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

Jing Yao^{† 1 2} Xiaoyuan Yi^{* 2} Jindong Wang² Zhicheng Dou¹ Xing Xie²

Abstract

As Large Language Models (LLMs) are deployed globally, aligning them with pluralistic cultures is crucial for improving user engagement and preventing cultural conflicts. Recent work has curated, either synthesized or manually annotated, culture-specific corpora for alignment. Nevertheless, cultural theories suggest that they face two key challenges. (1) *Representativeness*: These corpora inadequately capture a culture’s core characteristics, causing insufficient cultural coverage and redundancy; (2) *Distinctiveness*: They fail to distinguish the unique nuances of a target culture from patterns shared across relevant ones, hindering precise culture modeling. To handle these challenges, we introduce **CAReDiO**, a novel data optimization framework that alternately optimizes culture-sensitive questions and responses according to two information-theoretic objectives in an in-context manner, enhancing both cultural representativeness and distinctiveness of constructed data. Extensive experiments across 15 cultures demonstrate that CAReDiO can create high-quality data with richer cultural nuances and enable efficient alignment for both small open-source or large proprietary LLMs with as few as 200 training samples, consistently outperforming previous datasets on both multi-choice and open-ended benchmarks. Our code and dataset are released at [ValueCompass/CAReDiO](#).

1. Introduction

As Large Language Models (LLMs) are widely integrated into human life (OpenAI, 2024; Dubey et al., 2024; Guo

[†]Jing Yao’s work was done at MSRA. ¹Renmin University of China, Beijing, China ²Microsoft Research Asia, Beijing, China. jingyao@microsoft.com. Correspondence to: Xiaoyuan Yi <xi-aoyuanyi@microsoft.com>.

et al., 2025), aligning their outputs with human values is imperative to mitigate safety risks and improve user engagement (Ouyang et al., 2022; Wang et al., 2024c). Prior studies, focusing on *universal* preferences, e.g., the HHH principle (Askell et al., 2021; Bai et al., 2022), overlook *cultural diversity* rooted in human values. As a result, current LLMs dominated by English corpus are often biased towards Western-centric opinions and cultures (Cao et al., 2023; Durmus et al., 2023), dissatisfying underrepresented cultural communities and causing unintended social tensions (Ryan et al., 2024). Therefore, *aligning LLMs with diverse cultural values has become both an ethical and practical necessity* (AlKhamissi et al., 2024; Tao et al., 2024).

Early efforts align LLMs with target cultures using an In-Context Learning (ICL) approach, via role-playing instructions or few-shot examples (Durmus et al., 2023; Kwok et al., 2024). While flexible, these methods suffer from unstable performance across tasks (especially for small models), high inference costs and privacy concerns (Saunders et al., 2022). Recently, *fine-tuning culture-aware LLMs has proven a practical alternative* (Li et al., 2024b) using large-scale local-language corpora (Gupta et al., 2023; Nguyen et al., 2024; Pipatanakul et al., 2023), but language alone is insufficient to encode cultural values (Choenni et al., 2024; Mukherjee et al., 2024; Rystrøm et al., 2025). Consequently, growing attention has shifted to building dedicated, culture-specific datasets, either synthesized or manually annotated (Li et al., 2024a;b). However, curating high-quality data demands massive annotation costs and lacks scalability. This raises a key research question: *Which data are more informative to enable cultural alignment at a minimal cost?*

To answer this question, we resort to culture theories, particularly 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. *The emic (internal) view* captures the highly shared beliefs central to a culture, while *the etic (external) view* highlights traits that differentiate one culture from others. However, existing cultural alignment datasets struggle to reflect both views due to two key challenges. **Challenge 1. Representativeness (emic)**: The dataset should prioritize

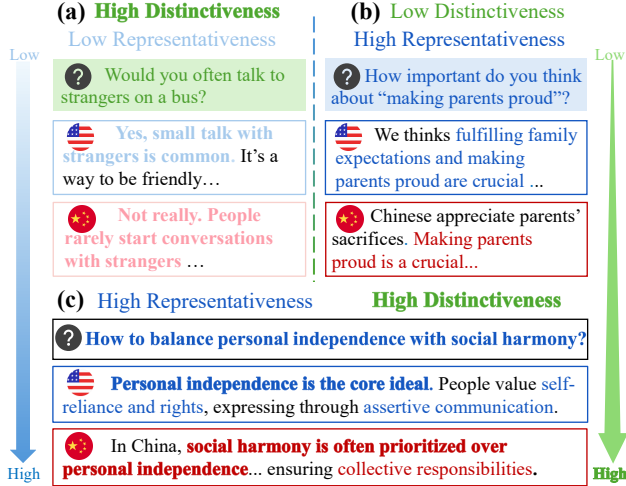


Figure 1. Most cultural data fails to capture both dimensions: (a) high distinctiveness (different answers for China and US) but low representativeness; (b) high representativeness (central cultural topic) but low distinctiveness. (c) A desired case of CAREDiO, scoring high in both dimensions.

samples that reflect the most salient and central aspects of the target culture, rather than less important or redundant cases; **Challenge 2. Distinctiveness (etic)**: The dataset should also capture the unique nuances of the target culture, instead of patterns shared across multiple related cultures (e.g., China, Japan and Korea). As shown in Fig. 1, the two dimensions are not always naturally tied together across samples. Failing to handle these challenges hinders the precise modeling of culture-specific stimuli, hence hurting the efficiency and effectiveness of cultural alignment.

This work proposes **CAREDiO**¹, a novel LLM-empowered in-context data optimization framework for the automatic construction of representative and distinctive cultural data. CAREDiO alternately generates and refines cultural questions and responses to fulfill two information-theoretic objectives: i) an *information gain objective*, inspired by Cultural Consensus Theory (Weller, 2007), to identify data samples that better reduce the LLM’s cultural uncertainty and gain agreement, *improving representativeness*; ii) a *culture divergence objective*, which is grounded in Cognitive Conflict Theory (Limón, 2001) and theoretically performs a point-wise optimization of different cultures’ JS divergence, to enhance sample distinguishability from non-target cultures, *improving distinctiveness*. CAREDiO is model-agnostic and 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 representative and distinctive data for any specific culture (shown in Fig. 2). Experiments show such data better captures culture characteristics and boundaries, enabling

¹Cultural Alignment via **R**epresentativeness and **D**istinctiveness Guided Data **O**ptimization.

more effective and efficient alignment across cultures.

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

2. Related Work

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 social science instruments to probe cultural values, including the World Value Survey (WVS) (AlKhamissi et al., 2024), Hofstede framework (Cao et al., 2023; Masoud et al., 2025; Wang et al., 2024d; Kharchenko et al., 2024; Sukiennik et al., 2025), European Value Survey (EVS) (Tao et al., 2024) and GlobalOpinionQA (Durmus et al., 2023). Recent benchmarks extend these frameworks with open-ended QA data (Karinshak et al., 2024; Banerjee et al., 2025). Besides, other cultural dimensions have been investigated: NORMSAGE (Fung et al., 2023) and NormAd (Rao et al., 2024) for social norms, EtiCor (Dwivedi et al., 2023) for social etiquette and the manually curated CulturalBench (Chiu et al., 2024) across comprehensive domains. In most evaluations, *even advanced LLMs exhibit Western-centric biases (Wang et al., 2024b), underscoring the urgency of cultural alignment*.

Approaches to Cultural Alignment Early efforts focus on In-Context Learning (ICL) alignment (Dong et al., 2024), including prompting LLMs to consider culture-specific perspectives (Durmus et al., 2023), role-playing with demographic attributes (Kwok et al., 2024; Kharchenko et al., 2024) or cultural value descriptions (Choenni & Shutova, 2024) and native language instructions (Durmus et al., 2023; Cao et al., 2023). While flexible, they depend on strong ICL capabilities and prior cultural knowledge, limiting their effectiveness for weaker LLMs (Saunders et al., 2022). Recently, fine-tuning LLMs with cultural datasets (Li et al., 2024a;b) and cultural learning-inspired strategies (Yuan et al., 2024; Liu et al., 2025) has become an effective solution, *highlighting the need for high-quality cultural datasets*.

Datasets for Cultural Alignment Existing cultural datasets fall into four categories, yet *none of them explicitly optimize for or can naturally guarantee both cultural representativeness and distinctiveness*. (1) *Large-scale*

local-language corpora adapt English-centric models regionally (Pires et al., 2023; Nguyen et al., 2024; Pipatanakul et al., 2023; Abbasi et al., 2023). Nonetheless, language data that is not curated for culture provides limited cultural specificity. (2) *Culture-related data filtered from large-scale web data or social media*, such as CRAFT (Wang et al., 2024a), CultureInstruct (Pham et al., 2025), CultureBank (Shi et al., 2024) and CultureAtlas (Fung et al., 2024). While broad in coverage, they representativeness or cross-cultural distinctions, leading to potential data redundancy. (3) *Manually curated datasets* emphasize data quality but are costly to scale, including NORMSAGE (Fung et al., 2023) 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 commonsense (Nguyen et al., 2023), and WVS survey for values (Haerpfer et al., 2020). In practice, achieving both properties requires expert annotations, while existing benchmarks often rely on crowd-sourcing with limited quality. (4) *LLM-synthesized datasets* (Yuan et al., 2024). CultureLLM (Li et al., 2024a) and CulturePark (Li et al., 2024b) augment WVS with model-generated opinions. CultureSPA (Xu et al., 2025) synthesizes WVS-style questions with shifted answers under culture-unaware and -aware settings. These datasets heavily rely on pre-defined surveys that capture important topics but lack explicit mechanisms to enforce cross-cultural distinctiveness.

In summary, *prior datasets lack explicit formulation and optimization for representativeness and distinctiveness, limiting data efficiency and effectiveness for cultural alignment.*

3. Methodology

3.1. Formalization and Overview

Denote $p_{\theta}(\mathbf{y}|\mathbf{x})$ as an LLM parameterized by θ , which generates a response $\mathbf{y}$ to a given question $\mathbf{x}$; $p_{c_1}(\mathbf{x}, \mathbf{y}), \dots, p_{c_{K+1}}(\mathbf{x}, \mathbf{y})$ as the true distributions of $K+1$ different cultures. Given a target culture $\mathbf{c}$, our goal is to construct a set of cultural question-response pairs $q_{\mathbf{c}}^* = \{(\mathbf{x}^*, \mathbf{y}^*)\}$ with satisfactory *representativeness* and *distinctiveness* to achieve effective and sample-efficient cultural alignment of the backbone model p_{θ} . For this purpose, we need to solve the objective below:

$$q_{\mathbf{c}}^* = \underset{(\mathbf{x}, \mathbf{y})}{\operatorname{argtopN}} \{p_{\mathbf{c}}(\mathbf{x}, \mathbf{y}) - \gamma * \frac{1}{K} \sum_{c_k \neq \mathbf{c}} p_{c_k}(\mathbf{x}, \mathbf{y})\}, \quad (1)$$

where γ is a hyperparameter. This objective helps identify data samples $(\mathbf{x}, \mathbf{y})$ that i) have a large probability mass $p_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$, *i.e.*, salient for the target culture $\mathbf{c}$, and ii) have a smaller probability of being shared by the non-target cultures, ensuring both representativeness and distinctiveness.

Nevertheless, *each true culture distribution $p_{c_k}(\mathbf{x}, \mathbf{y})$, $k = 1, \dots, K+1$, is unavailable.* Inspired by culture theo-

ries, we propose CAREDiO, an LLM-empowered in-context framework to approximate and optimize Eq.(1), *without relying on any true cultural distributions.* As shown in Fig. 2, CAREDiO consists of three core components: i) an *information gain objective* to identify samples that maximally reduce p_{θ} 's uncertainty about the target culture and increase the saliency $p_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$, *improving representativeness*; ii) a *culture divergence objective* to encourage samples distinguishable from non-target cultures and decrease $p_{c_k}(\mathbf{x}, \mathbf{y})$, $c_k \neq \mathbf{c}$, *improving distinctiveness*; and iii) an *iterative schema* that alternately refines the question $\mathbf{x}$ and corresponding response $\mathbf{y}$ to optimize the two objectives until convergence, leading to informative cultural data. We elaborate on each module in the following subsection.

3.2. The CAREDiO Framework

Representativeness Optimization via Information Gain

To improve *representativeness*, the main difficulty in solving Eq.(1) is that $p_{\mathbf{c}}(\mathbf{y}|\mathbf{x})^2$ is unavailable, thus we can neither sample $\mathbf{y}$ from the true distribution nor obtain the density of $\mathbf{y}$. Fortunately, *Cultural Consensus Theory* (CulCT) (Weller, 2007) from cognitive anthropology and cultural psychology indicates that culturally salient elements correspond to shared cognition among people with cultural competence. Building upon this theory, we approximate representativeness optimization as a problem of *consensus elicitation*.

We incorporate LLMs to mimic a group of culturally competent raters of the target culture, $p_{\omega_1}, \dots, p_{\omega_M}$, using an ICL alignment method such as role-playing (Kwok et al., 2024; Li et al., 2024b). To ensure diversity across raters, each p_{ω_i} could be either a heterogeneous LLM or the same one with different demographic role-plays. Then, for each rater p_{ω} , we quantify how strongly a response $\mathbf{y}$ reflects its cognition of culture $\mathbf{c}$ by the Mutual Information (MI) between $\mathbf{y}$ and $\mathbf{c}$: $I_{\omega}(\mathbf{c}; \mathbf{y}|\mathbf{x} = \mathbf{x}) = \mathbb{E}_{p_{\omega}(\mathbf{y}|\mathbf{x})} \mathbb{E}_{p_{\omega}(\mathbf{c}|\mathbf{y}, \mathbf{x})} [\log p_{\omega}(\mathbf{c}|\mathbf{x}, \mathbf{y}) - \log p_{\omega}(\mathbf{c}|\mathbf{x})]$. Concretely, for a given sample $(\mathbf{x}, \mathbf{y})$, we compute point-wise MI as the cognition rating score of p_{ω} , *i.e.*, $\Delta_{\omega}^{\omega}(\mathbf{y}|\mathbf{x})$:

$$\Delta_{\omega}^{\omega}(\mathbf{y}|\mathbf{x}) = \hat{I}_{\omega}(\mathbf{c}; \mathbf{y}|\mathbf{x}) = \log p_{\omega}(\mathbf{c}|\mathbf{x}, \mathbf{y}) - \log p_{\omega}(\mathbf{c}|\mathbf{x}). \quad (2)$$

For open-sourced LLMs, probabilities are available, while for black-box LLMs, we approximate each term by prompting them to return scores. We then aggregate the scores to measure consensus and find the highest-scoring $\mathbf{y}$:

$$\Delta_{\mathbf{c}}^{\omega_1, \dots, \omega_M}(\mathbf{y}|\mathbf{x}) = \frac{1}{M} \sum_i \Delta_{\omega_i}^{\omega_i}(\mathbf{y}|\mathbf{x}). \quad (3)$$

In this way, for a given question, we identify *representative responses* that reflect the *shared cognition* of the target culture $\mathbf{c}$. Eq.(2) can be regarded as Information-Directed Sampling (IDS) (Hao et al., 2022), identifying responses $\mathbf{y}$

²We conduct a dual refinement process for both $\mathbf{x}$ and $\mathbf{y}$ but only describe the process of $\mathbf{y}$ for brevity.

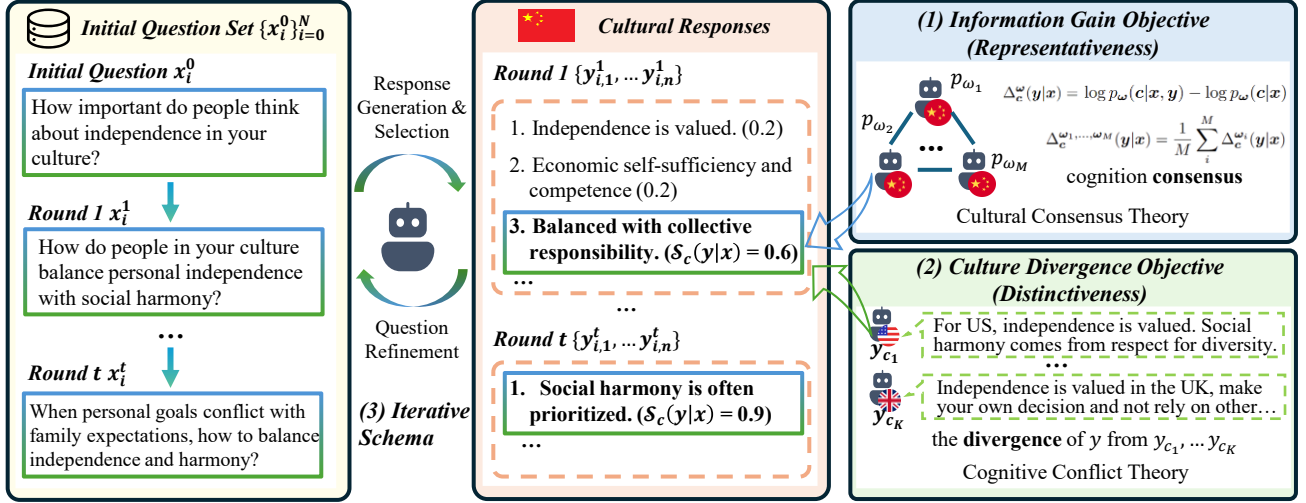


Figure 2. The CAREDiO framework, including two information-theoretic objective modules for representativeness and distinctiveness enhancement, as well as an iterative schema to alternately optimize questions and responses.

that reinforce p_ω 's understanding of culture c . p_ω could be either p_θ (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., 2024).

As the foundation of this representativeness optimization method, we validate in Appendix B.1 that i) LLMs can effectively simulate cultural raters via role-playing and ii) multiple LLM-simulated roles can elicit higher-quality cultural consensus rather than artifact amplification. We then give the following conclusion:

Proposition 3.1. *For two responses y_1, y_2 toward the same x , if their scores in Eq.(2) satisfy $\Delta_c^\omega(y_1|x) > \Delta_c^\omega(y_2|x)$, under mild conditions, using y_1 for fine-tuning leads to a larger gradient in learning the true distribution $p_c(x, y)$ than using y_2 : $\|\nabla_{\theta} l_{\theta}(y_1, x)\| > \|\nabla_{\theta} l_{\theta}(y_2, x)\|$.*

Proof. See Appendix. A.

Distinctiveness Optimization by Culture Divergence

As demonstrated in Fig. 1, highly representative samples do not guarantee distinguishability from other non-target ones, particularly among closely related cultures (e.g., China, Japan and Korea). To ensure *distinctiveness*, we capture cultural boundaries and construct distinguishable y by resorting to Cognitive Conflict Theory (CogCT) (Cosier & Rose, 1977), which suggests cognitive conflicts among cultures can provoke self-reflection on their own culture.

We set the target culture $c = c_{K+1}$, and $c_1, \dots, c_K$ as other non-target ones for convenience, and use the generalized JS divergence (Engleson & Azizpour, 2021) between the data distribution q_c and $p_{c_1}, \dots, p_{c_K}$, i.e., $GJS_{\alpha, w}[q(y|x), p_{c_1}(y|x), \dots, p_{c_K}(y|x)]$, to measure distinctiveness. $\alpha, w = (w_1, \dots, w_K) > 0$ are weights for each distribution with $\alpha + \sum_{i=1}^K w_i = 1$. Nevertheless, the

difficulty in Eq.(1) still exists that each true culture distribution p_{c_k} is unattainable. Therefore, we use the formulation $\Gamma_{c_1, \dots, c_K}(y|x)$ to approximate distinctiveness instead:

$$\Gamma_{c_1, \dots, c_K}(y|x) = \phi(y, x) \left[\log \frac{\phi(y, x)}{1 - \phi(y, x)} + \log \frac{1 - \alpha}{2\alpha} \right] + \log(1 - \phi(y, x)), \quad (4)$$

where $\phi(y, x)^3$ is a classifier to estimate the probability that the response y to x does NOT come from any of the K non-target cultures. As high-quality cultural data for fine-tuning is rare, we implement the classifier using an LLM p_ω and the OpenAI text-embedding-3-small model. Given (x, y) , p_ω generates a response y_c for c and $(y_{c_1}, \dots, y_{c_K})$ for non-target cultures. Then, we calculate $\phi(y, x) = \frac{\exp(\text{sim}(e_y, e_c))}{\text{sim}(e_y, e_c) + \sum_{k=1}^K \text{sim}(e_y, e_{c_k})}$, where e_y, e_c, e_{c_k} are text embeddings. To better ensure distinction, we can also introduce culture-specific LLMs to generate responses y_c and $(y_{c_1}, \dots, y_{c_K})$ for various cultures rather than relying on only p_ω . We validate the accuracy, robustness of the classifier and compare alternative classifier variants like *llm-as-judge* in Appendix B.2 and E.7.

For the above approximation, we provide a conclusion:

Proposition 3.2. *For any given x and y , if the classifier error $|\phi(y, x) - p^*(y \notin p_{c_1}, \dots, p_{c_K}|y, x)| < \epsilon$, and the classifier is not over-confident, i.e., $\phi(y, x) < \eta$, then maximizing Eq.(4) is an approximated point-wise maximization of the lower bound of $GJS_{\alpha, w}[q(y|x), p_{c_1}(y|x), \dots, p_{c_K}(y|x)]$, 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.

³We abbreviate $p_\phi(y \notin p_{c_1}, \dots, p_{c_K}|y, x)$ or $p_\phi(x \notin p_{c_1}, \dots, p_{c_K}|y, x)$ as $\phi(y, x)$.

Therefore, we can use a reliable classifier to score and identify distinctive samples following Eq.(4), which maximizes a lower bound of the true distinctiveness and elicits cultural differences, even without access to real culture distributions. The plausibility of this approximation relies on the classifier error bounds $(\epsilon, \eta, \mathcal{E})$, which we validate in Appendix B.2.

Eventually, we combine the two objectives and use the following score for data optimization:

$$\mathcal{S}_c(\mathbf{y}|\mathbf{x}) = \lambda_1 \cdot \Delta_{c_1, \dots, c_K}^{\omega_1, \dots, \omega_M}(\mathbf{y}|\mathbf{x}) + \lambda_2 \cdot \Gamma_{c_1, \dots, c_K}(\mathbf{y}|\mathbf{x}) \quad (5) \\ + \lambda_3 \cdot \mathbb{E}_{(\mathbf{x}', \mathbf{y}') \sim q_c} [\mathcal{K}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}'))].$$

$\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, and $\mathcal{K}(\cdot, \cdot)$ is cosine similarity over text embeddings to ensure diversity. We only describe response optimization for brevity but conduct a dual process for $\mathcal{S}_c(\mathbf{x}|\mathbf{y})$.

Iterative Optimization Schema Starting from an initial set of cultural questions $\{\mathbf{x}_i^0\}_{i=0}^N$, we alternately optimize the questions $\mathbf{x}$ and responses $\mathbf{y}$ to maximize Eq.(5). We use an LLM p_ω and a classifier $\phi(\mathbf{y}, \mathbf{x})$ (in all experiments, $\theta = \omega$). At the t -th iteration, given the question $\mathbf{x}^{t-1}$, we prompt p_ω with diverse roles of the target culture to generate responses $(\mathbf{y}_1, \dots, \mathbf{y}_n)$. To promote distinctiveness, we also provide p_ω with responses from non-target cultures, $\{\mathbf{y}_{c_1}, \mathbf{y}_{c_2}, \dots, \mathbf{y}_{c_K}\}$, encouraging response with unique nuances of culture c . We score each $\mathbf{y}_i$ using Eq.(5) and select the highest-scoring one as $\mathbf{y}^t$. Given $\mathbf{y}^t$, we provide p_ω with all candidate responses $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ and their scores, then p_ω refines $\mathbf{x}^{t-1}$ to increase the generation probability of representative and distinctive responses and suppress low-scoring ones, improving $\mathcal{S}_c(\mathbf{x}|\mathbf{y})$. To be compatible with black-box LLMs, the entire process is performed via in-context learning without parameter updates, until convergence or early stopping. The full algorithm is in Algo. 1, prompts and implementations in Appendix C.2.

Justification for Our Improvement Since ground-truth cultural distributions are often unavailable, our motivation is to precisely elicit effective cultural data without assessing true distributions. Though relying on LLMs, CAREDiO substantially differs from heuristic or engineering-driven data synthesis. It mitigates rather than amplifies model biases by anchoring the optimization process to two information-theoretic objectives: i) The representativeness objective ensembles diverse perspectives to mitigate biases of single LLMs. ii) The distinctiveness objective distinguishes ambiguous data. This advantage is empirically validated through extensive comparisons with existing cultural data synthesis baselines in Sec. 4.3.

3.3. CARDSet Construction and Alignment

To validate CAREDiO, we instantiate it to construct a cultural alignment dataset **CARDSet**. We initiate questions

Algorithm 1 The CAREDiO Framework

- 1: **Input:** Maximum number of iterations T , the LLM p_ω , the classifier $\phi(\mathbf{y}, \mathbf{x})$, target culture c , non-target culture $c_1, \dots, c_K$
 - 2: **Output:** Optimized data $q_c^* = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=0}^N$
 - 3: Initialize a set of cultural questions $\{\mathbf{x}_i^0\}_{i=0}^N$
 - 4: **for** $t = 1, 2, \dots, T$ **do**
 - 5: **for** $i = 1, 2, \dots, N$ **do**
 - 6: Generate responses $\{\mathbf{y}_{i,1}^t, \dots, \mathbf{y}_{i,n}^t\}$ with p_ω
 - 7: Calculate $\mathcal{S}_c(\mathbf{y}_{i,j}^t|\mathbf{x}_i^{t-1})$ for each $\mathbf{y}_{i,j}^t$ by Eq.(5)
 - 8: Select the highest-scoring one as $\mathbf{y}_i^t$
 - 9: Use p_ω to refine $\mathbf{x}_i^{t-1}$ to $\mathbf{x}_i^t$ with a larger score $\mathcal{S}_c(\mathbf{x}_i^t|\mathbf{y}_i^t)$
 - 10: **end for**
 - 11: **end for**
-

$\{\mathbf{x}_i^0\}_{i=1}^N$ covering core cultural aspects (See Appendix. C.1 for topic details, C.2, C.3 and C.4 for generation details.) To compute consensus score in Eq.(3), we prompt p_ω to role-play a group of culturally competent individuals, including 15 *general people with various demographics*; 3 *cultural experts with different backgrounds*; and 3 *cross-cultural researchers*. We vary p_ω to synthesize multiple versions of CARDSet for comparison experiments in Sec.4.3.

To control training cost and ensure diversity, we rank samples by $\mathcal{S}_c(\mathbf{x}, \mathbf{y})$ and sequentially select them under a pre-defined computational budget, filtering out semantically similar samples using a threshold $\tau = 0.85$. With responses for non-targeted cultures as dispreferred ones, we can fine-tune cultural LLMs via SFT or DPO. For fair comparison, we follow prior work to use SFT in all experiments.

4. Experiments

4.1. Experimental Settings

Benchmarks We consider four complementary cultural benchmarks. (1) **CulturalBench** (Chiu et al., 2024), a manually curated benchmark of 1,227 four-choice questions on cultural knowledge, spanning 45 regions with *Easy* and *Hard* variants, where the *Hard* decomposes each question into four binary judgments and requires all to be answered correctly. *Accuracy* is reported. (2) **Prism** (Kirk et al., 2025) consists of conversations between 1,500 participants across 75 countries and 21 LLMs. We retain value-related questions with i) cultural topics and ii) culturally variant responses. *Response quality* is rated on a 1-5 scale by both Gemini-2.5-Pro and native annotators. (3) **GlobalOpinionQA** (Durmus et al., 2023) compiles items from two surveys. To avoid benchmark overlap, we retain only the Global Attitudes surveys (GAS) subset. Each item consists of a multi-choice question and the choice distributions across countries. We report *accuracy* as whether the model’s pre-

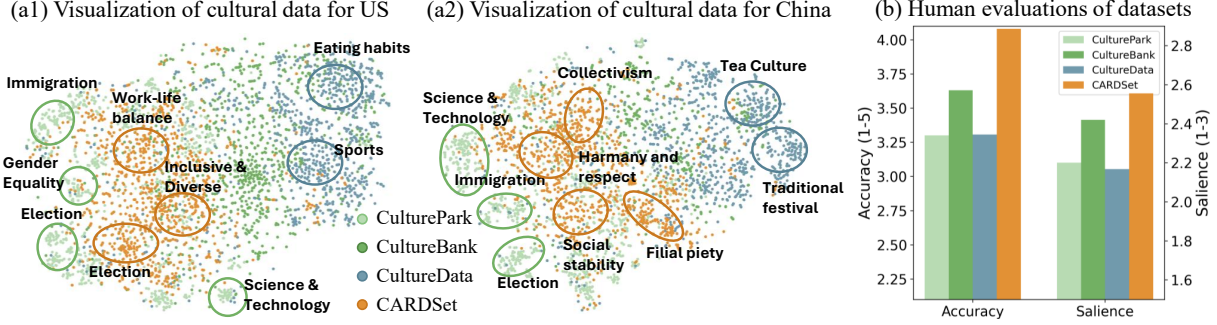


Figure 3. TSNE visualization and human evaluation of cultural datasets.

diction matches the top-1 human choice. (4) **World Value Survey (WVS)** (Haerpfer et al., 2020), a survey covering 13 value topics and various countries. Following (Xu et al., 2025), we measure the *consistency* between an LLM’s predictions and the culture’s real answers. More details and human evaluation protocols are in Appendix. D.1.

Baselines We evaluate our method on LLMs from different families and scales, including proprietary GPT-4.1 and GPT-5, open-sourced LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Team, 2024) and Gemma-3-27B-IT (Team et al., 2025). For all backbones, we include a **Role-Play** baseline that simulates culture-specific individuals. We further compare cultural LLMs finetuned with different datasets. As in Tab. 1, six cultural datasets are included. **CultureLLM** (Li et al., 2024a) and **CulturePark** (Li et al., 2024b) are augmented from real WVS data using GPT-4-Turbo. **CultureSPA** (Xu et al., 2025) is synthesized WVS-style QA data. Compared to our framework grounded in information-theoretic objectives, these are simple prompt engineering synthesis. **CultureBank** (Shi et al., 2024) and **CultureInstruct** (Pham et al., 2025) are culture-relevant text filtered from Tiktok/Reddit and DOLMA corpus (Soldaini et al., 2024). **CultureData** is constructed by merging public manual cultural datasets, such as NORMBANK (Ziems et al., 2023). For fair comparisons, we employ 1,000 samples per culture for all datasets except CultureInstruct which is a mixed dataset lacking explicit cultural labels. Additional details about baselines and our implementations can be found in Appendix D.3, D.4.

CAREDiO denotes our framework and CARDSet is the dataset synthesized via CAREDiO. Thus, Sec. 4.2 evaluates the raw samples in CARDSet. In subsequent sections, metrics are reported on LLMs finetuned with CARDSet.

4.2. Cultural Dataset Analysis

Before cultural alignment, we compare the quality of CARDSet built by CAREDiO with existing cultural datasets.

Quantitative Analysis In Tab. 1, we compare the datasets across multiple metrics. Samples in CARDSet are sub-

Table 1. Statistics of cultural datasets. ‘Avg.L’ is the average length of all samples. ‘#Cult. Points’ means the number of distinct cultural aspects in each dataset, where an aspect refers to a meaningful cultural concept expression (e.g., family ties, parenting style). ‘Sim / SB (Self-BLEU)’ measures the intra-dataset similarity at semantic and lexical level, on which lower values indicate better cross-cultural distinctiveness. ‘Cult. Sim’ is the cosine similarity between subsets of different cultures.

Datasets	#sample	Avg.L ↑	#Cult. Points ↑	Sim ↓	SB ↓	Cult. Sim ↓
CultureLLM	1,000 per cult.	48.6	245.2	0.246	0.616	0.246
CulturePark	1,000 per cult.	68.6	494.6	0.235	0.406	0.223
CultureSPA	1,000 per cult.	46.7	517.6	0.261	0.410	0.264
CultureBank	18,396 total	87.4	442.0	0.229	0.167	0.187
CultureInstruct	46,878 total	<u>191.1</u>	-	-	-	-
CultureData	1,000 per cult.	16.8	<u>1521.0</u>	0.199	0.330	0.127
CARDSet	1,000 per cult.	200.4	2027.0	0.251	<u>0.324</u>	0.202

stantially longer (Avg.L = 200.4) and contain more culture-relevant content (with more cultural points #Cult. Points = 2027.0 extracted by GPT-4.1). This suggests that CARDSet capture richer and more informative cultural factors. Moreover, CARDSet exhibits lower intra- and inter-cultural similarity, indicating that CARDSet encodes more diverse and unique cultural knowledge compared to prior baselines.

Visualization Analysis Fig. 3 presents a t-SNE visualization of cultural text embeddings. CulturePark covers important but shared value topics, such as ‘Gender Equality’, limiting its ability to distinguish subtle cultural variations. CultureData captures culture-specific elements, like the ‘Tea Culture’ for China, but these are often superficial and loosely connected to core values. In contrast, CARDSet 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. For the whole distribution, CARDSet also locates at a joint region of other datasets, indicating broader coverage and core factors.

Furthermore, a fundamental question is how the two objectives in CAREDiO help reshape the internal knowledge of the LLM to synthesize better cultural data. To this end, we conduct a visualization analysis in Appendix E.1.

Human Evaluation To complement automatic evaluation,

Table 2. Cultural alignment performance across four benchmarks. ‘%Imp. C.’ denotes the ratio of improved cultures, confirming consistent gains across multiple cultures rather than being driven by a few cultures; ‘Average’ reports 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. † and ‡ denote statistically significant improvements ($p < 0.05$) over the Role-Play baseline and the strongest baseline (underlined), respectively.

Family	Method	CulturalBench-Easy		CulturalBench-Hard		Prism		GlobalOpinionQA		WVS		Average
		Acc	% Imp. C.	Acc	% Imp. C.	Rating (1-5)	% Imp. C.	Acc	% Imp. C.	Acc	% Imp. C.	
Proprietary LLMs	GPT-5	88.79	-	59.54	-	2.187	-	46.27	-	62.38	-	60.14
	GPT-5 + Role-Play	89.55	12/14*	59.99	11/14*	4.519	14/14	60.08*	14/14	70.44*	8/8	74.09*
Qwen2.5-7B-Instruct	Raw Model	72.01	-	38.90	-	2.103	-	53.28	-	59.68	-	53.18
	Role-Play	72.38	8/14	36.73	8/14	<u>3.364</u>	4/14	55.83	13/14	64.96	8/8	59.44
	CultureLLM	71.79	7/14	34.79	6/14	3.121	14/14	<u>57.11</u>	12/14	65.49	7/8	58.32
	CulturePark	71.99	7/14	34.41	6/14	3.107	14/14	56.47	11/14	57.83	2/8	56.57
	CultureSPA	70.92	8/14	36.25	6/14	3.108	14/14	52.83	6/14	62.00	7/8	56.83
	CultureBank	72.28	8/14	27.34	2/14	3.193	14/14	56.43	13/14	62.47	7/8	56.48
	CultureInstruction	72.77	8/14	27.75	2/14	3.346	14/14	57.78	14/14	63.25	7/8	57.69
	CultureData	<u>72.83</u>	7/14	<u>40.11</u>	10/14	3.354	14/14	57.44	14/14	64.69	8/8	<u>60.43</u>
	CAREDiO	73.48	11/14	40.20†	11/14*	3.871†‡	14/14	56.23†	12/14	<u>65.26</u>	8/8	62.51†‡
	Raw Model	<u>82.11</u>	-	46.59	-	2.174	-	51.83	-	64.77	-	57.76
Gemma-3-27B-IT	Role-Play	81.33	9/14	48.28	10/14	<u>4.571</u>	14/14	54.84	10/14	67.22	4/8	68.62
	CultureLLM	80.46	8/14	46.31	9/14	4.441	14/14	58.15	10/14	66.99	4/8	68.14
	CulturePark	81.85	9/14	46.34	8/14	4.474	14/14	59.74	12/14	65.95	4/8	<u>68.67</u>
	CultureSPA	81.40	9/14	48.00	10/14	4.431	14/14	56.59	11/14	67.76	6/8	68.48
	CultureBank	81.82	10/14	41.88	6/14	4.323	14/14	55.89	9/14	67.11	5/8	66.63
	CultureInstruction	76.18	3/14	18.19	0/14	3.525	14/14	<u>59.22</u>	13/14	61.42	2/8	57.10
	CultureData	81.83	10/14	44.28	8/14	4.032	14/14	58.33	12/14	68.02	6/8	66.62
	CAREDiO	82.56	10/14	48.88	10/14	4.627*	14/14	58.25†	13/14	<u>67.96</u>	6/8	70.04†
	Raw Model	89.82	-	59.45	-	2.131	-	52.69	-	60.91	-	61.10
GPT-4.1	Role-Play	89.29	8/14	63.47	11/14*	<u>4.270</u>	14/14	53.76	9/14	69.85	8/8	<u>72.35</u>
	CultureBank	90.80*	11/14	60.00	9/14	4.226	14/14	56.76	14/14	68.11	8/8	72.04
	CAREDiO	<u>90.32</u>	10/14	63.54*	11/14*	4.336†	14/14	<u>56.64†</u>	13/14	<u>69.66†</u>	8/8	73.37†‡

we recruit native annotators from each culture to assess the data quality on 1) Accuracy (1-5), reflecting the cultural consensus level of the data; and 2) Salience (1-3), cultural representativeness and importance of the data. As shown in Fig. 3 (b), CARDSet consistently outperforms other datasets on both criteria, validating that our method produces culturally accurate and salient data. Details on annotator recruitment and guidance are provided in Appendix D.5.

4.3. Cultural Alignment Performance

Setup. We evaluate cultural alignment across 15 cultures and four LLM backbones. Tab. 2 reports scores averaged over cultures and the ratio of improved cultures on four benchmarks, using Qwen2.5-7B-Instruct, Gemma-3-27B-IT and GPT-4.1 as the aligned backbone. A critical difference diagram, per-culture results and experiments on Llama-3.1-8B-Instruct are provided in Appendix E.2. Fig. 4 (a) presents human evaluation results (win rate aggregated across US, China, Japan and Poland), with Qwen2.5-7B-Instruct being aligned. Fig. 4 (b) compares the datasets built by different LLMs under our CAREDiO framework, with fixed Qwen2.5-7B-Instruct being the downstream aligned model. Observing these results, we have two key findings.

First, *CAREDiO consistently improves cultural alignment across LLM families and scales, including strong proprietary LLMs.* Compared to raw models (Qwen2.5-7B, Gemma-3-27B, GPT-4.1) and role-playing baselines, CAREDiO achieves substantial gains on all benchmarks. Notably,

these improvements are broadly distributed across cultures rather than driven by a small subset, as evidenced by high improvement ratios (e.g., 14/14 on Prism, 13/14 on GlobalOpinionQA, and 8/8 on WVS). Fig. 4 (b) shows that our framework is robust to model choice. While more capable synthesis models lead to further improvement when available, data generated by the aligned backbone itself still yields clear gains. This demonstrates that our improvement of cultural alignment does not solely derive from knowledge distillation but also the information-theoretic objectives designed for representativeness and distinctiveness.

Second, *CAREDiO outperforms existing cultural datasets on most benchmarks, especially CultureBench-Hard and Prism passing significant t-test ($p < 0.05$).* CulturalBench covers a wide range of manually curated cultural knowledge and Prism reflects real-world interactions. Superior performance on these data highlights the practical usability of our method. On GlobalOpinionQA and WVS, CAREDiO shows significant improvements over Role-Play, CultureBank and CultureInstruction while lags slightly behind CultureLLM, we guess it is because CultureLLM augments real WVS data and thus has an inherent advantage in similar evaluations (see Appendix D.6 for discussion about possible data leakage). We consider synthesizing more survey-style data to narrow this gap. Notably, for advanced GPT-5, CAREDiO achieves comparable or better results on Prism. In the future, our framework can use GPT-5 for data synthesis to further improve itself once it is available for fine-tuning.

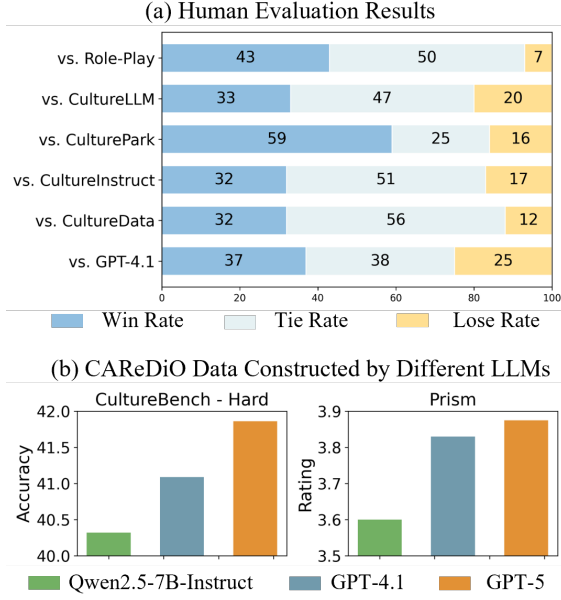


Figure 4. Results of cultural alignment.

Human Evaluation We conduct human evaluation with *native annotators* from the US, China, Japan and Poland (three per culture). Due to unaffordable cost of large-scale human annotation, we prioritize these four countries to ensure diversity in both culture sphere (e.g., Christianity, Catholicism and Confucianism) and resource levels (e.g., low-resource one: Poland). Showing Prism questions and responses generated by different methods, they rate the consensus level from 1 (conflict with the culture) to 5 (highly aligned with the culture). For each culture, we label 100 samples and report the average win rate in Fig. 4 (a). *Responses generated by CAREDiO are consistently preferred over baselines*. This demonstrates that our approach not only improves automatic benchmark scores but also produces outputs perceived as more culturally aligned by native users. More annotation details and human agreement are in Appendix D.5.

Table 3. Ablation study results on each independent objective.

Method	CB-Easy	CB-Hard	Prism	GlobalOpinion	Average	Improve.
Qwen2.5-7B	72.01	38.90	2.103	53.28	51.56	-
Role-Play	72.38	36.73	3.364	55.83	58.05	+12.59%
Only Rep	74.53	38.70	3.478	54.94	59.43	+15.26%
Only Dist	72.18	38.14	3.436	54.95	58.50	+13.45%
CAREDiO	73.48	40.20	3.871	56.23	61.83	+19.91%

4.4. Ablation Study

We perform an ablation study about both objectives, comparing two variants against the backbone LLM and the full CAREDiO. (1) **Only Rep** uses Eq.(3) for scoring and sample selection, optimizing representativeness only; (2) **Only Dist** uses Eq.(4) for scoring and sample selection, optimizing

distinctiveness only. Results are shown in Table 3.

(i) *Both objectives independently improve cultural alignment. Only Rep yields an average gain of 15.3%, while Only Dist achieves a 13.5%. (ii) Jointly optimizing both (full CAREDiO) achieves the best performance. (iii) Optimizing representativeness alone outperforms distinctiveness alone, indicating that current LLMs contain biased cultural knowledge while Eq.(3) more effectively mitigates them. (v) Adding distinctiveness brings additional gains, suggesting that Eq.(4) helps clarify cultural boundaries and reduce ambiguity. (More analytical discussion in Appendix E.3)*

Table 4. Confusion matrices of cross-cultural evaluation on GlobalOpinionQA, using CulturePark and CAREDiO for alignment.

(a) CulturePark on GlobalOpinionQA

	China	Japan	Korea	Δ Acc $\uparrow$	Conflict Acc $\downarrow$
China	52.41	53.35	51.59	-0.06	38.34
Japan	42.74	53.38	52.18	5.92	30.86
Korea	49.47	50.47	51.41	1.44	35.13

(b) CAREDiO on GlobalOpinionQA

	China	Japan	Korea	Δ Acc $\uparrow$	Conflict Acc $\downarrow$
China	53.88	50.84	49.47	3.73	34.69
Japan	45.25	54.56	53.70	5.08	33.73
Korea	45.23	45.54	52.46	7.07	27.29

4.5. Analysis Experiments

Across these experiments, we use Qwen2.5-7B-Instruct both for data synthesis and as the model being aligned, reporting the results averaged over 15 cultures.

Distinctiveness among Related Cultures To assess whether the distinctiveness objective sharpens cultural boundaries, we analyze three culturally proximate regions: China, Japan and Korea. We first compare CAREDiO-synthesized data with those from CulturePark. As visualized in Fig. 12, *CAREDiO produces substantially more separable clusters, while CulturePark exhibits cross-cultural overlap* (inter-cluster centroid distance 0.47 vs. 0.29). We further evaluate cultural LLMs aligned with them via cross-cultural confusion matrices. Tab. 4 (a) and (b) show that: (i) *CAREDiO-aligned models perform better on the target culture while worse on non-target cultures*, achieving improved cultural specificity (Δ Acc: 5.3 vs. 2.4); (ii) On conflict questions expecting different answers from distinct cultures, *CAREDiO also exhibits lower non-target alignment, suggesting reduced cultural confusion*. Together, these results confirm that CAREDiO enables finer-grained distinction among closely related cultures. We present comparison with more baselines and full results in Appendix E.4.

Number of Training Samples We continuously increase the training samples from 100 to 2,000, prioritizing high-

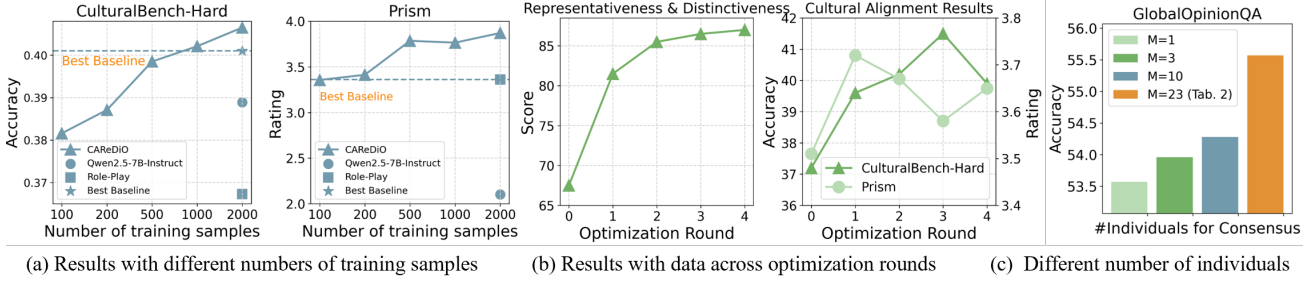


Figure 5. Analysis of three hyperparameters: number of training samples, optimization round and #individuals in consensus elicitation.

scoring samples under Eq. (5). As shown in Fig. 5 (a), performance improves steadily, with **early samples contributing larger gains**, supporting the efficiency of our approach. On Prism, CAREDiO 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 compare the cultural datasets generated across different optimized rounds (hyperparameter T in Algo 1) and analyze their impacts on downstream tasks. As shown in Fig. 5, the scores of optimization objectives improve along the iterative process, especially in the first round. This suggests that CAREDiO can produce high-quality cultural data with limited synthesis cost.

Interestingly, we observe divergent performance trends across downstream benchmarks. On the knowledge-intensive CultureBench-Hard, more optimization rounds yield larger improvements. Conversely, performance on the open-ended Prism dataset saturates quickly and exhibits a marginal decline after Round 1. We attribute this to the shifting granularity of the synthesized data. As iterations proceed, the cultural data becomes increasingly high-quality and specific. This can provides clearer knowledge for CB-Easy/Hard. In contrast, open-ended tasks like Prism demand highly broad question generalizability, which can be marginally constrained.

To prevent over-specialization, CAREDiO explicitly enforces question generalizability through: (1) constraining initial questions from being too specific (Fig. 6), (2) preserving generality during prompt refinement (Fig. 10), and (3) implementing an early-stopping mechanism.

Number of Individuals in Consensus Elicitation We analyze the number of individuals in consensus elicitation (hyperparameter M in Eq.(3)). As shown in Fig. 5 c), *increasing M consistently improves performance*, validating the role of consensus for representativeness and mitigating bias. Importantly, even a small M significantly outperforms the single-individual setting, making CAREDiO robust under limited resources. Full results are in Appendix E.6.

Hyperparameter Sensitivity in Eq.(5) We conduct a controlled ablation study about weights $\lambda_1, \lambda_2, \lambda_3$ for representativeness, distinctiveness and diversity, respectively. While fixing two as 1, we vary another independently from 0 to 3. As shown in Tab. 5, we find: (1) Combining representativeness and distinctiveness equally performs the best, confirming our default design choice. (2) The diversity term plays an important role: setting it to 0 leads to redundant high-score samples and degraded performance, while overly large values remove informative samples. Based on this, we recommend $\lambda_1, \lambda_2, \lambda_3 = 1$ as a robust default configuration.

Table 5. Sensitivity analysis on the hyperparameter weights λ_1, λ_2 , and λ_3 in Eq.(5), which are all set as 1 by default.

λ_1 (Rep.)	λ_2 (Dist.)	λ_3 (Div.)	Avg. (same as Tab. 3)
1	1	1	61.83
0	1	1	58.50
2	1	1	60.77
3	1	1	60.31
1	0	1	59.43
1	2	1	60.58
1	3	1	59.84
1	1	0	59.37
1	1	2	61.46
1	1	3	60.68

More analyses on result variance, classifier robustness, token cost quantification and case study are in Appendix E.5, E.7, E.8, E.9.

5. Conclusion

In this paper, we addresses the challenges of representativeness and distinctiveness in existing cultural datasets by introducing CAREDiO, a novel data optimization framework. It applies an iterative in-context schema to generate and refine questions and responses for two information-theoretic objectives, enhancing representativeness and distinctiveness. Using the constructed dataset CARDSet covering 15 cultures, we demonstrate CAREDiO outperforms baseline datasets in both multiple-choice and open-ended benchmarks.

Furthermore, we provide an in-depth discussion of the limitations and future directions in Appendix G.

Impact Statement

This paper introduces CAREDiO, a novel framework to enhance the cultural alignment of Large Language Models (LLMs). As LLMs are increasingly deployed in real-world applications, their ability to respect and appropriately respond to diverse cultural contexts has direct implications for fairness, inclusivity, and user trust. 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 itself is general and can be extended to a broader range of cultural contexts, potentially contributing to more equitable access to culturally aware AI systems. Furthermore, our method is also targeted at the efficiency of cultural alignment, which makes it more feasible to support underrepresented cultures with limited resources.

In addition, we emphasize the importance of responsible development and reproducibility. Due to space limitations, we provide technical details such as derivations, implementations, and some experimental settings in the Appendix. We submit the core code of our method as the supplementary materials for clarity and commit to releasing the necessary code and data upon acceptance to support reproducibility.

References

- Abbasi, M. A., Ghafouri, A., Firouzmandi, M., Naderi, H., and Bidgoli, B. M. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*, 2023.
- Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S. S., Aji, A. F., O’Neill, J., Modi, A., and Choudhury, M. Towards measuring and modeling “culture” in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15763–15784, 2024.
- AlKhamissi, B., ElNokrashy, M., Alkhamissi, M., and Diab, M. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12404–12422, 2024.
- Askill, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.
- Bai, Y., Jones, A., Ndousse, K., Askill, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Banerjee, S., Layek, S., Shrawgi, H., Mandal, R., Halder, A., Kumar, S., Basu, S., Agrawal, P., Hazra, R., and Mukherjee, A. Navigating the cultural kaleidoscope: A hitchhiker’s guide to sensitivity in large language models. 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. 7580–7617, 2025.
- Barber, D. and Agakov, F. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.
- Cao, Y., Zhou, L., Lee, S., Piqueras, L. C., Chen, M., and Hershcovich, D. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pp. 53–67, 2023.
- Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., et al. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms. 2024.
- Choenni, R. and Shutova, E. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *arXiv preprint arXiv:2408.16482*, 2024.
- Choenni, R., Lauscher, A., and Shutova, E. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15042–15058, 2024.
- Cosier, R. A. and Rose, G. L. Cognitive conflict and goal conflict effects on task performance. *Organizational behavior and human performance*, 19(2):378–391, 1977.
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askill, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

- Dwivedi, A., Lavania, P., and Modi, A. Eticor: Corpus for analyzing llms for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6921–6931, 2023.
- Engleson, E. and Azizpour, H. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.
- Feng, R., Gao, S., Chen, X., Chen, L., and Shang, S. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22413–22430, 2025.
- Fiske, S. T. T. and Taylor, S. E. Social cognition: From brains to culture. 2020.
- Fung, Y., Chakrabarty, T., Guo, H., Rambow, O., Muresan, S., and Ji, H. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15217–15230, 2023.
- Fung, Y., Zhao, R., Doo, J., Sun, C., and Ji, H. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*, 2024.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Gupta, K., Thérien, B., Ibrahim, A., Richter, M. L., Anthony, Q., Belilovsky, E., Rish, I., and Lesort, T. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*, 2023.
- Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. World values survey wave 7 (2017-2020) cross-national data-set. (*No Title*), 2020.
- Hao, B., Lattimore, T., and Qin, C. Contextual information-directed sampling. In *International Conference on Machine Learning*, pp. 8446–8464. PMLR, 2022.
- Hofstede, G. and Hofstede, G. Cultures and organizations: Software of the mind. third millennium edition, 2005.
- Karinshak, E., Hu, A., Kong, K., Rao, V., Wang, J., Wang, J., and Zeng, Y. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.
- Kharchenko, J., Roosta, T., Chadha, A., and Shah, C. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*, 2024.
- Kim, E., Suk, J., Oh, P., Yoo, H., Thorne, J., and Oh, A. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3335–3346, 2024.
- Kirk, H. R., Whitefield, A., Rottger, P., Bean, A. M., Margatina, K., Mosquera-Gomez, R., Ciro, J., Bartolo, M., Williams, A., He, H., et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multi-cultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2025.
- Kwok, L., Bravansky, M., and Griffin, L. D. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*, 2024.
- Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838, 2024a.
- Li, C., Teney, D., Yang, L., Wen, Q., Xie, X., and Wang, J. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216, 2024b.
- Limón, M. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and instruction*, 11(4-5):357–380, 2001.
- Liu, C. C., Korhonen, A., and Gurevych, I. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3114–3134, 2025.
- Mallen, A., Brumley, M., Kharchenko, J., and Belrose, N. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023.
- Masoud, R., Liu, Z., Ferianc, M., Treleaven, P. C., and Rodrigues, M. R. Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8474–8503, 2025.

- Miyamoto, Y., Yoo, J., Levine, C. S., Park, J., Boylan, J. M., Sims, T., Markus, H. R., Kitayama, S., Kawakami, N., Karasawa, M., 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.
- Mostowlansky, T. and Rota, A. Emic and etic. 2020.
- Mukherjee, A., Caliskan, A., Zhu, Z., and Anastasopoulos, A. 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.
- Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., 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.
- Nguyen, T.-P., Razniewski, S., Varde, A., and Weikum, G. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pp. 1907–1917, 2023.
- Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., et al. Seallms-large language models for southeast asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 294–304, 2024.
- OpenAI. Gpt-4 technical report, 2024.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Pham, V. T., Li, Z., Qu, L., and Haffari, G. Cultureinstruct: Curating multi-cultural 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.
- Pipatanakul, K., Jirabovonvisut, P., Manakul, P., Sripaisarnmongkol, S., Patomwong, R., Chokchainant, P., and Tharnpipitchai, K. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*, 2023.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pp. 226–240. Springer, 2023.
- Rao, A., Yerukola, A., Shah, V., Reinecke, K., and Sap, M. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*, 2024.
- Ryan, M. J., Held, W., and Yang, D. Unintended impacts of llm alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16121–16140, 2024.
- Rystrøm, J. H., Kirk, H. R., and Hale, S. A. Multilingual!=multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pp. 74–85, 2025.
- Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.
- Shi, W., Li, R., Zhang, Y., Ziem, C., Yu, S., Horesh, R., De Paula, R. A., and Yang, D. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, 2024.
- Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.
- Sukiennik, N., Gao, C., Xu, F., and Li, Y. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*, 2025.
- Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.
- Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Team, Q. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2, 2024.

- Triandis, H. C., Bontempo, R., Leung, K., and Hui, C. H. A method for determining cultural, demographic, and personal constructs. *Journal of Cross-Cultural Psychology*, 21(3):302–318, 1990.
- Wang, B., Lin, G., Liu, Z., Wei, C., and Chen, N. Craft: Extracting and tuning cultural instructions from the wild. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 42–47, 2024a.
- Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., and Lyu, M. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6349–6384, 2024b.
- Wang, X., Duan, S., Yi, X., Yao, J., Zhou, S., Wei, Z., Zhang, P., Xu, D., Sun, M., and Xie, X. On the essence and prospect: An investigation of alignment approaches for big models. *arXiv preprint arXiv:2403.04204*, 2024c.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.
- Wang, Y., Zhu, Y., Kong, C., Wei, S., Yi, X., Xie, X., and Sang, J. Cdeval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 1–16, 2024d.
- Weller, S. C. Cultural consensus theory: Applications and frequently asked questions. *Field methods*, 19(4):339–368, 2007.
- Xu, S., Leng, Y., Yu, L., and Xiong, D. Self-pluralising culture alignment for large language models. 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. 6859–6877, 2025.
- Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., and Zhou, T. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.
- Yuan, J., Di, Z., Zhao, S., Cui, Z., Wang, H., Yang, G., and Naseem, U. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*, 2024.
- Ziems, C., Dwivedi-Yu, J., Wang, Y.-C., Halevy, A., and Yang, D. Normbank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7756–7776, 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 $\mathbf{x}$; $\mathbf{c}_1, \dots, \mathbf{c}_K, \mathbf{c}_{K+1}$ as $K+1$ different cultures. Our goal is to find the best question-response pair $(\mathbf{x}^*, \mathbf{y}^*)$ which satisfies: i) $(\mathbf{x}^*, \mathbf{y}^*) = \operatorname{argmax}_{(\mathbf{x}, \mathbf{y})} p_{\mathbf{c}_i}(\mathbf{x}, \mathbf{y})$ where $\mathbf{c}_i$ is the target culture; and ii) $(\mathbf{x}^*, \mathbf{y}^*) = \operatorname{argmin}_{(\mathbf{x}, \mathbf{y})} \frac{1}{K} \sum_{k, k \neq i} p_{\mathbf{c}_k}(\mathbf{x}, \mathbf{y})$. Requirement i) follows our *Representativeness* rule and Requirement ii) is in line with *Distinctiveness*. We should show how it can be approximated and solved.

Representativeness Optimization The major challenge of $(\mathbf{x}^*, \mathbf{y}^*) = \operatorname{argmax}_{(\mathbf{x}, \mathbf{y})} p_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$ 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 that each large enough LLM $p_\theta(\mathbf{y}|\mathbf{x})$ possesses sufficient competence but it is usually unelicited. Therefore, we approximate the Representativeness objective as a *consensus elicitation* problem, and find the best $\mathbf{y}$ ⁴ that maximizes $I_\theta(\mathbf{c}; \mathbf{y}|\mathbf{x} = \mathbf{x})$:

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

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

$$\Delta_{\mathbf{c}}^\theta(\mathbf{y}) = \hat{I}_\theta(\mathbf{c}; \mathbf{y}|\mathbf{x}) = \log p_\theta(\mathbf{c}|\mathbf{x}, \mathbf{y}) - \log p_\theta(\mathbf{c}|\mathbf{x}) \quad (7)$$

Eq.(7) represents a form of Information-Directed Sampling (IDS) (Hao et al., 2022), which helps find the response $\mathbf{y}$ that reinforces the LLM p_θ ’ understanding of culture $\mathbf{c}$. When p_θ 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_θ is a larger LLM (potentially with better culture competence), e.g., GPT-5, we conduct typical knowledge distillation. In practice, we use multiple LLMs to perform a better consensus elicitation:

$$\Delta_{\mathbf{c}}^{\theta_1, \dots, \theta_M}(\mathbf{y}) = \sum_i^M \Delta_{\mathbf{c}}^{\theta_i}(\mathbf{y}), \quad (8)$$

where each θ_i could be either a heterogeneous LLM or the same one with different individual settings. With this approximated representativeness objective, we then give the following conclusion:

Theorem 1. For two samples $\mathbf{y}_1, \mathbf{y}_2$ toward the same question $\mathbf{x}$, assume their probabilities under the true cultural distribution do not differ significantly, i.e., $|p_{\mathbf{c}}(\mathbf{y}_1|\mathbf{x}) - p_{\mathbf{c}}(\mathbf{y}_2|\mathbf{x})| < \epsilon$, which holds for culturally plausible candidate answers, if their scores from Eq.(7) satisfy $\Delta_{\mathbf{c}}^\theta(\mathbf{y}_1) > \Delta_{\mathbf{c}}^\theta(\mathbf{y}_2)$, then using $\mathbf{y}_1$ for fine-tuning leads to a larger gradient than using $\mathbf{y}_2$: $\|\nabla_{\theta} l_\theta(\mathbf{y}_1, \mathbf{x})\| > \|\nabla_{\theta} l_\theta(\mathbf{y}_2, \mathbf{x})\|$.

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

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{\mathbf{c}}} [\log p_\theta(\mathbf{y}|\mathbf{x})]. \quad (9)$$

Then, we have the gradient for the parameters θ :

$$\nabla_{\theta} \mathcal{L}(\theta) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{\mathbf{c}}} [\nabla_{\theta} l_\theta(\mathbf{y}, \mathbf{x})], \quad l_\theta(\mathbf{y}, \mathbf{x}) = \log p_\theta(\mathbf{y}|\mathbf{x}). \quad (10)$$

We then consider the gradient magnitude $\|\nabla_{\theta} l_\theta(\mathbf{y}, \mathbf{x})\|$ and demonstrate $\|\nabla_{\theta} l_\theta(\mathbf{y}, \mathbf{x})\| \propto [\log p_{\mathbf{c}}(\mathbf{y}|\mathbf{x}) - \log p_\theta(\mathbf{y}|\mathbf{x})]$. Suppose the model conducts softmax to compute the probability of output in the last layer, with the parameters $\mathbf{w}$. For

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

brevity, we consider $\mathbf{y}$ as one single token (which is typical in multiple-choice questions), and thus we have:

$$\begin{aligned} z_{\mathbf{y}}(\mathbf{x}) &= \mathbf{w}_{\mathbf{y}}^T \cdot h(\mathbf{x}), \\ p_{\theta}(\mathbf{y}|\mathbf{x}) &= \frac{\exp(z_{\mathbf{y}}(\mathbf{x}))}{\sum_i \exp(z_i(\mathbf{x}))}, \\ \log p_{\theta}(\mathbf{y}|\mathbf{x}) &= z_{\mathbf{y}}(\mathbf{x}) - \log \sum_i \exp(z_i(\mathbf{x})) \end{aligned} \quad (11)$$

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

$$\frac{\partial \log p_{\theta}(\mathbf{y}|\mathbf{x})}{\partial w_i} = (\mathbb{I}[i = \mathbf{y}] - p_{\theta}(i|\mathbf{x})) \cdot h(\mathbf{x}). \quad (12)$$

Thus, we have $\|\nabla_{\theta} \log p_{\theta}(\mathbf{y}, \mathbf{x})\| = \|\mathbb{I}[i = \mathbf{y}] - p_{\theta}(i|\mathbf{x})\| \cdot \|h(\mathbf{x})\|$. Since the number of candidate tokens is usually huge for an LLM, the probability $p_{\theta}(i|\mathbf{x})$ for tokens $i \neq \mathbf{y}$ is usually very small. For simplicity, we consider the case $i = \mathbf{y}$ that mainly affects the gradient magnitude. Defining $\delta(\mathbf{x}, \mathbf{y}) = \log p_c(\mathbf{y}|\mathbf{x}) - \log p_{\theta}(\mathbf{y}|\mathbf{x})$, we present $\log p_{\theta}(\mathbf{y}|\mathbf{x}) = \log p_c(\mathbf{y}|\mathbf{x}) - \delta(\mathbf{x}, \mathbf{y})$. Substituting $p_{\theta}(\mathbf{y}|\mathbf{x})$ with this term and ignore $\|h(\mathbf{x})\|$ that has no correlation with $\mathbf{y}$, we obtain

$$\|\nabla_{\theta} \log p_{\theta}(\mathbf{y}, \mathbf{x})\| = \|\mathbb{I}[i = \mathbf{y}] - p_{\theta}(i|\mathbf{x})\| \cdot \|h(\mathbf{x})\| \propto 1 - e^{\log p_c(\mathbf{y}|\mathbf{x}) - \delta(\mathbf{x}, \mathbf{y})} \quad (13)$$

For plausible training samples $(\mathbf{x}, \mathbf{y})$ from the target culture $\mathbf{c}$, they have a large score in $\log p_c(\mathbf{y}|\mathbf{x})$ and only yield a narrow probability difference range. Thus, $\delta(\mathbf{x}, \mathbf{y})$ plays a major role in determining the value of Eq.(13). This indicates that using a $(\mathbf{y}, \mathbf{x})$ with a larger score $\delta(\mathbf{x}, \mathbf{y})$ to fine-tune the LLM p_{θ} can leads to a larger gradient magnitude, accelerating the cultural learning towards $p_c(\mathbf{x}, \mathbf{y})$.

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

$$\begin{aligned} \delta(\mathbf{x}, \mathbf{y}) &= \log p_c(\mathbf{y}|\mathbf{x}) - \log p_{\theta}(\mathbf{y}|\mathbf{x}) \\ &\approx \log p_{\theta}(\mathbf{y}|\mathbf{x}, \mathbf{c}) - \log p_{\theta}(\mathbf{y}|\mathbf{x}) \quad (\text{using Bayes Equation}) \\ &= \log \frac{p_{\theta}(\mathbf{y}|\mathbf{x}, \mathbf{c}) \cdot p_{\theta}(\mathbf{x}, \mathbf{c})}{p_{\theta}(\mathbf{y}|\mathbf{x}) \cdot p_{\theta}(\mathbf{x}, \mathbf{c})} \\ &= \log \frac{p_{\theta}(\mathbf{c}|\mathbf{y}, \mathbf{x}) \cdot p_{\theta}(\mathbf{y}|\mathbf{x}) \cdot p_{\theta}(\mathbf{x})}{p_{\theta}(\mathbf{y}|\mathbf{x}) \cdot p_{\theta}(\mathbf{c}|\mathbf{x}) \cdot p_{\theta}(\mathbf{x})} \\ &= \log p_{\theta}(\mathbf{c}|\mathbf{y}, \mathbf{x}) - \log p_{\theta}(\mathbf{c}|\mathbf{x}) \\ &= \Delta_{\mathbf{c}}^{\theta}(\mathbf{y}|\mathbf{x}), \end{aligned} \quad (14)$$

which indicates that using samples $\mathbf{y}$ with a larger score from Eq.(7) to finetune p_{θ} can approximately leads to a larger gradient magnitude, accelerating the cultural learning towards $p_c(\mathbf{x}, \mathbf{y})$.

Distinctiveness Optimization To optimize $(\mathbf{x}^*, \mathbf{y}^*) = \operatorname{argmin}_{(\mathbf{x}, \mathbf{y})} \frac{1}{K} \sum_{k, k \neq i} p_{\mathbf{c}_k}(\mathbf{x}, \mathbf{y})$, we refer to the *Cognitive Conflict Theory*, and elicit cultural differences from conflicts. For a given question $\mathbf{x}$, assume we have collected a set of $\mathbf{y}$ by Eq.(7), which forms an empirical distribution $q(\mathbf{y}|\mathbf{x})$ 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, $\mathbf{c}_1, \dots, \mathbf{c}_K$.

Concretely, we use the following objective:

$$\phi(\mathbf{y}, \mathbf{x}) \left[\log \frac{\phi(\mathbf{y}, \mathbf{x})}{1 - \phi(\mathbf{y}, \mathbf{x})} + \log \frac{1 - \alpha}{2\alpha} \right] + \log(1 - \phi(\mathbf{y}, \mathbf{x})) = \Gamma(\mathbf{y}|\mathbf{x}), \quad (15)$$

$$q^*(\mathbf{y}|\mathbf{x}) = \operatorname{argtop}_{\mathbf{y}} \Gamma(\mathbf{y}|\mathbf{x}) \quad (16)$$

where $\phi(\mathbf{y}, \mathbf{x})$ is a classifier to give the probability that the response $\mathbf{y}$ 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 $\mathbf{x}$ and $\mathbf{y}$, if the classifier error $|\phi(\mathbf{y}, \mathbf{x}) - p^*(\mathbf{y} \notin p_{c_1}, \dots, p_{c_K} | \mathbf{y}, \mathbf{x})| < \epsilon$, and the classifier is not over-confident, i.e., $\phi(\mathbf{y}, \mathbf{x}) < \eta$, then maximizing Eq.(15) is an approximated point-wise maximization of the lower bound of $GJS_{\alpha, \mathbf{w}}[q(\mathbf{y}|\mathbf{x}), p_{c_1}(\mathbf{y}|\mathbf{x}), \dots, p_{c_K}(\mathbf{y}|\mathbf{x})]$, 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 $\mathbf{x}$, assume we have collected a set of $\mathbf{y}$ by Eq.(7), which forms an empirical distribution $q(\mathbf{y}|\mathbf{x})$ 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, $c_1, \dots, c_K$. We optimize:

$$q^*(\mathbf{y}|\mathbf{x}) = \underset{q}{\operatorname{argmax}} GJS_{\alpha, \mathbf{w}}[q(\mathbf{y}|\mathbf{x}), p_{c_1}(\mathbf{y}|\mathbf{x}), \dots, p_{c_K}(\mathbf{y}|\mathbf{x})], \quad (17)$$

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

For brevity, we omit $\mathbf{x}$. Since $1 - \alpha = \sum_{i=1}^K w_i$, define $\beta_i = \frac{w_i}{1-\alpha}$, and the average non-target culture distribution as $\hat{p} = \frac{\sum_{i=1}^K w_i p_{c_i}}{1-\alpha}$, we have $m = \alpha p + \sum_{i=1}^K w_i p_{c_i}$, and thus $GJS_{\alpha, \mathbf{w}}[q, p_{c_1}, \dots, p_{c_K}] = \alpha \text{KL}[q||m] + (1 - \alpha) \sum_{i=1}^K \beta_i \text{KL}[p_{c_i}||m] = \alpha q + (1 - \alpha)\hat{p}$. Then, we further have:

$$\begin{aligned} & GJS_{\alpha, \mathbf{w}}[q, p_{c_1}, \dots, p_{c_K}] \\ &= \alpha \mathbb{E}_q[\log q - \log m] + (1 - \alpha) \sum_{i=1}^K \beta_i \mathbb{E}_{p_{c_i}}[\log p_{c_i} - \log m] \\ &= \alpha \mathbb{E}_q[\log q - \log m] + (1 - \alpha) \left[\mathbb{E}_{\hat{p}} \log \hat{p} - \sum_{i=1}^K \beta_i \mathbb{E}_{p_{c_i}} \log m + \sum_{i=1}^K \beta_i \mathbb{E}_{p_{c_i}} \log p_{c_i} - \mathbb{E}_{\hat{p}} \log \hat{p} \right] \\ &= \alpha \mathbb{E}_q[\log q - \log m] + (1 - \alpha) \left[\mathbb{E}_{\hat{p}} \log \hat{p} - \sum_{i=1}^K \beta_i \mathbb{E}_{p_{c_i}} \log m + GJS_{\beta}[p_{c_1}, \dots, p_{c_K}] \right] \\ &= \alpha \text{KL}[q||m] + (1 - \alpha) \text{KL}[\hat{p}||m] + (1 - \alpha) GJS_{\beta}[p_{c_1}, \dots, p_{c_K}] \\ &= GJS_{\alpha}[q, m] + (1 - \alpha) GJS_{\beta}[p_{c_1}, \dots, p_{c_K}] \\ &\geq GJS_{\alpha}[q, m]. \end{aligned} \quad (18)$$

Once the mix weight $\mathbf{w}$ is determined, $GJS_{\beta}[p_{c_1}, \dots, p_{c_K}]$ only relies on $p_{c_1}, \dots, p_{c_K}$, irrelevant to q . Therefore, we only maximize $GJS_{\alpha}[q, m]$. However, each true p_{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 $\mathbf{y}$ for a fixed question $\mathbf{x}$. When $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x})$, $\mathbf{s} = 0$, when $\mathbf{y} \sim \hat{p}(\mathbf{y}|\mathbf{x})$, $\mathbf{s} = 1$. We then maximize the mutual information $I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \mathbf{x})$. We then also have:

$$\begin{aligned} & I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \mathbf{x}) \\ &= p(\mathbf{s} = 0|\mathbf{x}) \text{KL}[p(\mathbf{y}|\mathbf{s} = 0, \mathbf{x})||p(\mathbf{y}|\mathbf{x})] + p(\mathbf{s} = 1|\mathbf{x}) \text{KL}[p(\mathbf{y}|\mathbf{s} = 1, \mathbf{x})||p(\mathbf{y}|\mathbf{x})] \\ &= \alpha \text{KL}[q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})] + (1 - \alpha) \text{KL}[\hat{p}(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})] \\ &= \alpha \text{KL}[q(\mathbf{y}|\mathbf{x})||m(\mathbf{y}|\mathbf{x})] + (1 - \alpha) \text{KL}[\hat{p}(\mathbf{y}|\mathbf{x})||m(\mathbf{y}|\mathbf{x})] \\ &= GJS_{\alpha}[q, m]. \end{aligned} \quad (19)$$

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

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

$$I(\mathbf{s}; \mathbf{y}|\mathbf{x} = \mathbf{x}) \geq \mathbb{E}_{p(\mathbf{y}|\mathbf{x})} \mathbb{E}_{p(\mathbf{s}|\mathbf{y}, \mathbf{x})} [\log q_{\phi}(\mathbf{s}|\mathbf{y}, \mathbf{x}) - \log p(\mathbf{s}|\mathbf{x})]. \quad (20)$$

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

$$\begin{aligned}
 & \hat{I}(\mathbf{s}, \mathbf{y} | \mathbf{x}) \\
 & \geq p(\mathbf{s} = 0 | \mathbf{y}, \mathbf{x}) \left[\log \frac{q_\phi(\mathbf{s} = 0 | \mathbf{y}, \mathbf{x})}{q_\phi(\mathbf{s} = 1 | \mathbf{y}, \mathbf{x})} + \log \frac{1 - \alpha}{\alpha} \right] + \log \frac{q_\phi(\mathbf{s} = 1 | \mathbf{y}, \mathbf{x})}{2} \\
 & \approx \phi(\mathbf{y}, \mathbf{x}) \left[\log \frac{\phi(\mathbf{y}, \mathbf{x})}{1 - \phi(\mathbf{y}, \mathbf{x})} + \log \frac{1 - \alpha}{2\alpha} \right] + \log(1 - \phi(\mathbf{y}, \mathbf{x})) \\
 & = \Gamma(\mathbf{y}),
 \end{aligned} \tag{21}$$

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

From the derivation above, we conclude that optimizing $\Gamma(\mathbf{y})$ is equivalent to optimizing a point-wise lower bound of $\text{GJS}_{\alpha, \mathbf{w}}[q, p_{c_1}, \dots, p_{c_K}]$. Assume the error of this classifier $|\phi(\mathbf{y}, \mathbf{x}) - p(\mathbf{s} = 0 | \mathbf{y}, \mathbf{x})| < \epsilon$ and the classifier is not over-confident, *i.e.*, $\phi(\mathbf{y}, \mathbf{x}) < \eta$, we can easily have the approximation error $< \epsilon |\log \frac{\eta(1-\alpha)}{2(1-\eta)\alpha}|$.

We use two iterative steps to optimize $\mathcal{S}(\mathbf{x}, \mathbf{y})$.

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 $\mathcal{S}(\mathbf{x} | \mathbf{y})$. This step mainly involves: i) enhancing $p_c(\mathbf{x})$, 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 $p_c(\mathbf{y} | \mathbf{x})$, ii) the representativeness of $(\mathbf{x}, \mathbf{y})$; iii) the distinctiveness.

B. Supplements for Methodology

B.1. Plausibility and Effectiveness of Consensus Elicitation

To obtain cultural data with higher representativeness, we draw on Cultural Consensus Theory (CulCT) (Weller, 2007) and approximate this objective as a consensus elicitation problem. This approximation relies on two assumptions: (1) each LLM p_ω possesses sufficient cultural competence to identify responses that reflect the target culture; and (2) one single LLM may encode incomplete and biased knowledge of the target culture, whereas multiple LLMs prompted with different personas $p_{\omega_1}, \dots, p_{\omega_M}$ can help extract cultural knowledge from diverse perspectives and vote for a more representative consensus, mitigating bias. Below, we discuss the rationale behind these assumptions and empirically validate both.

(1) **Prompting an LLM to simulate cultural roles can indeed assign it with higher cultural competence.** Known as In-Context Alignment (ICA), conditioning an LLM with attributes, such as role, persona and culture, has been a widely adopted method in cultural alignment and shown efficacy (Durmus et al., 2023; Kwok et al., 2024; Kharchenko et al., 2024; Choenni & Shutova, 2024). Moreover, our experiments in Table 2 also demonstrate that the Role-Play baseline significantly improves the original base model, *e.g.*, 57.75 $\rightarrow$ 68.62 on Gemma-3-27B-IT, showing that ICA indeed enhances an LLM’s cultural competence. In Tab. 7, we compare a role-playing general model (GPT-5) with region-specific LLMs *skt/A.X-4.0-Light* for South Korea and *Llama-SEA-LION-v3.5-8B-R* for Singapore. The results show that GPT-5 with Role-Play demonstrates highly competitive performance and significantly outperforms local models on complex tasks.

(2) **The performance of cultural alignment can be further enhanced by extracting the consensus from multiple role-playing LLMs.** Using GPT-5 as the backbone LLM, for each culture, we prompt it to simulate seven roles with distinct personas: i) three general people with different demographics randomly sampled from the WVS data of the target culture; ii) three cultural experts such as sociologist; and iii) one cross-cultural researcher who comes from another culture but studies global cultural patterns. First, we evaluate these personas individually on GlobalOpinionQA and measure their mutual agreement in Table 4. Then, we extract their consensus on each test question and compare the performance between individual roles and the consensus results in Table 9. Two key observations emerge:

- **Different roles show divergent cultural perceptions.** We observe obvious differences in the outputs of distinct roles. This indicates that incorporating multiple personas can help activate diverse cultural knowledge.

Table 6. Notation Table

Variable	Description
$\mathbf{x}$	the question
$\mathbf{y}$	the response
$(\mathbf{x}, \mathbf{y})$	a cultural data sample
θ	parameters of a target LLM to be aligned
$p_{\theta}(\mathbf{y} \mathbf{x})$	an LLM to be aligned
$p_c(\mathbf{x}, \mathbf{y})$	the true data distribution of the target culture c
$p_{c_k}(\mathbf{x}, \mathbf{y})$	the true data distribution of the non-target culture c_k
K	the number of non-target cultures
q_c^*	the goal set of cultural data, with satisfactory representativeness and distinctiveness
p_{ω}	an LLM that simulates an individual with cultural competence of the target culture c
p_{ω_i}	the i -th LLM for simulation
$I(\mathbf{c}; \mathbf{y} \mathbf{x})$	the mutual information between $\mathbf{c}$ and $\mathbf{y}$ given $\mathbf{x}$ as the condition
$\Delta_c^{\omega}(\mathbf{y} x)$	the approximated representativeness score by an LLM p_{ω}
$\Delta_c^{\omega_1, \dots, \omega_M}(\mathbf{y} x)$	the approximated representativeness score by the group of simulating LLMs
M	the number of cultural individuals participating in consensus elicitation
$\Gamma_{c_1, \dots, c_K}(\mathbf{y} \mathbf{x})$	the approximated distinctiveness score
$\phi(x, y)$	the classifier that estimates the probability that y does NOT come from $c_1, \dots, c_K$
$\text{sim}(\cdot, \cdot)$	cosine similarity between text embeddings
ϵ	the classifier error
η	the classifier confidence
$\mathcal{E}$	the approximate error bound of Proposition 2
$S_c(\mathbf{y} \mathbf{x})$	the score of $\mathbf{y}$ for data optimization
$\lambda_1, \lambda_2, \lambda_3$	hyperparameters to trade-off representativeness and distinctiveness
$\{x_i^0\}_{i=1}^N$	the initial question set for cultural data optimization
t	the iteration index
N	the total number of samples for optimization
n	the number of candidate responses generated in each optimization step
m	the number of candidate questions generated in each optimization step

 Table 7. Performance comparison between a role-conditioned non-local advanced LLM (GPT-5) and specialized local LLMs under the role-play setting (*skt/A.X-4.0-Light* for South Korea and *Llama-SEA-LION-v3.5-8B-R* for Singapore).

Model	GlobalOpinionQA		CultureBench-Hard	
	South Korea	Singapore	South Korea	Singapore
local LLM + Role-Play	57.22	70.96	14.63	34.78
GPT-5 + Role-Play	60.04	67.74	53.66	47.83

- **Consensus of multiple roles further improves stability and accuracy.** Obviously, aggregating multiple roles yields better and more stable (lower std across cultures) results, with especially larger gains in low-resource settings, e.g., Russia and Poland, where model biases are more prominent.

Table 8. The average and minimum agreement among multiple roles.

	United States	Germany	Japan	Russia	Poland	Mexico	Nigeria
Average Agreement	0.86	0.85	0.81	0.80	0.81	0.79	0.80
Minimum Agreement	0.83	0.78	0.72	0.69	0.74	0.70	0.74

Table 9. Performance comparison between single LLM and multi-LLM consensus on the GlobalOpinionQA dataset, using GPT-5 as the LLM backbone.

Method	Avg.	Std	United States	Germany	Japan	Russia	Poland	Mexico	Nigeria
Single Role	58.63	2.758	64.5	62.81	56.47	56.65	59.71	59.4	58.84
Multi Consensus	59.92	2.573	65.17	64.99	58.24	59.21	62.63	60.45	59.94
Δ	1.29	-	0.67	2.18	1.77	2.56	2.92	1.05	1.10

These results collectively demonstrate the effectiveness and reliability of consensus elicitation (CulCT) from multiple LLMs. Moreover, note that this process does not rely on perfect persona simulation. We recognize precisely simulating individuals is an unresolved challenge, but it is NOT our purpose. Instead, CulCT only assumes the presence of diverse agents with sufficient cultural competence to identify shared cultural knowledge. The experiments above show that such approximated persona playing is sufficient to obtain cultural improvement.

B.2. Empirical Verification of Cultural Classifier and Error Bound

Error bound of Classifier Since the true culture distributions are unattainable, we introduce a classifier, which estimates the probability that a response y does NOT come from any of the K non-target cultures, and approximate the distinctiveness objective using Eq.(4). Correspondingly, we provide Proposition 3.2 as i) an interpretation of Eq.(4), showing that our method truly exploits the differences between true cultural distributions, and ii) a guarantee that our approximated implementation indeed optimizes the distinctiveness objective. Although Prop. 3.2 holds in theory, it also indicates that the tightness of the approximation error bound $\mathcal{E}$ depends on two classifier-related factors. We therefore conduct empirical verification about whether these conditions hold in practice.

First, we construct an evaluation set of 800 samples (x, y, p^*, c) covering four distinct cultures (200 samples for each), where (x, y) are question-response pairs sampled from our synthetic data and p^* are annotated and averaged across three native speakers per culture. Following the consensus-rating protocol described in Appendix D.5, we use a 5-point scale and convert the rating score into $p^* \in [0, 1]$ ($1 \rightarrow 0.1, 2 \rightarrow 0.3, 3 \rightarrow 0.6, 4 \rightarrow 0.85, 5 \rightarrow 0.95$). Using the classifier $\phi(x, y)$ implemented by p_ω and the openai text-embedding-3-small model, we measure the error bound ϵ and confidence η for each culture. The classifier error ϵ is calculated as the average $|\phi(x, y) - p^*|$ across the 200 samples, and the classifier confidence η as the max value of $\phi(x, y)$.

As shown in Table 10, the approximation error $\mathcal{E}$ is smaller than 0.05 for most countries. This is very small compared to the typical range of GJS, which spans $[0, \log K]$ for $K + 1$ distributions ($K = 14$ in this work, and the maximum of GJS is 2.64). Thus, the approximation is sufficiently accurate for practical use. Besides, the efficacy of Eq.(4) (distinctiveness optimization) has also been empirically verified by the strong overall performance of CAREDiO across all benchmarks, further supporting that Proposition 2 holds in practice.

Accuracy and Robustness of Classifier We further conduct experiments to assess the reliability of our classifier. Besides, we claim that the classifier can be implemented in multiple ways, thus we also compare the current embedding-based version with another alternative, i.e., llm-as-judge using GPT-5-thinking as the backbone. The results are shown in Table 10, where we report five metrics: i) average probability error $\epsilon = |\phi(x, y) - p^*|$; ii) maximum confidence η ; iii) the approximation error bound $\mathcal{E}$ mentioned in Proposition 2; iv) classification accuracy, Acc.; and v) classification F1.

We can see: a) *Our simple classifier based on cultural response similarity is highly accurate (with $F1 > 0.9$).* We infer this is because semantic differences between cultures are not small, even simple methods perform well, while previous work has

Table 10. Error bound and accuracy of cultural classifiers implemented with different architectures.

	Cultural Response Similarity				LLM-as-Judge (GPT-5-Thinking)			
	United States	China	Japan	Poland	United States	China	Japan	Poland
Accuracy	0.9051	0.9350	0.9450	0.9350	0.9650	0.9550	0.9700	0.9100
F1-score	0.9064	0.9366	0.9442	0.9319	0.9648	0.9577	0.9714	0.9100
ϵ	0.1682	0.1803	0.0973	0.0964	0.1981	0.1358	0.1206	0.1635
η	0.9519	0.9559	0.9304	0.9675	1.0000	1.0000	1.0000	0.8490
$\mathcal{E}$	0.0160	0.0330	0.0290	0.0485	0.7952	0.5452	0.5061	0.1902

largely ignored distinctiveness. b) Note that in Eq.(4) we do not use the binary prediction label, but rather the probability $\phi(x, y)$. Therefore, the approximation error bound $\mathcal{E}$ matters more, which is very small (smaller than 0.05) compared to its upper bound (2.64) in this work. c) LLM-as-judge with the advanced GPT-5-Thinking achieves a little better accuracy and F1, but also larger probability error ϵ , and thus leads to much worse error bound $\mathcal{E}$. This is because these off-the-shelf LLMs often fail to provide continuous score/probability.

Overall, these results above demonstrate that our simple classifier is sufficiently stable and accurate for our optimization objective. It is also worth noting that $\phi(x, y)$ can be implemented in different architectures. This is not our core contribution, but just an implementation of our novel optimization method (Algorithm 1). We can explore more structures in the future.

Sensitivity to Text Embedders To evaluate the sensitivity of our classifier to the underlying text embedding mode, we compare our default text-embedding-3-small against two widely adopted alternatives: gemini-embedding-2-preview and qwen3-embedding-8B. For each triplet (x, y, y_{c_k}) (x is the question, y and y_{c_k} are responses for the target and non-target cultures), we compute the similarity between the two responses with different embedders and quantify their statistical agreement. Across cultures, we observe strong consistency between embedders, as shown in Tab. 11. These high correlations indicate that different embedders yield highly consistent similarity rankings. Since our classifier relies on relative similarity, this consistency ensures stable decision for the distinctiveness objective. Therefore, our method is robust to the choice of embedding models.

Table 11. Statistical agreement between the default embedding model (text-embedding-3-small) and alternative embedders (gemini-embedding-2-preview and qwen3-embedding-8B).

Embedding Models	Spearman Correlation	Pearson Correlation
vs. gemini-embedding-2-preview	0.8934	0.8197
vs. qwen3-embedding-8B	0.8975	0.8105

C. Supplements for CARDSet Construction

C.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.
- **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.

C.2. Supplements for Optimization Prompts

CAReDiO is an in-context data optimization framework, without any training. The primary prompts used in the framework are illustrated in the following figures. Corresponding to Algorithm 1,

- Fig 6 shows the prompt for question initialization with the Self-Instruct approach (Line 3).
- Fig. 7 shows the prompt for response generation, maximizing representativeness and distinctiveness (Line 6).
- Fig. 8 shows the prompt for point-wise MI calculation in Eq.(2) (Line 7).
- Fig 9 includes the prompts for role-playing of individuals in three types.
- Fig 10 presents the prompt for question revision in Line 9.

We will also open-source the code and synthetic datasets for reproducibility.

You are a {country} culture expert with deep knowledge of cultures, values and social practices of {country} and other countries.

Given a cultural value topic from the Schwartz Theory of Basic Values, Hofstede's Cultural Dimensions, or World Value Survey, etc, your task is to generate 3-5 diverse questions or scenarios around this topic to explore the representative values, beliefs and social norms of {country} culture.

You should follow these guidelines for generation:

1. Topic Relevance: Question should explore cultural characteristics around the given topic.
2. Culture Consensus: Question should focus on values, beliefs and social practices that are representative, widely accepted and frequently appeared in {country} culture. Avoid niche, rare, or extreme viewpoints.
3. Culture Representativeness: Question should capture generalized values, beliefs and social practices of {country} culture that can be guidance across various situations, not be too specific to a single event or context.
4. Diversity: Each generated question should cover different aspects of the topic and differ from the provided example questions.
5. Neutral phrasing: Do not explicitly mention '{country}' in the question, so that it can be used to interview people from any cultural backgrounds to uncover the distinctiveness, i.e., avoid using like "In {country}, how do people...?" or "In {country} culture, what is the view on...?" in the question.

Generated questions should be one of the following types:

- Scenario-based question: Present a realistic value-related scenario and ask how people might react.
- Value-oriented question: Ask how people might prioritize or value certain principles or beliefs.
- Open-ended question: Encourage people to share their thoughts, feelings, and experiences related to a cultural aspect in detail.
- Likert-scale attitude question: Ask respondents to indicate their degree of agreement or priority regarding specific cultural value statements, like such as questions in World Value Survey or other surveys.

Here is the cultural topic that you need to generate questions for. (The provided definition is only a reference but not a complete description, you can further explore it based on your knowledge)

Cultural Topic: {topic}

Example Questions

[example 1]: {example_1}

[example 2]: {example_2}

Please generate 3-5 new questions following the guidelines above. Output exactly in the following format, without any other text.

[question 1]: <text of the question> (scenario-based question / value-oriented question / open-ended question / Likert-scale attitude question)

[question 2]: <text of the question> (...)

... (up to 5 questions)

The generated questions are:

""""

Figure 6. Prompt for initializing questions using the Self-Instruct approach.

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 7. Prompts for generating representative and distinctive responses.

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 8. Prompts for scoring the point-wise MI for each response.

C.3. Supplements for Question Formats

To align with the practical usage of LLMs, we consider four common question formats in our synthesis process, listed as follows.

- Scenario-based question: Present a realistic value-related scenario and ask how people might react.

General people with various demographics sampled from WVS

“Now you are a {country} citizen with the following demographic profile: {demographic_info} You understand {country} culture based on your own experiences and viewpoints as this persona, including the cultural values, beliefs and social practices. Your task is ...”

Cultural experts, such as sociologists and cultural psychologist

“You are a {country} culture expert, with 15 years of experience as a {role}. Thus, you have deep knowledge of {country} culture values, traditions and social practices. Your task is to ... from your own viewpoints as this expert”

Cross-cultural researchers

“You are a {foreign} culture expert, with 15 years of experience as a cross-cultural researcher. Thus, you have deep knowledge of cultural values, beliefs and social practices of various cultures across the world. Your task is to ... from your own viewpoints as this foreign culture expert”

Figure 9. Prompts for role-playing of individuals in three types, used in the point-wise MI calculation.

Table 12. 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

- Value-oriented question: Ask how people might prioritize or value certain principles or beliefs.
- Open-ended question: Encourage people to share their thoughts, feelings, and experiences related to a cultural aspect in detail.
- Likert-scale attitude question: Ask respondents to indicate their degree of agreement or priority regarding specific cultural value statements, like such as questions in World Value Survey or other surveys.

C.4. Supplements for Question Initialization

First, we leverage the LLM p_ω to initiate questions $\{x_i^0\}_{i=1}^N$ around diverse topics from core culture aspects, including cultural values, e.g., Hofstede Cultural Dimensions (Hofstede & Hofstede, 2005), norms and behavioral practices (See Appendix C.1 for topic details). Specifically, several questions are first generated for each topic, then we employ the Self-Instruct approach (Wang et al., 2023) to extend more distinct questions for each (100 in this work). Four common question formats are adopted to align with the practical usage of LLMs: scenarios-based, value-oriented, open-ended and multiple-choice questions (Detailed formats in Appendix C.3).

With questions $\{x_i^0\}_{i=1}^N$, we perform the iterative data optimization process as shown in Algo. 1. Concretely, we implement Eq.(3) by prompting p_ω to role-play a group of individuals from the target culture c . To incorporate comprehensive knowledge and enhance reliability, these individuals are set in three types: (i) *general people with various demographics* sampled from the WVS data of c ; (ii) *cultural experts with different backgrounds*, such as sociologists; and (iii) *cross-cultural researchers*. We use 15 general people, 5 cultural experts, and 3 cross-cultural researchers. We use different p_ω to synthesize multiple versions of CARDSet for comparison experiments in Tab. 2 and Fig. 4 (b).

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 10. Prompts for refining questions.

D. Supplements for Experimental Settings

D.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 and the statistics are reported in Tab. 12.

(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 compute the accuracy of the model’s prediction based on the ground truth.
- **WVS** (Haerpfer et al., 2020): 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., 2025): $Align(P_\theta, C) = (1 - \frac{Euclidean(\mathcal{A}_{P_\theta}, \mathcal{A}_C)}{max_distance}) \times 100$. $\mathcal{A}_{P_\theta}$ 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.

D.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.

D.3. More Details about Baselines

Role-Play: We instruct LLMs to simulate individuals from specific cultural backgrounds by providing a system prompt *“You are a {country} chatbot that knows {country} culture very well. Please answer the following questions according to your knowledge about the culture of country.”*.

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., 2025) 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), Cancele, 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.

D.4. 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 synthesize $N = 100$ questions for each topic at first. For representativeness optimization, we set 15 general people, 5 cultural experts and 3 cross cultural researchers.

D.5. Details about Human Evaluation

Human Recruitment We recruited three native annotators per country individually 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. The human evaluation task in Section 4.3 also follows the "Accuracy" annotation task. 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 Since this task is inherently subjective, we do not enforce a single "correct" label. Instead, we focus on minimizing noise due to the inconsistent understanding of the rubric. To this end, we provide annotators with the above fine-grained scoring rubrics and conduct mandatory training sessions before annotation. This ensures that disagreements stem from genuine cultural differences rather than misunderstandings of the guidelines.

We evaluate inter-annotator agreement across tasks and cultures, reporting the results in Table 13. The average pairwise agreement reaches 85%, while full agreement across all three annotators was achieved in approximately 65–70% of

Table 13. Inter-annotator agreement on different methods across cultures.

IAA	China	Japan	Poland
Role-Play (Qwen2.5-7B)	0.879	0.596	0.654
Role-Play (GPT-4.1)	0.867	0.571	0.624
CulturePark	0.848	0.652	0.752
CAReDiO	0.806	0.880	0.795
Avg	0.850	0.675	0.706

cases, indicating a strong agreement for subjective NLP/LLM evaluation tasks (with most values exceeding 0.7). These results demonstrate that our annotation process is reliable and that annotator divergence remains within an acceptable and meaningful range. For each sample, we applied a majority-vote strategy to merge individual ratings into the final label.

D.6. Discussion on Data Leakage

As shown in the third paragraph of Sec. 4.3, we explicitly mention this potential data leakage and the possible inflated gains caused by it. We need to clarify that our data is fully synthesized from scratch and does not rely on any existing cultural benchmarks (e.g., survey-style WVS). This differs from prior work (e.g., CultureLLM, CulturePark), which directly use or augment WVS, leading to a risk of data leakage when evaluated on similar benchmarks.

In this case, we still keep these questionnaire-style benchmarks in our evaluation for two reasons below.

(1) **Comparing the impact of leakage by WVS.** Several baselines, e.g., CultureLLM, CulturePark, CultureSPA, are structurally integrated with WVS, and it’s infeasible to remove WVS-related data without fundamentally altering these baselines. However, CAReDiO uses only WVS demographics, not its questions or responses. As discussed in Sec.4.3 (Line 482-485), the evaluation on WVS helps show the limitation of baselines, which perform well only on these two benchmarks but fail on other OOD ones, indicating potential leakage and overfitting.

(2) **Comparing the generalization ability of different methods.** To mitigate possible leakage from WVS, we take two actions: (i) **We remove the WVS split in GlobalOpinionQA** and only keep the GAS subset for evaluation. (ii) **We intentionally introduce two additional benchmarks in different formats for comparison.** As shown in Table 2, our method achieves significantly larger gains on these non-WVS benchmarks than on WVS/GlobalOpinionQA, while baselines, especially the three rooted in WVS, only perform well on WVS (on non-WVS sets, even worse than Role-Play). This demonstrates CAReDiO’s improvements are not driven by leakage and supports its robustness and generality.

E. Supplements for Results

E.1. Supplements for Cultural Dataset Analysis

To demonstrate how the two objectives in CAReDiO help organize the internal knowledge of the LLM to synthesize better cultural data, we conduct a low-dimensional visualization of four groups of cultural samples: i) the cultural data directly generated by the backbone LLM; ii) optimized for balanced representativeness and distinctiveness; iii) only representativeness and iv) only distinctiveness. We compute the distributional metrics as shown in Tab. 14.

From these results, we form several key observations:

- The original distribution is highly scattered, as evidenced by the large std and pairwise distance. This indicates a limitation in covering the most important information within a small budget.
- All datasets optimized for only rep, only dist or both have obvious shifts from the original data distribution to be closer, while optimizing both objectives achieves the most significant shift. Observing from the visualization, we find the original data scatters around while the data optimized for both objectives shape a clearer cluster, which we infer highlights the core of the culture.
- Besides, we observe only dist shows a lower diversity than only rep. From the visualization, we see there are some small clusters in data optimized for distinctiveness. We infer it could highlight those distinctive ones.

Table 14. Quantitative metric changes under different optimization objectives in the low-dimensional embedding space.

Metric	General (Base)	Only Rep.	Only Dist.	Rep. + Dist. (CARDSet)
Shift Distance $\uparrow$	-	0.3078	0.3035	0.3253
Standard Deviation $\downarrow$	0.7605	0.6982	0.6815	0.6680
Pairwise Distance $\downarrow$	0.5841	0.4925	0.4696	0.4508

E.2. Supplements for Cultural Alignment Performance

In addition to reporting an average score over four benchmarks in Tab. 2, we conduct a Friedman test with Nemenyi post-hoc analysis to compare the average ranks of all methods and visualize the results through the Critical Difference (CD) diagram in Fig. 11. Our method achieves the best average rank of 2.0, outperforming all baselines (which rank 2.5, 3.83, 4.33, 4.67, 6.33, 7.00, 7.17, and 7.17, respectively). The Friedman test confirms a statistically significant difference among the compared methods ($p = 0.0011 < 0.05$). The Nemenyi post-hoc analysis further indicates that our method significantly outperforms several key baselines, including the backbone model, CulturePark, and CultureSPA.

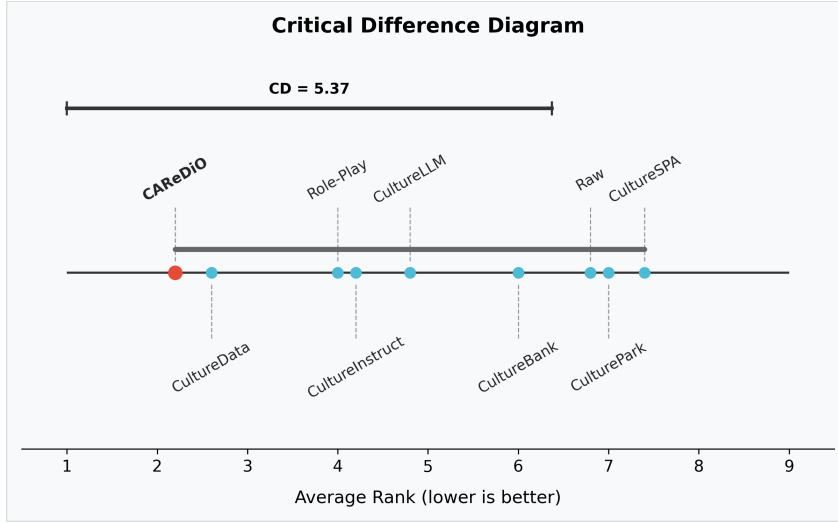


Figure 11. Critical Difference (CD) diagram representing the statistical significance of CAReDiO and alternative baselines ($p = 0.0011$).

Cultural alignment performance with Llama-3.1-8B-Instruct as the backbone is listed in Tab. 15. Since Prism is originally evaluated on a 1-5 Likert scale, we first rescale it to by multiplying by 20. Then, the ‘Average’ score is computed as the unweighted mean across all five benchmarks: CultureBenchmark-Easy, CultureBench-Hard, Prism, GlobalOption and WVS.

Table 15. 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 16, Table 17, Table 18, Table 19 and Table 20.

Table 16. Cultural alignment performance across various cultures on the GlobalOpinionQA dataset. Best results are in **bold**.

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

Table 17. Cultural alignment performance across various cultures on the WVS dataset. Best results are in **bold**.

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

CAReDiO: Cultural Alignment via Representativeness and Distinctiveness Guided Data Optimization

Table 18. Cultural alignment performance across various cultures on the CulturalBench-Easy dataset. Best results are in **bold**.

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

Table 19. Cultural alignment performance across various cultures on the CulturalBench-Hard dataset. Best results are in **bold**.

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 20. Cultural alignment performance across various cultures on the Prism dataset. Best results are in **bold**.

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

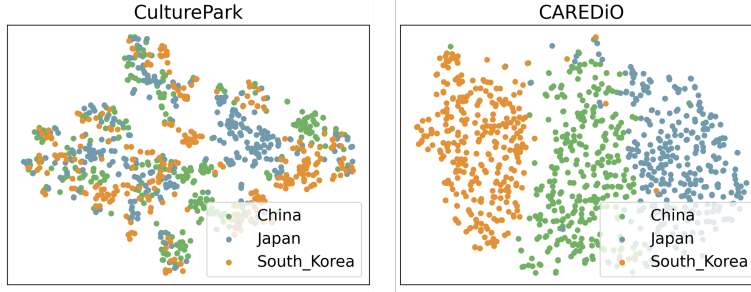


Figure 12. TSNE visualization of cultural alignment data for related cultures China/Japan/Korea, from CulturePark and CAREDiO respectively.

E.3. Supplements for Ablation Study

We perform an ablation study to evaluate the independent contributions of the representativeness and distinctiveness objectives in our framework. The main results and analysis are reported in Sec.4.4. Here, we further explain the effectiveness of CAREDiO from a more intuitive aspect.

More Intuitive Discussion about Gains To help readers understand, we provide a more intuitive explanation for why CAREDiO, especially the two objectives, can outperform other methods on cultural alignment.

Following the line of synthesizing cultural data with LLMs, we acknowledge two key assumptions (which often hold in practice): (i) The backbone LLMs, e.g., GPT-4.1, have encoded sufficient knowledge about specific cultures and can be elicited in a certain way. (ii) The cultural knowledge distribution learned by the LLMs is imperfect and biased, which can be more serious in weaker models. Based on these assumptions, we observed that most prior methods directly apply simple role-playing, and usually produce biased or ambiguous data.

This improvement by CAREDiO is obtained by explicitly shaping the data distribution along two theoretically grounded criteria derived from cultural theory:

- *Representativeness*: CAREDiO utilizes multi-LLM ensemble to maximize the shared understandings of the target culture while minimizing irrelevant noise (corner cases) of generated samples (Eq.(2)-(3)), which helps mitigate the bias of single LLM role-playing. Thus, optimizing only representativeness yields improvements. In Table 2, with (weaker) Qwen2.5-7B-Instruct as backbone, on difficult tasks (CB-Hard), Role-Play performs much worse than CAREDiO (40.20 > 36.73).
- *Distinctiveness*: CAREDiO captures characteristics unique to the target culture rather than generic patterns shared across related cultures (Eq.(4)). This further distinguishes ambiguous data samples located near cultural boundaries. This is verified in Appendix E.4. For closely related cultures, e.g., Japan/Korea/China, CAREDiO achieves larger performance gaps between target and non-target cultures, $\Delta=5.3 > 2.4$, while ensuring better performance on each target one.

Therefore, through CAREDiO is an in-context optimization framework, its concrete designs are grounded in information-theoretic derivation to produce higher-quality cultural data than baselines, explaining its superior empirical performance.

E.4. Supplements for Distinctiveness Analysis among Related Cultures

We incorporate the distinctiveness objective in CAREDiO to help identify clearer cultural boundaries and resolve ambiguity. To validate its effectiveness, we conduct a focused analysis on culturally proximate regions, i.e., China, Japan and Korea, and evaluate whether CAREDiO improves fine-grained distinctiveness among them. Analysis is conducted from two perspectives below.

Distinctiveness of synthesized training data We embed CAREDiO-synthesized data for China, Japan, Korea, and those from the CulturePark benchmark, using OpenAI text-embedding-3-small API, then conduct TSNE dimensionality reduction. The visualization is shown in Fig. 12. We observe that **CAREDiO produces much more clearly separated clusters, while CulturePark exhibits substantial cross-cultural overlap**. To quantify this, we compute the average inter-cluster centroid

distance: CAReDiO achieves 0.47 compared to CulturePark’s 0.29. This substantial margin indicates that CAReDiO’s synthesized data provides significantly finer distinctiveness than the baseline.

Table 21. Confusion matrices of cross-cultural evaluation on GlobalOpinionQA and Prism, for CulturePark and our CAReDiO framework.

(a) CulturePark on GlobalOpinionQA						(b) CAReDiO on GlobalOpinionQA					
	China	Japan	Korea	Δ Acc $\uparrow$	Conflict Acc $\downarrow$		China	Japan	Korea	Δ Acc $\uparrow$	Conflict Acc $\downarrow$
China	52.41	53.35	51.59	-0.06	38.34	China	53.88	50.84	49.47	3.73	34.69
Japan	42.74	53.38	52.18	5.92	30.86	Japan	45.25	54.56	53.70	5.08	33.73
Korea	49.47	50.47	51.41	1.44	35.13	Korea	45.23	45.54	52.46	7.07	27.29

(c) CulturePark on Prism					(d) CAReDiO on Prism				
	China	Japan	Korea	Δ Rating $\uparrow$		China	Japan	Korea	Δ Rating $\uparrow$
China	1.0	0.7283	0.7214	0.2752	China	1.0	0.7122	0.6948	0.2965
Japan	0.7283	1.0	0.7297	0.2710	Japan	0.7122	1.0	0.6756	0.3061
Korea	0.7214	0.7298	1.0	0.2744	Korea	0.6948	0.6756	1.0	0.3148

Cross-Cultural Confusion Matrices We further evaluate models aligned to each target culture using either CAReDiO or other baseline datasets, then measure how each model performs on non-target cultures to derive confusion matrices. We assessed two cultural benchmarks: GlobalOpinionQA and Prism.

Table 21(a) and (b) illustrate the detailed pairwise confusion matrices on GlobalOpinionQA against CulturePark. The first three columns report the accuracy on each culture; Δ is the average performance gap between the target culture and non-target cultures. As for Con.Acc, we extract the subset from GlobalOpinion where the non-target culture shows different answers and report the model’s alignment with the answers from the non-target culture, thus the lower the better.

We can see: (i) *Models aligned via CAReDiO perform better on the target culture while worse on non-target cultures*, achieving improved cultural specificity than CulturePark (larger performance gaps Δ , $5.3 > 2.4$). ii) On the conflicting subset (different answers are expected from distinct cultures), CAReDiO models exhibit lower alignment on non-target cultures (Acc. on Conflicting, $31.9 < 34.8$).

Table 21(c) and (d) present the evaluation results on the Prism dataset. Using 200 open-ended questions sampled from the Prism dataset, we measured pairwise similarity between responses produced by China-, Japan-, and Korea-aligned models. CAReDiO models display lower cross-cultural similarity (higher Δ) than CulturePark ($0.31 > 0.27$), confirming better separation across related cultures.

Comprehensive Comparison across All Baselines To ensure evaluating fairness, we extend our distinctiveness evaluation to all remaining baselines and summarize the definitive distinctiveness metrics (Δ and Conflict Acc) in Tab. 22. On GlobalOpinionQA, CAReDiO gets the largest Δ and lower Con.Acc, showing better alignment and reduced confusion. As for Prism, CAReDiO shows better distinctiveness ($\Delta=0.3058$) than synthesis baselines and manual CultureData. CultureInstruct uses substantially larger data (46,878), making it less directly comparable. Results consistently support our conclusions.

Across all analyses, cluster distinguishability, centroid distances, confusion matrices, and cross-cultural response similarity, **CAReDiO consistently demonstrates stronger ability to make fine-grained distinctions between closely related cultures.**

E.5. Analysis of Results Variance

Detailed per-culture evaluation results across all four benchmarks have been included in Table 18, 19, 20, 16 and Table 17. To address the concern that the reliability of major conclusions delivered by the average results in Table 2 could be impacted by the high variance raised from particular cultures or question types, we analyze the mean and std of CAReDiO’s performance across all benchmarks and cultures.

Table 22. Comprehensive distinctiveness evaluation across all cultural alignment baselines. Δ metrics measure the performance gap between target and non-target cultures ($\uparrow$ higher is better), while Conflict Acc denotes the accuracy on the conflicting subsets ($\downarrow$ lower is better). The best and second-best results are highlighted in **bold** and underlined.

Method	GlobalOpinionQA Δ Acc ($\uparrow$)	GlobalOpinionQA Conflict Acc ($\downarrow$)	Prism Δ Rating ($\uparrow$)
CultureLLM	3.36	34.6	0.26
CulturePark	<u>4.21</u>	34.8	0.27
CultureSPA	0.20	32.9	0.22
CultureBank	3.88	33.1	<u>0.31</u>
CultureInstruct	3.41	<u>32.9</u>	0.35
CultureData	3.53	32.4	0.29
CAREDiO (Ours)	5.30	31.9	<u>0.31</u>

Table 23. The average and standard deviation of improvement over the backbone LLM across cultures.

	Average	Std	% of Improved Culture	% of Decreased Culture
CB-Easy	1.47	6.88	73.30%	26.70%
CB-Hard	1.30	4.85	73.30%	26.70%
Prism	35.35	4.61	100%	0
GlobalOpinionQA	2.95	3.10	100%	0
WVS	5.00	2.47	100%	0

Average and standard deviation of gains across cultures As shown in Table 23, on three benchmarks (Prism, GlobalOpinionQA and WVS), CAREDiO yields small/acceptable variance (std) compared to the average of the improvement over the backbone LLM. Besides, we also observe CAREDiO achieves improvements on most cultures, e.g., 100% improvement ratio on Prism, GlobalOpinionQA and WVS, 73.3% on CultureBench).

Table 24. The improvement ratio of CAREDiO across each question type and cultures.

Question Type	US	UK	Germany	Italy	China	Japan	South Korea	India
Multi-Choice Knowledge	2.60%	-4.76%	13.63%	14.69%	-9.73%	0.78%	-0.87%	-1.01%
Value Questionnaire	4.24%	5.41%	6.73%	3.49%	12.25%	3.06%	-3.26%	8.98%
Open-ended Question	62.12%	69.82%	93.44%	89.32%	101.41%	60.42%	84.38%	71.33%

Question Type	Singapore	Indonesia	Russia	Poland	Romania	Mexico	Nigeria
Multi-Choice Knowledge	-8.33%	8.88%	4.76%	11.78%	12.49%	4.68%	8.12%
Value Questionnaire	18.17%	-1.88%	13.49%	5.34%	6.72%	10.51%	2.07%
Open-ended Question	-	92.86%	94.45%	89.63%	73.83%	105.23%	99.34%

Improvement ratio across each question type We classify all benchmarks into three categories of question type: i) Multi-choice question about culture knowledge (CB-Easy, CB-Hard), ii) Value Questionnaire (GlobalOpinionQA and WVS) and iii) Open-ended question (Prism). As shown in Table 24 and Table 2, CAREDiO obtains consistent improvements on most cultures and question types, with two general conclusions:

(1) CAREDiO performs better on moderately low-resource cultures, e.g., Germany and Russia. This is because our representativeness objective (Eq.(3)) helps elicit consensus and mitigates LLMs’ inherent bias towards minority cultures.

(2) CAREDiO achieves the most substantial gains on open-ended questions, with improvements that are statistically significant (t -test, $p < 0.05$). This is because our synthesized data takes the form of (x, y) QA pairs. We adopt this format due to that most baselines are rooted in multiple-choice questions (e.g., CulturePark and CultureSPA) and tend to overfit to that format, and thus fail to generalize to more realistic open-ended QA tasks. However, CAREDiO still shows improvements across diverse question types, indicating that this form of data offers better generalization.

For questionnaire-style benchmarks, CAREDiO shows statistically significant improvements over role-playing, CultureBank,

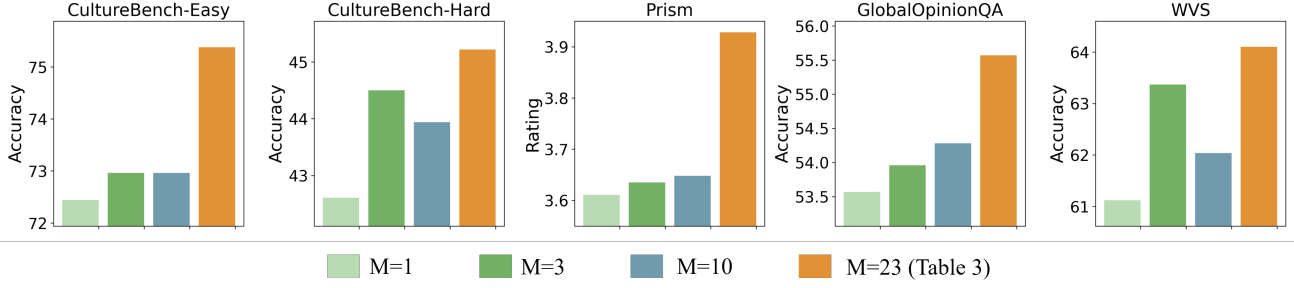


Figure 13. Performance comparison of CAREDiO with different numbers of simulated individuals (i.e., M in the figure) for representativeness optimization.

and CultureInstruct baselines on both datasets. For multiple-choice knowledge questions, improvements on the easy subset are not statistically significant, which is expected since these questions can largely be solved without alignment. On the more challenging CB-Hard subset, CAREDiO achieves statistically significant gains over most baselines, indicating its benefit in harder settings.

These findings confirm that our major conclusion—that CAREDiO consistently and significantly enhances cultural alignment—is robust and not driven by high-variance artifacts in specific cultures or question types.

E.6. Number of Individuals in Representativeness Optimization

Drawing on the Cultural Consensus Theory, we incorporate multiple LLM-simulated individuals to perform consensus elicitation for representativeness optimization. We conduct experiments to analyze the impact of the number of individuals involved in this process, i.e., the hyper-parameter M in Eq.(3). Again, we use Qwen2.5-7B-Instruct for both data synthesis and for being the aligned model, and report the results averaged across multiple countries (Germany, Italy, China, Japan, Russia and Nigeria).

As shown in Fig. 13, *introducing multiple LLM-simulated individuals indeed significantly enhances the performance*, demonstrating the effectiveness of consensus elicitation for optimizing representativeness and mitigating the bias embedded in the original LLM. *As a larger number of individuals become available, the results are further improved*. Whereas, when the computation resource is limited, a small number of roles can also produce results much better than the single version.

Table 25. Statistics of cultural datasets. Cult. Points: Culture points representing distinctive cultural aspects extracted from the dataset by GPT-4.1; Sim and SB are cosine similarity and Self-BLEU within each dataset; Cult. Sim is cosine similarity between subsets of different cultures.

Datasets	Source	#sample	Avg.L $\uparrow$	#Cult. Points $\uparrow$	Sim $\downarrow$	SB $\downarrow$	Cult. Sim $\downarrow$
CulturePark	WVS augmentation	1,000 each	68.6	494.6	<u>0.235</u>	0.406	0.223
CultureData	manual annotation	1,000 each	16.8	1521.0	0.199	0.330	0.127
CARDSet (embed sim)	LLM-synthetic	1,000 each	200.4	2027.0	0.251	<u>0.324</u>	<u>0.202</u>
CARDSet (llm-judge)	LLM-synthetic	1,000 each	<u>192.0</u>	<u>2021.0</u>	0.257	0.321	0.205

Table 26. Alignment performance of CAREDiO using different classifiers to implement the distinctiveness objective. The best and second-best results are highlighted in bold and underlined.

Method	CB-Easy	CB-Hard	Prism	GlobalOpinion	WVS	Average
Qwen2.5-7B-Instruct	74.85	<u>42.38</u>	2.154	53.65	61.70	53.49
Role-Play	<u>76.77</u>	41.21	3.471	56.79	66.01	61.05
CAREDiO (embed sim)	76.92	42.04	4.002	<u>57.55</u>	66.43	64.13
CAREDiO (llm-judge)	76.34	42.59	<u>3.975</u>	57.75	<u>66.25</u>	<u>64.05</u>

E.7. Sensitivity Analysis of Distinctiveness Classifier

Since the true culture distributions are unattainable, we introduce a cultural classifier to approximate the distinctiveness objective following Proposition 3.2. This classifier can be implemented in multiple ways, provided that it offers reliable signals for distinguishing between cultures. In Appendix B.2, we compare our default text embedding-based classifier with an alternative llm-as-judge classifier, GPT-5-Thinking, and show that both achieve sufficiently high differentiation accuracy with no substantial discrepancies.

To further examine whether the downstream data synthesis and cultural alignment performance depend on the classifier architectures, we conduct an additional sensitivity experiment. Using Qwen2.5-7B-Instruct as both the data generator and the backbone LLM being aligned, we synthesize culture-specific datasets with two classifier variants: i) *text embedding-based classifier*: the generator paired with OpenAI text-embedding-3-small model; and ii) *LLM-as-judge classifier*: the generator acting as the classifier (we use this backbone Qwen2.5-7B-Instruct here rather than GPT-5-Thinking to ensure a fair comparison with the embedding-based setup). Then, we obtain two sets of culture-specific data for alignment, denoted as *CAReDiO (embed sim)* and *CAReDiO (llm-judge)* respectively. We experiment on six diverse cultures, i.e., US, UK, Germany, China, Russia and Mexico. (randomly selected these six cultures due to resource and time constraints.)

With these data for cultural alignment, we first compare their statistics. As shown in Table 25, both datasets contain richer cultural information than baselines. Then, we evaluate their alignment performance on the four distinct cultural benchmarks. Table 26 illustrates the scores averaged over the 6 cultures, again using Qwen2.5-7B-Instruct as both the data generator and the backbone LLM being aligned. We observe that: (i) *datasets synthesized under our framework, either using the embedding-based classifier or the llm-as-judge classifier, achieve significant improvements over the baseline*; and ii) the two classifier variants do not lead to significant differences. All these results demonstrate that *our framework is not sensitive to the classifier choice* as long as it can provide accurate feedback. This further indicates that the core contribution of our framework is the novel information-theoretic optimization framework, while various practical classifier implementations can serve as effective instantiations.

E.8. Analysis of Token or Human Annotation Cost

Regarding the absolute token or human annotation cost, it is difficult to standardize across methods due to two reasons. i) Most baselines rely on extensive data synthesis and filter an effective small set, such as CultureSPA. ii) Human-annotated datasets bypass LLM token costs entirely but require substantially higher cost for human labor, expert coordination and crowd-sourcing. For transparency, we quantify the empirical cost of deploying our CAReDiO framework. Synthesizing 1k samples per culture with CAReDiO needs 42.25M input and 12.5M output tokens, i.e., \$177.8 using GPT-5, which is modest and notably lower than human annotation.

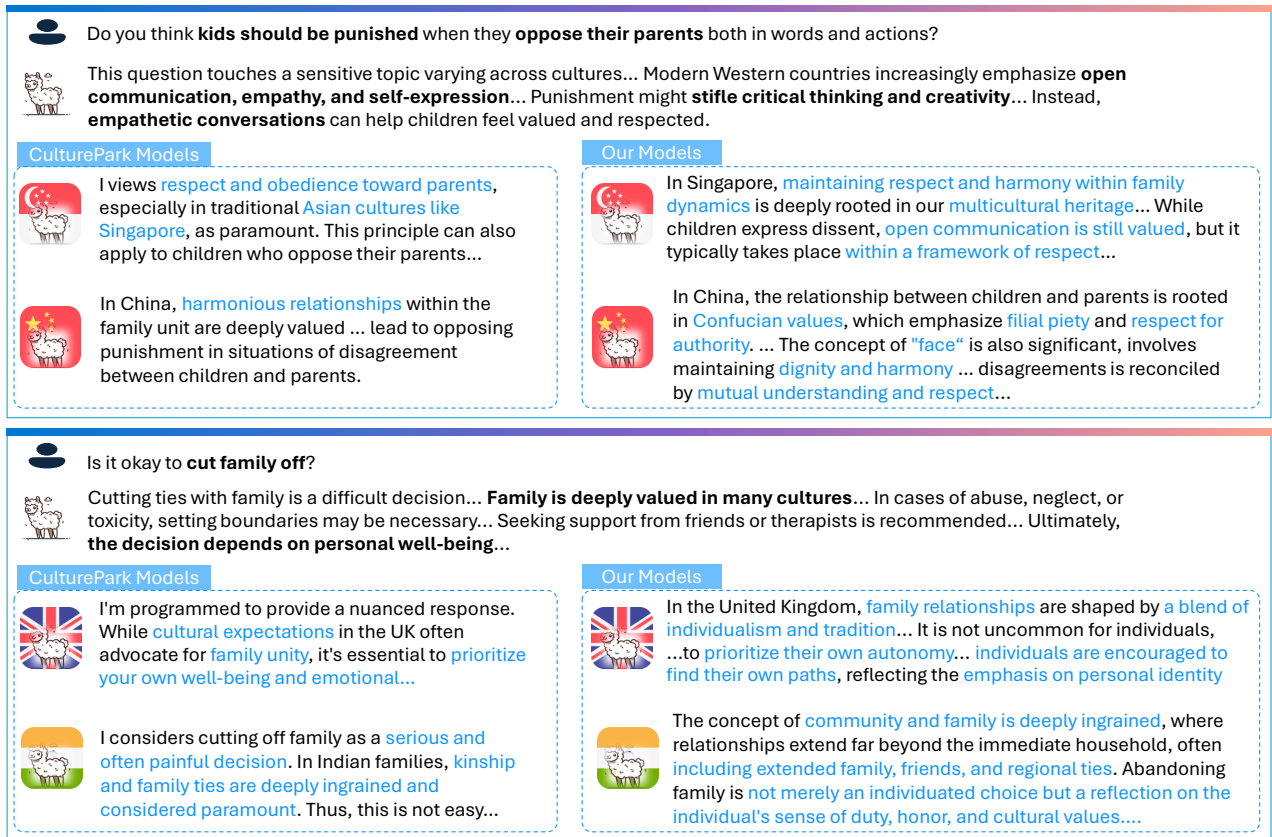
E.9. Case Study

Case Study We present case studies of culturally sensitive topics in Figure 14. Without cultural alignment, the original Llama-3.1-8B-Instruct usually returns general responses that lack cultural specificity. Due to the predominance of English-language training data, its response sometimes demonstrates a bias towards Western perspectives, which underscores the importance of cultural alignment to ensure the inclusivity of AI. We find that culture-specific models exhibit significantly improved adaptation to their respective cultural 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.

We conduct case studies to reveal the effectiveness of our framework for cultural alignment. More examples are presented in Tab. 27 and Tab. 28.

F. 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





Is it okay to **cut family off**?



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**...

CulturePark Models



I'm programmed to provide a nuanced response. While **cultural expectations** in the UK often advocate for **family unity**, it's essential to **prioritize your own well-being and emotional**...



I considers cutting off family as a **serious and often painful decision**. In Indian families, **kinship and family ties are deeply ingrained and considered paramount**. Thus, this is not easy...

Our Models



In the United Kingdom, **family relationships** are shaped by **a blend of individualism and tradition**... It is not uncommon for individuals, ...to **prioritize their own autonomy**... **individuals are encouraged to find their own paths**, reflecting the **emphasis on personal identity**



The concept of **community and family is deeply ingrained**, where relationships extend far beyond the immediate household, often **including extended family, friends, and regional ties**. Abandoning family is **not merely an individuated choice but a reflection on the individual's sense of duty, honor, and cultural values**....

Figure 14. Case studies on cultural alignment.

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.

G. Limitations

In this paper, we propose a novel data optimization framework to construct cultural datasets rich in representativeness and distinctiveness. Extensive experiments across 15 cultures and four diverse LLM backbones have verified the effectiveness of CAReDiO. Furthermore, we outline several limitations of our work and discuss potential future directions as follows.

(1) Our optimization framework currently relies on LLMs to generate cultural data. Thus, it is unavoidably affected by the imperfect and imbalanced cultural knowledge distribution embedded in current LLMs, and may struggle to collect data accurately enough for extremely low-resource cultures. This is a critical challenge shared across all LLM-based synthetic approaches. Instead, our method specially designs the mechanism to elicit more representative and distinctive cultural knowledge from LLMs, which helps to mitigate the challenge to some extent. Moreover, we also conduct experiments on some low-resource cultures, e.g., Indonesia and Nigeria, and observe improvements, empirically showing the effectiveness of our method to mitigate this limitation.

We are also considering two approaches that could be explored to alleviate this issue in the future. First, we can instantiate our CAReDiO framework with more advanced LLMs that might contain richer knowledge for low-resource cultures and leverage our algorithm to elicit satisfactory data. Second, based on the effectiveness of two core principles verified by us, i.e., representativeness and distinctiveness, we can extend CAReDiO to integrate human annotations, enabling collection of representative and distinctive data for cultures with limited knowledge in current LLMs.

(2) Due to constraints in computational budget and API resources, our experiments currently cover 15 distinct cultures from various regions. These have fully demonstrated this potential of CAReDiO. Given the vast diversity and complexity of global cultures, 15 groups represent only a fraction of the world’s pluralistic values and we are considering scale it to 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. Future iterations could introduce temporal weighting mechanisms into our information gain objective, allowing the framework to capture both historic cultural anchors and real-time societal shifts.

(4) Currently, we follow baselines to use supervised fine-tuning for downstream cultural alignment. But it is easy to collect dispreferred responses in the data optimization process. Thus, we can explore more effective fine-tuning techniques such as DPO.

(5) This paper conducted experiments merely in English. Though a substantial portion of cultural knowledge can be expressed cross-linguistically, certain norms, values, and practices are tightly coupled with native languages. Besides, there have already been some studies exploring the impact of language for cultural alignment (Mukherjee et al., 2024; Choenni et al., 2024). While the improvements achieved by CAReDiO demonstrate that English-based elicitation is already beneficial, exploring native-language prompting and multilingual extensions remains an important future direction that is orthogonal to our focus in this paper.

Table 27. Case studies to demonstrate the effectiveness of cultural alignment, comparing our method and baselines.

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 28. Model case studies.

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