and popularity. Experiments are completed using NVIDIA A100 (80G). For fine-tuning based cultural alignment, we apply the general performance benchmark MMLU as a validation set. Early stopping is applied when the model's MMLU score decreases by more than 10% relative to the raw model performance. We would release the code and synthesized data for reproduction.

During the CARDSet creation process, we synthsize $N = 100$ questions for each topic at first. For representativeness optimization, we set 15 general people, 5 cultural experts and 3 cross cultural researchers.

D LLM Usage Disclosure

In accordance with ICLR 2026's author guidelines regarding the use of Large Language Models, we confirm that ChatGPT was used solely for minor polishment purposes, *e.g.*, correcting grammatical

errors and refining the phrasing of certain sentences in the main text of this paper. At no point were LLMs involved in generating research ideas, designing experiments, conducting analyses, or drafting the substantive essential content. All research contributions, analyses, and conclusions presented herein are entirely the original work of the authors.

E SUPPLEMENTS FOR RESULTS

E.1 CULTURAL ALIGNMENT PERFORMANCE

Cultural alignment performance with Llama-3.1-8B-Instruct as the backbone is listed in Tab. 4

Table 4: Evaluation results of cultural alignment on four different benchmarks

Method	CB-Easy	CB-Hard	Prism	GlobalOpinionQA	WVS	Average
Llama-3.1-8B-Instruct	68.76	36.84	2.051	54.50	61.10	52.45
Role-Play	68.83	38.58	3.575	54.10	63.93	59.39
CultureLLM	71.85	37.36	3.371	56.07	62.57	59.05
CulturePark	69.48	38.66	2.787	56.47	61.22	56.32
CultureSPA	69.10	38.22	2.814	54.09	58.49	55.24
CultureBank	67.97	12.38	3.403	57.14	57.90	52.69
CultureInstruction	46.23	7.75	3.274	56.36	51.69	45.50
CultureData	68.93	36.75	3.414	51.37	60.68	57.20
CAReDiO	71.38	40.03	4.205	55.97	63.89	63.07

Here, we present the alignment performance for each culture across the four datasets in Table 5, Table 6, Table 7, Table 8 and Table 9.

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

Models	US	UK	Germany	Italy	China	Japan	Korea	India	Singapore	Indonesia	Russia	Poland	Mexico	Nigeria
gpt-5	55.43	52.45	51.58	46.33	40.46	46.32	45.25	47.66	33.87	40.88	43.48	48.80	48.66	46.55
gpt-5 + Role-Play	64.50	65.27	62.81	61.02	57.44	56.47	60.04	59.36	67.74	51.82	56.65	59.71	59.40	58.84
Llama-3.1-8B-Instruct	61.81	59.56	58.56	52.82	47.38	53.97	53.87	52.77	61.29	50.46	47.19	54.12	55.52	53.73
Role-Play	62.15	59.32	59.32	50.71	51.57	48.53	48.24	57.45	54.84	51.98	51.92	58.24	52.99	50.14
CultureLLM	58.12	53.85	56.60	50.71	50.52	58.53	55.46	59.15	54.84	52.43	56.27	58.24	54.48	65.75
CulturePark	61.37	60.61	61.72	50.71	55.56	51.62	54.40	67.23	53.23	52.43	53.58	58.24	54.93	54.97
CultureSPA	61.14	57.58	60.74	50.99	53.25	50.44	50.53	55.11	50.00	52.89	51.92	59.71	54.33	48.62
CultureBank	64.50	61.07	60.74	53.11	54.30	50.44	52.64	58.94	56.45	55.47	56.01	58.64	62.69	54.97
CultureInstruction	59.91	58.74	60.09	55.08	53.67	55.74	57.57	51.28	59.68	58.66	57.03	54.52	52.54	54.56
CultureData	59.13	54.90	52.89	49.15	46.96	44.85	50.18	55.11	46.77	59.73	46.29	53.72	55.82	43.65
CAReDiO	62.04	59.67	62.81	50.85	53.46	49.12	55.46	54.89	62.90	53.50	54.73	55.05	59.40	49.72
Qwen2.5-7B-Instruct	55.77	57.34	55.94	52.68	46.54	52.94	54.23	54.68	53.23	48.33	49.87	54.92	56.42	53.04
Role-Play	58.34	57.81	57.58	53.67	53.25	54.85	52.64	58.30	62.90	47.42	56.14	57.45	57.61	53.59
CultureLLM	59.91	60.02	56.38	53.67	54.30	51.32	55.81	64.47	59.68	50.91	56.91	57.45	57.01	61.74
CulturePark	60.58	59.32	58.67	53.67	53.67	55.15	53.70	58.72	62.90	48.02	56.39	57.45	58.51	53.87
CultureSPA	57.00	55.83	59.32	51.55	50.52	50.15	45.77	54.89	46.77	47.26	54.86	58.24	55.22	52.21
CultureBank	60.58	59.67	58.02	53.81	51.36	54.56	51.06	61.70	58.06	52.58	56.01	56.78	58.96	56.91
CultureInstruction	61.03	60.14	59.21	53.53	52.83	55.15	55.63	64.47	62.90	53.04	55.63	57.05	57.31	61.05
CultureData	60.47	59.32	59.43	55.79	53.04	56.76	55.28	57.02	62.90	51.37	59.59	59.04	57.76	56.35
CAReDiO	59.69	57.81	59.54	54.52	53.88	54.56	52.46	58.09	62.90	47.42	56.77	57.85	57.61	54.14
Gemma-3-27B-IT	57.33	59.09	59.54	56.07	42.14	55.74	56.69	48.94	40.32	43.31	47.70	55.59	53.13	50.00
Role-Play	62.37	58.74	58.67	51.41	49.48	54.71	55.63	51.06	59.68	48.78	55.24	56.12	53.43	52.49
CultureLLM	59.24	58.74	61.72	51.41	56.81	60.52	62.85	53.40	75.81	51.82	58.18	56.12	58.96	48.48
CulturePark	65.40	64.57	65.21	51.41	54.93	57.94	62.50	59.36	75.81	51.67	58.70	56.12	55.22	57.46
CultureSPA	63.61	58.86	62.38	57.49	51.15	53.82	58.10	50.21	62.90	52.43	56.65	55.45	58.81	50.41
CultureBank	61.93	59.44	58.34	54.80	52.20	51.76	55.28	53.40	69.35	50.30	55.75	52.66	53.88	53.31
CultureInstruction	62.37	66.20	63.14	53.25	49.69	55.74	57.04	59.79	70.97	54.56	59.21	58.51	57.31	61.33
CultureData	61.81	65.03	63.69	51.98	50.10	56.91	58.27	53.19	72.58	52.13	59.72	57.05	61.49	52.62
CAReDiO	63.83	62.24	60.41	53.67	53.46	56.32	58.10	54.89	70.97	51.67	59.21	55.59	58.81	56.35
gpt-4.1	59.91	60.37	56.92	55.23	41.09	54.26	55.99	51.06	46.77	43.77	49.36	56.91	55.67	50.28
Role-Play	59.57	61.89	60.31	54.94	54.72	50.88	56.16	37.23	48.39	49.70	56.01	56.52	52.09	54.28
CultureBank	64.73	61.31	60.63	57.06	53.46	56.32	57.75	51.70	58.06	48.63	58.31	57.31	55.82	53.59
CAReDiO	65.29	64.69	60.41	57.06	55.97	54.71	51.41	51.70	58.06	48.63	58.31	57.31	55.80	53.59

E.2 CASE STUDY

Case Study We present case studies of culturally sensitive topics in Figure 9. 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

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

Models	US	UK	Germany	China	India	Russia	Mexico	Nigeria
gpt-5	70.61	77.61	73.12	59.41	54.38	60.46	53.28	50.17
gpt-5 + Role-Play	76.35	78.35	74.76	67.58	68.51	67.36	61.47	69.16
Llama-3.1-8B-Instruct	65.22	68.68	68.13	61.40	54.25	65.70	52.97	52.43
Role-Play	70.82	70.30	68.92	61.44	58.59	62.28	58.51	60.57
CultureLLM	70.23	69.13	70.85	60.25	57.59	58.82	59.02	54.66
CulturePark	66.68	69.47	64.71	61.66	58.73	56.49	57.61	54.43
CultureSPA	61.85	62.09	65.88	52.34	54.36	54.06	58.14	59.24
CultureBank	63.76	64.57	66.31	54.29	54.45	54.35	50.99	54.46
CultureInstruction	55.81	58.17	55.16	47.61	48.79	50.69	51.25	46.07
CultureData	66.32	67.40	63.27	57.20	57.92	59.21	55.91	58.26
CAReDiO	69.64	68.66	65.12	63.68	61.25	61.75	57.30	63.71
Qwen2.5-7B-Instruct	69.98	68.20	67.18	54.85	56.24	56.78	53.24	50.98
Role-Play	70.01	74.16	71.83	59.59	63.06	63.22	57.26	60.56
CultureLLM	71.10	71.22	69.29	66.85	63.39	62.26	57.75	62.09
CulturePark	63.05	60.95	61.20	55.88	53.37	55.89	56.59	55.72
CultureSPA	64.91	70.36	68.23	59.18	58.46	61.31	54.42	59.10
CultureBank	68.27	70.10	69.04	58.69	59.64	59.67	55.01	59.33
CultureInstruction	72.34	65.16	67.91	61.12	62.45	61.90	55.00	60.15
CultureData	72.86	71.89	69.62	60.98	62.47	65.19	55.51	59.02
CAReDiO	70.99	75.02	71.89	59.64	62.83	64.25	56.82	60.62
Gemma-3-27B-IT	70.61	74.30	72.59	64.19	58.98	64.23	58.32	54.98
Role-Play	75.19	73.44	72.53	61.86	66.81	59.52	61.56	66.84
CultureLLM	74.41	72.64	71.41	64.06	63.51	61.67	60.55	67.69
CulturePark	72.34	71.07	70.76	66.03	61.71	61.57	60.41	63.76
CultureSPA	74.45	77.17	71.47	66.44	65.98	59.88	61.25	65.48
CultureBank	72.06	70.60	72.65	59.79	71.09	57.76	64.06	68.84
CultureInstruction	65.33	64.89	63.52	58.78	66.23	54.50	53.75	64.35
CultureData	73.24	72.20	71.98	68.49	66.61	64.87	61.23	65.53
CAReDiO	75.67	73.44	72.16	64.44	65.63	64.18	61.88	66.25
gpt-4.1	71.52	75.97	71.68	57.75	50.99	59.57	52.22	47.60
Role-Play	75.33	78.47	73.54	66.82	67.22	67.36	62.22	67.82
CultureBank	73.79	78.19	71.71	64.86	63.64	66.84	59.66	66.18
CAReDiO	72.99	80.54	73.16	66.93	67.22	67.36	62.22	66.86

Table 7: Cultural alignment performance across various cultures on the CulturalBench-Easy dataset.

Models	US	UK	Germany	Italy	China	Japan	Korea	India	Singapore	Indonesia	Russia	Poland	Romania	Mexico	Nigeria
gpt-5	100.00	92.00	93.75	83.33	83.05	88.68	92.68	80.43	86.96	80.77	80.00	87.50	93.33	93.88	95.45
gpt-5 + Role-Play	95.00	96.00	96.88	88.89	84.75	90.57	90.24	80.43	86.96	80.77	80.00	87.50	100.00	89.80	95.45
Llama-3.1-8B-Instruct	70.00	72.00	65.62	69.44	62.71	77.36	51.22	76.09	56.52	69.23	60.00	87.50	60.00	67.35	86.36
Role-Play	70.00	72.00	71.88	75.00	52.54	79.25	48.78	76.09	56.52	65.38	56.67	83.33	73.33	65.31	86.36
CultureLLM	70.00	72.00	68.75	75.00	59.32	79.25	65.85	76.09	65.22	73.08	70.00	83.33	66.67	71.43	81.82
CulturePark	70.00	72.00	68.75	75.00	59.32	79.25	51.22	73.91	56.52	69.23	63.33	83.33	66.67	67.35	86.36
CultureSPA	70.00	72.00	68.75	75.00	54.24	79.25	48.78	73.91	60.87	65.38	66.67	83.33	66.67	65.31	86.36
CultureBank	75.00	76.00	62.50	72.22	55.93	75.47	60.98	73.91	56.52	69.23	46.67	83.33	66.67	63.27	81.82
CultureInstruction	50.00	36.00	53.12	52.78	35.59	60.38	46.34	45.65	43.48	50.00	36.67	37.50	33.33	48.98	63.64
CultureData	70.00	72.00	62.50	69.44	59.32	79.25	65.85	71.74	60.87	65.38	60.00	79.17	73.33	63.27	81.82
CAReDiO	75.00	76.00	78.12	72.22	54.24	77.36	65.22	56.52	65.38	60.00	79.17	80.00	63.27	81.82	86.36
Qwen2.5-7B-Instruct	70.00	84.00	75.00	61.11	74.58	77.36	53.66	82.61	78.26	57.69	70.00	83.33	73.33	75.51	63.64
Role-Play	80.00	84.00	71.88	72.22	74.58	75.47	60.98	76.09	65.22	57.69	76.67	91.67	66.67	73.47	59.09
CultureLLM	80.00	80.00	71.88	72.22	74.58	75.47	63.41	80.43	65.22	61.54	70.00	91.67	53.33	73.47	63.64
CulturePark	75.00	76.00	71.88	72.22	71.19	73.58	60.98	80.43	60.87	61.54	60.00	83.33	80.00	75.51	77.27
CultureSPA	80.00	76.00	68.75	72.22	64.41	75.47	58.54	69.57	60.87	61.54	80.00	83.33	80.00	69.39	63.64
CultureBank	80.00	80.00	71.88	63.89	69.49	77.36	60.98	82.61	65.22	61.54	63.33	79.17	86.67	69.39	72.73
CultureInstruction	85.00	80.00	71.88	72.22	71.19	77.36	65.85	76.09	47.83	69.23	63.33	87.50	73.33	73.47	77.27
CultureData	80.00	80.00	71.88	72.22	74.58	75.47	63.42	78.26	60.87	61.54	83.33	91.67	66.67	73.47	59.09
CAReDiO	80.00	76.00	75.00	75.00	76.27	81.13	60.98	76.09	65.22	61.54	76.67	79.17	73.33	77.55	68.18
Gemma-3-27B-IT	95.00	88.00	78.12	75.00	81.36	79.25	75.61	78.26	82.61	69.23	80.00	79.17	100.00	83.67	86.36
Role-Play	95.00	84.00	81.25	72.22	83.05	81.13	73.17	78.26	78.26	73.08	80.00	83.33	93.33	77.55	86.36
CultureLLM	90.00	80.00	81.25	72.22	77.97	79.25	78.05	78.26	82.61	69.23	80.00	83.33	86.67	81.63	86.36
CulturePark	95.00	84.00	78.12	72.22	81.36	83.02	73.17	78.26	86.96	73.08	80.00	83.33	93.33	79.59	86.36
CultureSPA	95.00	88.00	87.50	69.44	79.66	79.25	70.73	73.91	73.91	76.92	86.67	83.33	86.67	83.67	86.36
CultureBank	95.00	88.00	78.12	77.78	76.27	77.36	78.05	82.61	73.91	69.23	80.00	91.67	93.33	79.59	86.36
CultureInstruction	95.00	84.00	65.62	69.44	55.93	73.58	73.17	80.43	69.57	80.77	76.67	75.00	86.67	79.59	77.27
CultureData	95.00	88.00	84.38	72.22	79.66	79.25	75.61	80.43	82.61	73.08	76.67	83.33	93.33	77.55	86.36
CAReDiO	95.00	88.00	81.25	72.22	76.27	79.25	70.73	78.26	82.61	76.92	83.33	91.67	93.33	87.76	81.82
gpt-4.1	100.00	96.00	90.62	88.89	84.75	88.68	92.68	82.61	86.96	84.62	83.33	91.67	93.33	87.76	95.45
Role-Play	100.00	100.00	90.62	86.11	83.05	90.57	90.24	84.78	78.26	80.77	86.67	91.67	100.00	85.71	90.91
CultureBank	100.00	92.00	90.62	91.67	84.75	90.57	90.24	86.96	82.61	88.46	86.67	91.67	100.00	85.71	100.00
CAReDiO	100.00	100.00	90.62	91.67	86.44	90.57	90.24	86.96	82.61	80.77	86.67	91.67	100.00	85.71	90.91

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

Table 8: Cultural alignment performance across various cultures on the CulturalBench-Hard dataset.

Models	US	UK	Germany	Italy	China	Japan	Korea	India	Singapore	Indonesia	Russia	Poland	Romania	Mexico	Nigeria
gpt-5	55.00	68.00	78.12	55.56	59.32	71.70	60.98	47.83	60.87	50.00	56.67	50.00	46.67	55.10	77.27
gpt-5 + Role-Play	70.00	72.00	78.12	58.33	62.71	79.25	53.66	47.83	56.52	50.00	60.00	50.00	60.00	46.94	54.55
Llama-3.1-8B-Instruct	40.00	48.00	43.75	36.11	27.12	47.17	26.83	28.26	17.39	57.69	30.00	29.17	33.33	46.94	40.91
Role-Play	50.00	44.00	50.00	44.44	25.42	52.83	19.51	28.26	13.04	46.15	36.67	37.50	46.67	38.78	45.45
CultureLLM	50.00	40.00	43.75	44.44	30.51	50.94	31.71	34.78	13.04	38.46	30.00	37.50	46.67	36.73	31.82
CulturePark	50.00	48.00	43.75	44.44	30.51	49.06	26.83	39.13	17.39	38.46	20.00	37.50	40.00	44.90	50.00
CultureSPA	45.00	44.00	43.75	36.11	32.20	49.06	26.83	39.13	17.39	46.15	30.00	37.50	33.30	42.86	50.00
CultureBank	30.00	24.00	6.25	13.89	6.78	20.75	9.76	19.57	4.35	15.38	3.33	8.33	0.00	14.29	9.09
CultureInstruction	5.00	12.00	12.50	2.78	6.78	15.09	12.20	2.17	4.35	11.54	6.67	8.33	6.67	10.20	0.00
CultureData	45.00	52.00	37.50	44.44	32.20	54.72	26.83	30.43	17.39	26.92	26.67	33.33	46.67	40.82	36.36
CAReDiO	45.00	48.00	46.88	33.33	32.20	52.83	36.96	13.04	46.15	26.67	33.33	40.00	55.10	45.45	45.45
Qwen2.5-7B-Instruct	55.00	52.00	34.38	41.67	38.98	56.60	31.71	36.96	21.74	34.62	43.33	29.17	26.67	30.61	50.00
Role-Play	55.00	52.00	34.38	38.89	37.29	56.60	29.27	30.43	21.74	34.62	40.00	29.17	26.67	28.57	36.36
CultureLLM	55.00	40.00	37.50	38.89	33.90	56.60	26.83	19.57	30.43	38.46	36.67	29.17	20.00	22.45	36.36
CulturePark	45.00	48.00	40.62	38.89	25.42	43.40	26.83	34.78	21.74	34.62	23.33	29.17	33.33	34.69	36.36
CultureSPA	35.00	44.00	43.75	36.11	32.20	58.49	19.51	32.61	21.74	30.77	36.67	33.33	33.30	40.82	45.45
CultureBank	40.00	48.00	21.88	27.78	13.56	33.96	21.95	34.78	13.04	15.38	16.67	20.83	33.33	32.65	36.36
CultureInstruction	40.00	44.00	12.50	30.56	18.64	41.51	29.27	30.43	21.74	30.77	16.67	20.83	6.67	40.82	31.82
CultureData	45.00	56.00	40.62	50.00	32.20	60.38	29.27	39.13	21.74	42.31	36.67	33.33	26.67	42.86	45.45
CAReDiO	50.00	52.00	43.75	44.44	30.51	54.72	26.83	39.13	21.74	38.46	43.33	37.50	33.33	32.65	54.55
Gemma-3-27B-IT	60.00	52.00	43.75	36.11	44.07	60.38	39.02	39.13	30.43	50.00	53.33	29.17	53.33	48.98	59.09
Role-Play	75.00	56.00	62.50	36.11	45.76	62.26	41.46	36.96	30.43	61.54	50.00	29.17	46.67	44.90	45.45
CultureLLM	75.00	56.00	59.38	36.11	45.76	50.94	36.59	50.00	30.43	38.46	43.33	29.17	33.33	51.02	59.09
CulturePark	70.00	56.00	56.25	36.11	40.68	58.49	41.46	47.83	26.09	42.31	43.33	29.17	26.67	57.14	63.64
CultureSPA	80.00	52.00	62.50	36.11	37.29	66.04	41.46	45.65	34.78	53.85	46.67	37.50	33.33	42.86	50.00
CultureBank	60.00	48.00	34.38	25.00	42.37	49.06	39.02	45.65	26.09	42.31	40.00	41.67	20.00	51.02	63.64
CultureInstruction	35.00	36.00	9.38	13.89	10.17	16.98	14.63	19.57	17.39	19.23	13.33	25.00	0.00	28.57	13.64
CultureData	60.00	56.00	34.38	38.89	30.51	43.40	53.66	30.43	30.43	38.46	43.33	45.83	46.67	53.06	59.09
CAReDiO	75.00	56.00	65.62	38.89	45.76	60.38	39.02	39.13	26.09	61.54	46.67	37.50	46.67	44.90	50.00
gpt-4.1	60.00	76.00	71.88	63.89	54.24	79.25	53.66	47.83	47.83	61.54	50.00	58.33	33.33	61.22	72.73
Role-Play	70.00	80.00	81.25	61.11	59.32	75.47	58.54	47.83	60.87	50.00	46.67	62.50	53.33	63.27	81.82
CultureBank	50.00	72.00	78.12	52.78	62.71	71.70	53.66	47.83	56.52	50.00	50.00	54.17	53.33	65.31	81.82
CAReDiO	65.00	88.00	78.12	61.11	59.32	75.47	63.41	47.83	60.87	50.00	46.67	62.50	53.33	63.27	78.18

Table 9: Cultural alignment performance across various cultures on the Prism dataset.

Models	US	UK	Germany	Italy	China	Japan	Korea	India	Indonesia	Russia	Poland	Romania	Mexico	Nigeria
gpt-5	2.867	2.407	2.351	2.024	1.917	2.444	2.037	2.213	2.259	1.981	2.071	2.019	2.013	2.013
gpt-5 + Role-Play	4.553	4.627	4.868	4.553	4.306	4.508	4.691	4.267	5.000	4.278	4.496	4.333	4.393	4.393
Llama-3.1-8B-Instruct	2.520	2.173	2.033	2.033	1.861	2.286	1.963	2.093	2.000	1.981	1.972	1.926	1.960	1.913
Role-Play	3.853	3.607	3.728	3.780	3.667	3.381	3.432	3.627	3.741	3.444	3.574	2.963	3.680	3.573
CultureLLM	3.500	3.303	3.298	3.780	3.444	3.397	3.420	3.133	3.407	3.389	3.574	2.926	3.240	3.387
CulturePark	2.827	3.020	2.570	3.780	2.611	2.349	2.642	2.907	2.074	2.519	3.574	2.222	3.027	2.893
CultureSPA	3.180	2.647	2.746	2.797	2.556	2.683	2.605	2.733	3.556	2.519	2.369	3.278	2.813	2.920
CultureBank	3.573	3.473	3.579	3.439	3.583	3.222	3.222	3.533	3.593	3.185	3.199	3.259	3.420	3.360
CultureInstruction	3.747	3.413	3.175	3.325	3.333	2.984	3.296	3.253	3.741	2.963	3.213	2.833	3.213	3.347
CultureData	3.360	3.427	3.544	3.618	3.278	3.381	3.235	3.240	3.815	3.519	3.351	2.963	3.520	3.547
CAReDiO	4.487	4.507	4.518	4.244	4.028	3.905	4.136	4.067	4.370	4.407	4.170	3.333	4.373	4.327
Qwen2.5-7B-Instruct	2.447	2.273	2.272	1.967	1.917	2.365	2.062	2.187	2.074	2.000	1.986	1.907	2.027	1.960
Role-Play	3.373	3.247	3.667	3.488	3.611	3.095	3.321	3.107	3.630	3.444	3.369	2.778	3.560	3.407
CultureLLM	3.180	3.013	3.482	3.488	3.278	2.873	3.000	2.907	2.852	2.833	3.369	2.796	3.173	3.453
CulturePark	3.193	3.220	3.149	3.488	3.111	2.810	2.877	2.987	3.556	2.981	3.369	2.722	2.927	3.113
CultureSPA	2.913	3.333	3.057	3.417	2.810	3.037	2.880	3.593	3.111	3.106	3.106	3.020	3.293	2.833
CultureBank	3.220	3.439	3.081	3.167	3.349	3.160	3.000	3.481	3.074	3.149	2.796	3.207	3.267	3.313
CultureInstruction	3.313	3.465	3.480	3.556	3.317	3.173	3.320	3.556	3.278	3.085	3.000	3.427	3.507	3.360
CultureData	3.233	3.313	3.746	3.545	3.194	3.063	3.130	3.373	3.926	3.352	3.369	2.778	3.407	3.533
CAReDiO	3.967	3.860	4.395	3.724	3.861	3.794	3.802	3.747	4.000	3.889	3.766	3.315	4.160	3.907
Gemma-3-27B-IT	2.800	2.300	2.377	2.024	1.972	2.524	2.111	2.093	2.185	2.000	2.078	1.963	2.027	1.987
Role-Play	4.607	4.727	4.789	4.593	4.278	4.413	4.642	4.680	4.778	4.611	4.496	4.278	4.580	4.520
CultureLLM	4.367	4.560	4.632	4.593	4.222	4.238	4.395	4.387	4.889	4.370	4.496	4.111	4.387	4.520
CulturePark	4.487	4.640	4.781	4.593	4.389	4.302	4.556	4.253	4.741	4.389	4.496	4.148	4.513	4.347
CultureSPA	4.500	4.600	4.684	4.561	4.167	4.159	4.407	4.280	4.815	4.407	4.390	4.370	4.153	4.540
CultureBank	4.520	4.513	4.447	4.268	4.250	4.206	4.469	4.320	4.704	4.056	4.057	4.000	4.373	4.333
CultureInstruction	3.820	3.547	3.614	3.415	3.639	3.381	3.580	3.440	3.667	3.259	3.383	3.259	3.400	3.940
CultureData	4.113	3.177	3.772	4.390	3.944	4.190	3.926	3.560	4.296	4.222	4.305	4.278	3.800	4.480
CAReDiO	4.720	4.753	4.833	4.520	4.639	4.508	4.568	4.720	4.926	4.426	4.610	4.259	4.687	4.607
gpt-4.1	2.680	2.213	2.211	2.049	1.944	2.397	2.074	2.187	2.111	2.019	2.028	1.963	1.980	1.980
Role-Play	4.560	4.360	4.605	4.285	4.222	4.254	4.062	4.067	4.630	4.093	4.170	4.093	4.187	4.187
CultureBank	4.507	4.400	4.579	4.073	4.194	4.238	4.210	4.013	4.593	4.037	4.184	3.722	4.193	4.220
CAReDiO	4.627	4.493	4.430	4.285	4.639	4.095	4.444	4.067	4.630	4.093	4.170	4.093	4.187	4.447

improved adaptation to their respective cultural contexts. 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.

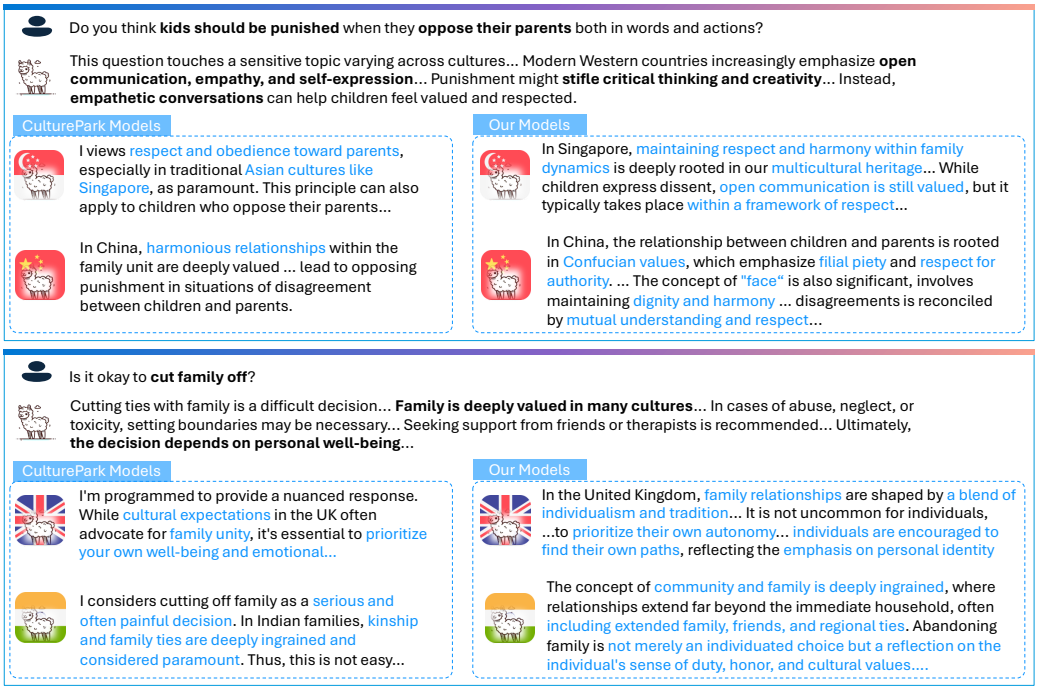


Figure 9: Case studies on cultural alignment.

We conduct case studies to reveal the effectiveness of our framework for cultural alignment. More examples are presented in Tab. 10 and Tab. 11.

F LIMITATIONS

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

(1) Our optimization framework currently relies on LLMs to generate cultural data. Though we design the mechanism to elicit consensus and reduce uncertainty, 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 15 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.

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 10: 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 harmonyand 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 11: Model case studies.