# CARE: Multilingual Human Preference Learning for Cultural Awareness

**Geyang Guo**[α], **Tarek Naous**[α], **Hiromi Wakaki**[β], **Yukiko Nishimura**[β], **Yuki Mitsufuji**[β,γ], **Alan Ritter**[α], **Wei Xu**[α]
[α]Georgia Institute of Technology, [β]Sony Group Corporation, [γ]Sony AI
{guogeyang, tareknaous}@gatech.edu
{hiromi.wakaki, yukiko.b.nishimura, yuhki.mitsufuji}@sony.com
{alan.ritter, wei.xu}@cc.gatech.edu

## Abstract

Language Models (LMs) are typically tuned with human preferences to produce helpful responses, but the impact of preference tuning on the ability to handle culturally diverse queries remains understudied. In this paper, we systematically analyze how native human cultural preferences can be incorporated into the preference learning process to train more culturally aware LMs. We introduce **CARE**, a multilingual resource containing 3,490 culturally specific questions and 31.7k responses with human judgments. We demonstrate how a modest amount of high-quality native preferences improves cultural awareness across various LMs, outperforming larger generic preference data. Our analyses reveal that models with stronger initial cultural performance benefit more from alignment, leading to gaps among models developed in different regions with varying access to culturally relevant data. CARE is publicly available at https://github.com/Guochry/CARE.


Figure 1: Example LM responses to culture-specific questions in the native languages. The base LM (🦙) `Llama3.1-8B` fails to respond appropriately, while its aligned versions on CARE generate better responses for Chinese (🦙), Arab (🦙), and Japanese (🦙) cultures.

## 1 Introduction

After large-scale pre-training, Language Models (LMs) typically undergo a crucial post-training phase (aka "alignment"), which includes supervised fine-tuning and preference tuning, to improve their ability to follow instructions and alignment with human preferences (Ouyang et al., 2022; Rafailov et al., 2024). However, *the effects of alignment on the cultural awareness* of multilingual LMs (i.e., how well they understand and generate appropriate responses to diverse, culturally relevant queries) remains understudied, partly because frontier open-weight models with multilingual support (e.g., Llama-3, Qwen-2.5) have only become available very recently around mid-2024. Consequently, existing studies have either examined off-the-shelf aligned LMs (e.g., Aya, Tulu 2) (Naous et al., 2024; Ryan et al., 2024) or explored instruction fine-tuning with culturally relevant data (Li et al., 2024a).

In this paper, we systematically analyze the extent to which and under what conditions preference optimization (e.g., DPO, KTO, SimPO) can help LMs align with regional user expectations along five key dimensions: cultural entities, cultural commonsense, social norms, opinions, and literacy (see examples in Figure 1). To enable our study, we create CARE, *a multilingual culture-specific*



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32855–32884
November 4-9, 2025 

| Dataset | Culture-Specific | Topics | Manually-Curated | Human[1] Preferences | Languages | Format |
|---|---|---|---|---|---|---|
| *Cultural Evaluation* | | | | | | |
| Fork (Palta and Rudinger) | ✓ | food-related customs | ✓ | ✗ | en | multiple-choice |
| CulturalBench (Chiu et al.) | ✓ | daily life, social etiquette, wider society | ✗ | ✗ | en | multiple-choice |
| GeoMLAMA (Yin et al.) | ✓ | cultural commonsense | ✓ | ✗ | en, fa, hi, sw, zh | multiple-choice |
| BLEnD (Myung et al.) | ✓ | food, sports, family, education, holidays, work-life | ✓ | ✗ | 13 languages | multiple-choice, free text |
| Include (Romanou et al.) | ✓ | general knowledge, social science, professional certifications, etc. | ✓ | ✗ | 44 languages | multiple-choice |
| CAMeL (Naous et al.) | ✓ | beverage, clothing, food, location, religion, sports, etc. | ✓ | ✗ | ar | masked prompts |
| CANDLE (Nguyen et al.) | ✓ | geography, religion, occupation, food, clothing, etc. | ✗ | ✗ | en | assertions |
| LLM-GLOBE (Karinshak et al.) | ✓ | cultural values | ✓ | ✗ | zh, en | multiple-choice, free text |
| CulturalTeaming (Chiu et al.) | ✓ | red-teaming cultural questions | ✓ | ✗ | en | multiple-choice |
| SHADES (Mitchell et al.) | ✓ | culture-specific stereotypes | ✓ | ✗ | 16 languages | stereotyped statements |
| JMMMU (Onohara et al.) | ✓ | art, heritage, history, business, science, medicine, etc. | ✓ | ✗ | ja | multiple-choice |
| *Cultural Fine-tuning* | | | | | | |
| Aya (Singh et al.) | ✓✗ | news, stories, recipes, scientific texts, etc. | ✓ | ✗ | 65 languages | free text |
| CIDAR (Alyafeai et al.) | ✓✗ | technology, translation, poetry, grammar, etc. | ✗ | ✗ | ar | free text |
| Cameleval (Qian et al.) | ✗ | information provision, reasoning, creative writing, etc. | ✓ | ✓ | ar | free text |
| PRISM (Kirk et al.) | ✓ | cross-cultural controversies | ✓ | ✓ | en | multi-turn conversations |
| Palm (Alwajih et al.) | ✓ | history, celebrations, sports, literature, etc. | ✓ | ✗ | ar | free text |
| CULTUREINSTRUCT (Pham et al.) | ✓ | art, cuisine, cultural norms, festivals, history, etc. | ✗ | ✗ | en | free text |
| CultureBank (Shi et al.) | ✓ | social norms, food, communication, festivals, etc. | ✗ | ✗ | en | free text |
| CulturePark (Li et al.) | ✓ | human belief, norm, custom | ✗ | ✗ | en | free text |
| CultureLLM (Li et al.) | ✓ | value survey | ✗ | ✗ | 9 languages | multiple-choice |
| **CARE (Ours)** | ✓ | cultural entities, opinion, norms, commonsense, literacy | ✓ | ✓ | ar, ja, zh | free text |

Table 1: Comparison of datasets for studying LMs' cultural awareness. CARE is a multilingual, *human-annotated preference* dataset specifically grounded in culture. (✓✗) indicates resources that include some cultural considerations (in the form of culture-related questions or by recruiting native annotators), while cultural coverage is not their only or primary focus (see § 3.1). Representative examples from each dataset are provided in Appendix A.5.

*dataset with human preference judgments* on 31.7k LLM/Human-written responses to 3,490 questions about Chinese, Arab, and Japanese cultures. Questions are drawn from a variety of instruction tuning datasets (Singh et al., 2024; Alyafeai et al., 2024) and evaluation resources (Chiu et al., 2024b; Palta and Rudinger, 2023) that we manually identified as culturally-relevant. We further collected additional questions to improve the coverage on regionally popular but less digitally documented commonsense knowledge and social norms, with help from native speakers who have overseas experience.

Leveraging CARE, we show that a *modest, high-quality set of native preferences* can enhance LMs' cultural awareness across model families and sizes, achieving competitive or stronger performance than generic preference data despite using 3–17× less data, and generalizing to out-of-domain cultural tasks (Myung et al., 2024; Romanou et al., 2025). We then conduct controlled studies on various LMs and uncover several important insights for cultural alignment: **(1)** Base model strength matters: aligning base models with stronger initial cultural awareness such as Qwen2.5-7B and Gemma2-9B lifts scores across all cultural categories, whereas models with weaker initial performance hardly improve after alignment, consistent with the post-training analyses of Ivison et al. (2023). **(2)** Data efficiency scales: even 25% of CARE can lead to a 74% improvement, with performance continuing to rise as the dataset scales to full size. **(3)** Cross-culture synergy exists: mixing samples about local and foreign cultures in preference learning improves overall awareness of local cultures (e.g., Chinese 4.691 → 4.944, Arabic 2.980 → 3.538).

Beyond these insights, we observe that performance on different cultural dimensions varies across model families. For example, the Chinese-centric Qwen2.5-72B (Yang et al., 2024a; Wen-Yi et al., 2024) excels on Chinese entities and social norms, however, it lags behind GPT-4o on Chinese cultural commonsense. In general, LLMs consistently perform better when queried about culture-specific phenomena in the native language (i.e., Arabic, Chinese, Japanese) rather than English, with the gap narrowing for cultural commonsense. A likely explanation is that everyday cultural commonsense is often unspoken in native contexts while more explicitly expressed and asked about in foreign languages (Yin et al., 2022). Our analysis with search engines further supports this, highlighting the importance of broad cultural and linguistic coverage in training data.

## 2 Related Work

**Cultural Evaluation.** The widespread use of LMs has sparked research interest in their relevance to diverse cultures (Adilazuarda et al., 2024; Shen et al., 2024; Pawar et al., 2024; Liu et al., 2024). Several studies investigate LMs' alignment to different cultures by examining their responses to social surveys that reflect human values and attitudes

[1]By "Human Preferences", we mean pairwise or listwise rankings of multiple candidate answers to the same question, which are the signal needed for preference tuning.

(Haerpfer et al., 2021; Cao et al., 2023). It has been consistently shown that LMs favor answers associated with Western culture (AlKhamissi et al., 2024; Abdulhai et al., 2023), even when prompted in different languages (Masoud et al., 2023; Wang et al., 2023; Rystrøm et al., 2025) or after preference optimization (Ryan et al., 2024). Another line of work develops culture-specific evaluation resources such as knowledge bases of cultural facts (Keleg and Magdy, 2023; Yin et al., 2022; Zhou et al., 2024b), entity-centric cultural benchmarks (Naous et al., 2024), and user self-reported cultural experiences (Shi et al., 2024). Other works have constructed QA datasets for different cultural aspects, such as culinary customs (Palta and Rudinger, 2023), norms (Rao et al., 2024; Zhan et al., 2024), social etiquette (Chiu et al., 2024a,b; Qiu et al., 2025), and more (Arora et al., 2024; Mousi et al., 2024). However, they are primarily designed for evaluation, often in a multiple-choice QA format, and are not well-suited for aligning LMs through preference optimization, which ideally relies on free-text QA and human preference data. (see Table 1 and §3.1).

**Cultural Fine-tuning.** Several resources (Muennighoff et al., 2022; Singh et al., 2024; Alyafeai et al., 2024; Lian et al., 2023; Bartolome et al., 2023; Ahmadian et al., 2024) have been developed to improve LMs' multilingual performance. While they include culturally relevant samples, they primarily offer general instructions and preferences for safety and helpfulness. A few other works have investigated cultural adaptation of LMs via fine-tuning strategies (Li et al., 2024a,b; Shi et al., 2024; Kirk et al., 2024; Choenni et al., 2024; Yuan et al., 2024; Pham et al., 2025; Alwajih et al., 2025; Xu et al., 2025; Yao et al., 2025). Unlike past studies (see Table 1), CARE provides human preferences on culture-specific topics in three languages (not translated from English), enabling a direct study of how multilingual preference optimization improves models' cultural awareness and by how much.

**Multilingual Preference Optimization.** Existing work aims to improve the alignment of LMs with human preferences across different languages, either by synthesising preference data (Pan, 2024; Hwang, 2024; 2A2I, 2024) or developing translation-based methods (She et al., 2024; Yang et al., 2024b). Only a few studies have released native multilingual human preference datasets that are neither translated from English nor rated by AI models. Notable examples include OpenAssistant (Köpf et al., 2023) and HelpSteer3-Preference (Wang et al., 2025b), both of which focus on general-purpose interactions. We focus on culturally-relevant preferences for an in-depth analysis of multilingual multicultural alignment.

## 3 Constructing CARE

We introduce CARE, a multilingual human preference dataset that consists of 3,490 culture-specific questions and 31.7k human/LLM responses *with* human preference ratings.

### 3.1 Limitations of Existing Data Resources

In this section, we notice limitations of existing resources and take action to fix them. We start with the Aya dataset (Singh et al., 2024), the largest multilingual instruction-tuning resource that contains human-written questions and answers. Though its samples are collected from native speakers, only part of its content focuses on culture, as many examples consist of general questions. After manual inspection and filtering out generic questions (e.g. "*How many hearts does an octopus have?*"), we yield 1,324 (out of 4,909) and 600 (out of 3,264) question and answer pairs that are culturally relevant in Chinese and Japanese but only 457 (out of 14,210) samples in Arabic. To expand the Arabic set, we apply the same filtering process to around 2,000 samples from CIDAR (Alyafeai et al., 2024), a human-written Arabic instruction dataset, resulting in 162 relevant samples.

We also examine four existing cultural knowledge bases, namely, CultureBank (Shi et al., 2024), CulturalBench (Chiu et al., 2024b), GeoMLAMA (Yin et al., 2022), and FORK (Palta and Rudinger, 2023). Since these datasets are exclusively in English, we manually translate them into the corresponding native language. We observe that these knowledge bases focus on broad regional coverage,

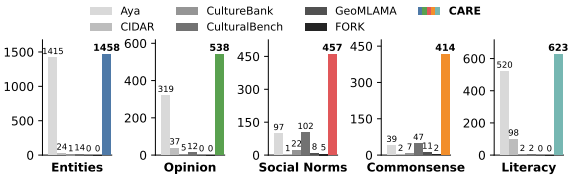



Figure 2: Overall coverage per cultural category. CARE provides 16.6× more social norm and commonsense questions compared to 2 instruction tuning datasets (Aya and CIDAR) and 4 cultural knowledge bases (CultureBank, CultureBench, GeoMLAMA, FORK) combined.

but supply limited samples per culture: namely, 60, 61, and 50 samples for Chinese, Arab, and Japanese cultures. Since these resources rely on multiple-choice or text-infilling formats, we reformatted all answers into free-text responses, adding explanations and removing any instances of stereotyping or overgeneralization.

We manually classify all questions into one of five categories: **(1) Cultural Entities**, where the question asks about culture-specific entities (Naous and Xu, 2025), **(2) Cultural Opinion**, where the question asks about a subjective interpretation for a cultural entity, **(3) Social Norms**, where the question is about social interactions between individuals (Huang and Yang, 2023), **(4) Cultural Commonsense**, where the question is about daily phenomena that locals may take for granted (Shen et al., 2024), and **(5) Literacy**, where the question is about the language, proverbs, or slang (Wuraola et al., 2024). After filtering all resources, we found an imbalanced coverage across categories, with a notable lack of data on social norms and commonsense, as shown in Figure 2. Also, to align LMs via preference optimization, we require human judgments of both appropriate and inappropriate responses, which are not offered by existing resources.

### 3.2 Multilingual Cultural Preference Data

To address the aforementioned shortcomings, we collect new data in three languages as follows:

**Social Norm and Commonsense.** To expand the samples on social norms and cultural commonsense, we ask native Chinese, Arabic, and Japanese speakers who are international college students undergoing culture transfer experiences themselves to curate such samples. To help brainstorm, the speakers are instructed to leverage international as well as regional social media platforms or forums (e.g., Twitter, Reddit, Zhihu, RedNote), and search for posts where users describe their culture shift experiences using search keywords such as "*most surprised abroad*", "*culture shock*", "*first time to*", etc. Taking inspiration from such discussions, the speakers create 190, 196, and 260 samples in Chinese, Arabic, and Japanese. This human-curated data is more authentic than synthetic data, preventing some inaccuracy and overgeneralization (e.g., *"In China, it is common to drink soup after the main dish."*) we have observed in the existing datasets.

**Culture-specific Human Preference Judgments.** Given these culture-specific questions, gathering human preferences on the cultural relevance of responses is crucial to performing cultural alignment through preference optimization. To obtain these cultural preferences, we present the native annotators with the zero-shot responses of 9 different LMs to each question within CARE. We specifically use the instruct version of recent multilingual LMs of `Llama3.3-70B` (Dubey et al., 2024), `Qwen2.5-72B` (Yang et al., 2024a), `Gemma2-27B` (Team et al., 2024), `Mistral-Large`, and `GPT-4o`. We also use their smaller-sized `Llama3.1-8B`, `Qwen2.5-7B`, `Gemma2-9B`, and `Mistral-7B`.

| Culture | $\alpha$ | $r$ | $\rho$ | $\tau$ |
|---|---|---|---|---|
| Arab | 0.86 | 0.90 | 0.88 | 0.71 |
| Chinese | 0.84 | 0.89 | 0.89 | 0.75 |
| Japanese | 0.92 | 0.93 | 0.92 | 0.77 |

Table 2: Inter-annotator agreement based on Krippendorff's $\alpha$, Pearson's $r$, Spearman's $\rho$, and Kendall's $\tau$.

We instruct annotators to rank the generated responses from the most to least culturally appropriate and assign a rating on a 1–10 scale (1: *poor* $\rightarrow$ 10: *excellent*). These ratings reflect how well responses align with the cultural expectations of native speakers and are used to construct preference pairs for cultural preference learning (§4.1).

**Human Annotators and Agreement.** To curate the samples in CARE, we recruited 2 Chinese, 2 Arab, and 5 Japanese native speakers who are familiar with the respective cultures. These annotators manually filtered the aforementioned resources and searched online platforms for additional samples, capturing broader community knowledge and ensuring that CARE reflects cultural experiences beyond our annotator pool. The annotators then rated the responses to the questions in CARE.

To ensure the quality of our preference ratings, we additionally recruited 4 Chinese, 1 Arab, and 4 Japanese annotators, who were not involved in the data curation process, to perform double annotation. Each additional annotator was given 75 randomly sampled questions (15 per cultural category), for which they were asked to provide their ratings of model responses. We then compare the ratings of the additional annotators to our initial set of annotators. As shown in Table 2, we find substantial inter-annotator agreements across various metrics, indicating consistent response preferences for questions within our considered cultural categories (see Appendix A.2 for per-category agreements).

Besides the 5-class cultural categories, we also

| Model | | Chinese | | | | | | Arabic | | | | | | Japanese | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* |
| Gemma2-9B | Vanilla | 4.89 | 8.46 | 7.55 | 6.67 | 4.72 | 6.49 | 4.73 | 6.56 | 6.33 | 5.60 | 3.28 | 5.33 | **4.46** | 6.50 | 5.67 | 6.50 | 2.33 | 5.09 |
| | Aligned | **5.50** | **8.53** | **7.90** | **7.26** | **5.17** | **6.89** | **5.30** | **6.82** | **6.93** | **6.55** | **3.57** | **5.84** | 4.40 | **6.60** | **5.80** | **6.67** | **2.93** | **5.28** |
| Llama3.1-8B | Vanilla | 3.14 | 4.16 | 4.62 | 3.93 | 3.03 | 3.78 | 4.08 | 3.62 | 2.87 | 3.07 | 2.07 | 3.30 | 2.20 | 3.93 | 2.47 | 2.57 | **1.97** | 2.63 |
| | Aligned | **3.86** | **5.90** | **5.36** | **5.56** | **3.69** | **4.88** | **4.33** | **4.43** | **3.70** | **4.50** | **2.36** | **3.86** | **3.20** | **4.67** | **2.80** | **3.20** | 1.67 | **3.11** |
| Qwen2.5-7B | Vanilla | 6.89 | 7.86 | 7.48 | 6.80 | 7.37 | 7.28 | **4.65** | 5.84 | 5.44 | 4.88 | 2.84 | 4.61 | 2.13 | 5.67 | 4.13 | 4.60 | **2.37** | 3.78 |
| | Aligned | **7.20** | **8.76** | **7.66** | **6.90** | **7.53** | **7.61** | 4.55 | **6.40** | **5.55** | **5.33** | **3.35** | **5.06** | **2.77** | **5.73** | **4.30** | **5.17** | 1.90 | **3.97** |
| Mistral-7B | Vanilla | **3.03** | 3.83 | 3.93 | 4.38 | **2.43** | 3.53 | 2.56 | 2.46 | 2.03 | **2.13** | 1.34 | 2.11 | 1.70 | **3.96** | 2.40 | **2.48** | 1.20 | 2.34 |
| | Aligned | 2.43 | **3.90** | **4.53** | **5.00** | 2.20 | **3.61** | **2.60** | **3.36** | **2.46** | 2.10 | **1.40** | **2.38** | **1.83** | 3.90 | **2.43** | 2.33 | **1.23** | **2.35** |

Table 3: Average scores (1: *poor* → 10: *excellent*) in responding to questions related to Chinese culture in Chinese, Arab culture in Arabic, and Japanese culture in Japanese. Performances are presented for vanilla LMs and LMs after cultural alignment using DPO on CARE. For each LM, the row labeled "*Vanilla*" corresponds to the original model (gray plot), and "*Aligned*" is after cultural preference learning (colored plot).

label each sample with the **Associated Culture** for experiments in §4.5: *Native* (questions about the local culture; i.e., Chinese, Arab, or Japanese), *Foreign* (questions about other cultures; e.g., US, German, etc.), or *General* (questions that are not specific to a particular culture). For fine-grained regional cultural knowledge evaluation (Appendix D.2), we also annotate samples for their **Geographic Scope** (*sub-nationwide*, *nationwide*, *continent-wide*, or *worldwide*). See Appendix A.1 for the annotation guideline.

**Data Split.** For the experiments in §4 and §5, we construct culture-specific test sets of 150 questions each for Chinese, Arab, and Japanese cultures, sampling 30 questions per cultural category. Remaining CARE questions (1,513 Chinese, 644 Arabic, 757 Japanese), along with rated responses, form the training set. This training data also includes questions about other cultures (e.g. U.S., German, etc.) beyond Chinese, Arab, and Japanese.

## 4 Aligning LMs for Cultural Awareness

We investigate whether, and to what extent, high-quality culture-specific data can enhance LMs. We perform human preference learning for cultural alignment of medium-sized LMs (7∼9B parameters) using CARE that fit in 8 Nvidia A40 GPUs.

### 4.1 Experiment Setup

**Human Preference Optimization.** The goal of cultural preference optimization is to align model outputs with culturally appropriate responses, informed by human judgments. Each training instance $\mathcal{D}$ consists of a prompt $x$ and a pair of responses $(y_w, y_\ell)$, where $y_w$ is the preferred and $y_\ell$ is the dispreferred. In CARE, we construct these pairs by taking the highest- and lowest-scored responses from the human annotations. We evaluate the following representative algorithms:

**DPO** (Rafailov et al., 2024) directly optimizes the log-odds difference between the preferred and dispreferred responses, measured against a reference model:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \pi_\theta(y_w \mid x) - \log \pi_{\text{ref}}(y_w \mid x)\big] - \beta\big[\log \pi_\theta(y_\ell \mid x) - \log \pi_{\text{ref}}(y_\ell \mid x)\big]\Big),$$

where $\sigma$ is the sigmoid and $\beta$ is a scaling factor.

**KTO** (Ethayarajh et al., 2024) introduces a reference baseline:

$$z_{\text{ref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\beta\, D_{\text{KL}}\big(\pi_\theta(y \mid x) \,\|\, \pi_{\text{ref}}(y \mid x)\big)\big]$$

and weighting desirable and undesirable examples with $\lambda_w, \lambda_\ell > 0$:

$$\mathcal{L}_{\text{KTO}} = -\lambda_w\, \sigma\Big(\beta\big[\log \pi_\theta(y_w \mid x) - \log \pi_{\text{ref}}(y_w \mid x)\big] - z_{\text{ref}}\Big) + \lambda_\ell\, \sigma\Big(z_{\text{ref}} - \beta\big[\log \pi_\theta(y_\ell \mid x) - \log \pi_{\text{ref}}(y_\ell \mid x)\big]\Big).$$

**SimPO** (Meng et al., 2024) removes the dependence on a reference model by directly comparing the length-normalized log-likelihoods of the two responses, with a margin $\gamma$:

$$\mathcal{L}_{\text{SimPO}} = -\log \sigma\Big(\tfrac{\beta}{|y_w|}\log \pi_\theta(y_w \mid x) - \tfrac{\beta}{|y_\ell|}\log \pi_\theta(y_\ell \mid x) - \gamma\Big).$$

Across these methods, we find that all achieve comparable performance on CARE (Table 4). Subsequently, we present most of the experiments in this paper with DPO unless otherwise specified.

**Baselines.** We compare cultural preference tuning on CARE with four groups of baselines: (1) General-domain preference datasets: we adopt the multilingual translations (2A2I, 2024; Pan, 2024; Hwang, 2024) of two synthetic English preference datasets, **OpenOrca** (Intel, 2024) (2.1k Ar, 2.0k Zh, 12.9k Ja pairs) and **UltraFeedback** (Cui et al., 2023a) (2.0k Ar, 1.7k Zh pairs). We also compare

| Approach (w/ data) | Gemma2-9B | | | Qwen2.5-7B | | | Llama3.1-8B | | | Mistral-7B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Arabic* | *Chinese* | *Japanese* | *Arabic* | *Chinese* | *Japanese* | *Arabic* | *Chinese* | *Japanese* | *Arabic* | *Chinese* | *Japanese* |
| *0-shot Prompting* | | | | | | | | | | | | |
| Vanilla | 5.331 | 6.490 | 5.093 | 4.618 | 7.286 | 3.780 | 3.304 | 3.784 | 2.627 | 2.114 | 3.534 | 2.339 |
| SFT (w/ Alpaca) | 5.443 | 6.416 | 3.447 | 4.689 | 5.093 | 3.387 | 3.141 | 3.709 | 2.433 | 1.287 | 2.100 | 1.673 |
| SFT (w/ CARE) | 5.463 | 6.440 | 3.493 | 4.700 | 5.396 | 3.219 | 3.440 | 3.813 | 2.673 | 1.360 | 2.627 | 1.653 |
| DPO (w/ UltraFeedback) | 5.765 | 6.380 | – | 4.845 | 7.547 | – | 3.880 | 4.160 | – | 2.220 | 3.307 | – |
| DPO (w/ OpenOrca) | 5.564 | 6.260 | 5.060 | 4.878 | 7.433 | 3.653 | 3.456 | 3.260 | 2.547 | 2.067 | 3.480 | 1.747 |
| DPO (w/ HelpSteer3) | – | 6.133 | 4.973 | – | 7.000 | 3.800 | – | 3.280 | 2.673 | – | **3.687** | 2.215 |
| DPO (w/ CARE) | 5.848 | **6.899** | 5.280 | **5.062** | 7.613 | 3.980 | 3.867 | 4.886 | 3.107 | 2.387 | 3.613 | 2.349 |
| KTO (w/ CARE) | **6.387** | 6.713 | **5.473** | 4.822 | 7.617 | **4.147** | 3.911 | 4.691 | 3.013 | **2.687** | 3.513 | **2.473** |
| SimPO (w/ CARE) | 5.932 | 6.647 | 5.033 | 4.765 | 7.427 | 3.847 | 3.917 | **4.946** | 2.947 | 2.253 | 3.480 | 2.093 |
| MAPO (w/ CARE) | 5.758 | 6.340 | 5.153 | 4.820 | **7.640** | 3.613 | **4.107** | 4.753 | **3.287** | 2.327 | 3.580 | 2.133 |
| *CoT Prompting* | | | | | | | | | | | | |
| Vanilla | 5.946 | 6.081 | 4.613 | 4.703 | 7.667 | 3.873 | 3.107 | 3.887 | **2.927** | 2.333 | **4.373** | **2.273** |
| DPO (w/ CARE) | **6.096** | **6.407** | **5.093** | **4.946** | **7.703** | **4.220** | **3.678** | **5.087** | 2.840 | **2.427** | 4.233 | 2.173 |
| *Role-Play Prompting* | | | | | | | | | | | | |
| Vanilla | 4.073 | 6.396 | 5.207 | 4.899 | **7.939** | 4.100 | 3.500 | 4.087 | 2.547 | **2.513** | 3.530 | **2.320** |
| DPO (w/ CARE) | **5.938** | **6.561** | **5.527** | **5.129** | 7.878 | **4.313** | **3.899** | **5.093** | **2.953** | 2.362 | **3.720** | 2.167 |

Table 4: Average scores (1: *poor* → 10: *excellent*) on Chinese, Arab, and Japanese cultures for a variety of prompting approaches, supervised fine-tuning, and preference learning using culture-specific (CARE) vs. general instruction-tuning (multilingual Alpaca) and preference (translated and filtered OpenOrca/UltraFeedback) data. SFT is performed on the instruction data only, while preference learning is conducted on the preference pairs.

with **HelpSteer3** (Wang et al., 2025b), a human-annotated multilingual preference dataset (1.1k Zh, 0.2k Ja pairs). All focus on truthfulness, honesty, and helpfulness. (2) Instruction-tuned models: we conduct SFT on our CARE and the multilingual **Alpaca** corpus (Cui et al., 2023b; Chen et al., 2023; fujiki, 2023). (3) Multilingual preference tuning: **MAPO** (She et al., 2024). (4) Prompting methods: **CoT** (Wei et al., 2022) and **role-play** (Kong et al., 2023). Implementation details are in Appendix B.

**LM-as-a-judge Evaluation and its Reliability.** For evaluation, we adopt the LM-as-a-judge strategy (Zheng et al., 2023), prompting `GPT-4o` to generate a rationale and assign a 1-10 score on how aligned the response is to the human reference. We validate this setup by comparing `GPT-4o`'s ratings with native annotators', where we achieve a high Pearson correlation of 0.93. More details and full evaluation prompts can be found in Appendix C.

## 4.2 Main Results

Table 3 presents the results of medium-sized LMs before and after cultural preference learning with CARE, while Table 4 compares results with various baselines. We see the following key findings:

Using CARE, culturally-aligned LMs achieve higher average scores (up to 29% improvement) compared to the vanilla checkpoints across all three cultures. This resonates with the findings of Zhou et al. (2024a) and shows that a relatively small amount of carefully curated data by humans can improve LMs' alignment. We also see a noticeable gap among LMs developed by different regions. The `Qwen2.5-7B` developed by China-based Alibaba performs the best in Chinese and can be further improved from 7.28 to 7.61 (out of 10) by conducting preference optimization with CARE. Interestingly, ***the aligned version of*** **`Gemma2-9B`** ***outperforms*** **`Qwen2.5-7B`** ***on Chinese social norms and commonsense*** (more on this in §5.2).

***Cultural preference learning strengthens stronger LMs but fails to address weaknesses in weaker LMs.*** `Qwen2.5-7B` and `Gemma2-9B`, the top-performing LMs on Chinese and Arabic, respectively, show consistent improvement across cultural categories after preference learning. However, preference learning shows limited gains where the model's initial performance is poor. `Mistral-7B` is not improved on Chinese entities and literacy, where its starting scores are only 3.03 and 2.43. Similarly, `Qwen2.5-7B`, does not benefit from preference learning on Arabic entities. Models also show limited progress on Japanese data, reflecting their shallow exposure to Japanese culture. Similarly, when the model possesses foundational cultural knowledge, utilizing the preference signals from the model itself via methods like MAPO rather than directly from human preferences can also lead to improvements. All these suggest that base models need basic cultural knowledge for preference learning to be effective.

***Culture-specific human preference data provides complementary benefits.*** Table 4 shows that preference tuning with CARE yields competitive

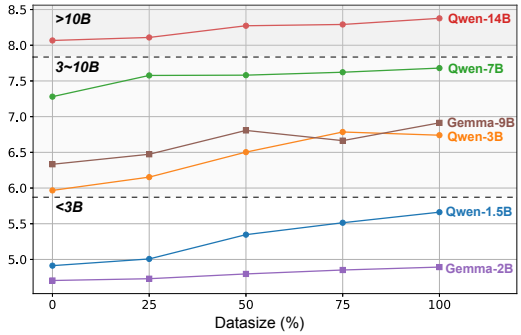



Figure 3: Impact of model size and preference data volume on cultural awareness performance. The average scores (1: *poor* → 10: *excellent*) of aligned models are plotted against different % of preference pairs in CARE. Improvements are achieved across different model sizes and data sizes, in comparison to the vanilla model (0%).

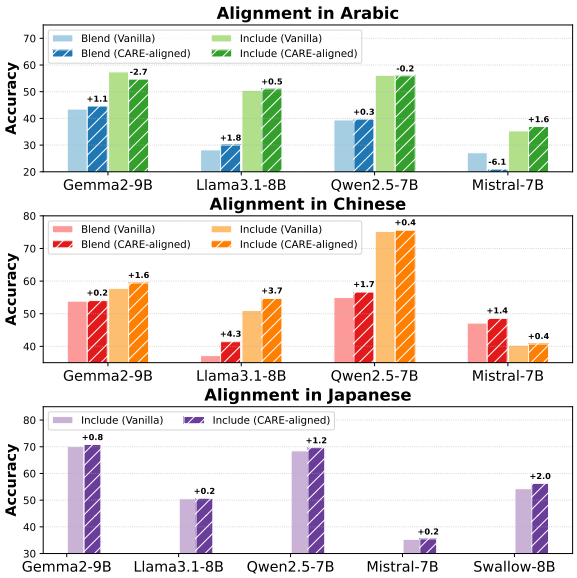



Figure 4: Accuracy on the Chinese, Arabic, and Japanese subsets of the short answer questions in Blend (Myung et al., 2024) and multi-choice questions in Include (Romanou et al., 2025). CARE preference tuning improves out-of-domain cultural tasks for most cultures and models, except on the Arabic subset of Include. Blend does not provide a Japanese subset.

or stronger results than training on larger general datasets (3–17× more pairs). Cultural preference learning with CARE also improves performance in both CoT and role-play prompting setups. While culture-specific SFT outperforms generic SFT, preference tuning yields greater gains, highlighting the value of ***human cultural preferences over ground-truth demonstrations alone***.

### 4.3 Scaling of Data and Model Sizes

We examine how scaling both model size and data volume influences the performance before and after cultural preference learning. Specifically, we consider the `Qwen2.5` `{1.5B, 3B, 7B, 14B}` and the `Gemma2` `{2B, 9B}` series, which offer multiple model sizes. For data sizes, we align models using DPO on different proportions `{0%, 25%, 50%, 75%, 100%}` of Chinese cultural preference pairs within CARE. As shown in Figure 3, consistent improvements are observed across different model and data sizes. Even a relatively small amount (e.g., 25%) of data in CARE can lead to improvements, particularly for smaller-sized LMs. Scaling up the data leads to progressively better performance, highlighting ***the benefits of employing more cultural preference data in future work***.

### 4.4 Cultural Generalization

We evaluate CARE-aligned models on other cultural tasks using culture-related data that is not included in CARE. Specifically, we use the short answer questions within the Blend (Myung et al., 2024) benchmark and multi-choice questions within the Include (Romanou et al., 2025) benchmark, both written in the native languages. From Figure 4, we observe that ***preference learning with culture-specific human preference can also benefit out-of-domain cultural tasks***, indicating the generalization capability of CARE. Appendix D.4 extends this analysis to additional fine-tuned models, and Appendix D.3 demonstrates that overall NLP capabilities remain unaffected.

### 4.5 Multi-cultural Alignment

We analyze the impact of incorporating data about different cultures in preference learning. To ensure a fair comparison, we select an equal number of samples for each trial. Specifically, we align `Llama3.1-8B-Instruct` with 675 Chinese and 547 Arabic pairs sampled from three different contexts: native (Chinese or Arab) culture, foreign cultures (non-Chinese or non-Arab), and mixed pairs (half native and half foreign). Results are shown in Figure 5. Tuning on foreign-culture pairs alone lowers native performance; adding native pairs raises them (Chinese 4.69 → 4.87, Arab 2.98 → 3.50), and combining native and foreign pairs further improves performance, indicating that ***geographically diverse preference data can further strengthen cultural awareness***.

### 4.6 Language-specific LLMs

We also explore the combined benefits of both language-specific training and cultural preference tuning. In particular, we conduct DPO tun-

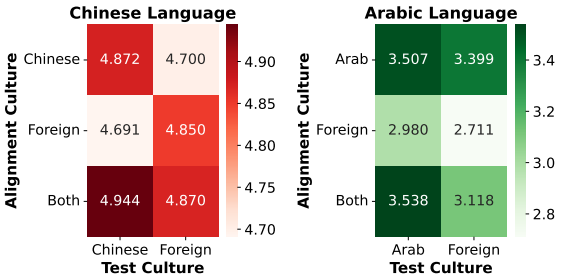



Figure 5: Average scores of `Llama3.1-8B-Instruct` on local and foreign cultures when aligned using data from native, foreign, or mixed cultures. The highest performance on local culture is achieved when mixing local and foreign samples during preference learning.

| Model | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | Avg. |
|---|---|---|---|---|---|---|
| | | | *Llama family* | | | |
| Llama3.1-8B | 2.20 | 3.93 | 2.47 | 2.57 | 1.97 | 2.63 |
| ↪ CARE Aligned | 3.20 | 4.67 | 2.80 | 3.20 | 1.67 | 3.11 |
| ↪ Swallow-8B | 5.63 | 6.40 | 5.47 | 6.33 | 4.63 | 5.69 |
| ↪ CARE Aligned | **5.77** | **6.57** | **5.60** | **6.70** | **5.40** | **6.01** |
| Llama3.3-70B | 6.17 | 6.07 | 4.77 | 6.13 | 5.97 | 5.82 |
| ↪ Swallow-70B | **7.20** | **7.37** | **7.13** | **7.53** | **7.27** | **7.30** |
| | | | *Gemma family* | | | |
| Gemma2-9B | 4.46 | 6.50 | 5.67 | 6.50 | 2.33 | 5.09 |
| ↪ CARE Aligned | 4.40 | 6.60 | 5.80 | 6.67 | 2.93 | 5.28 |
| ↪ Swallow-9B | 6.07 | **7.27** | 6.90 | **7.20** | 5.10 | 6.51 |
| ↪ CARE Aligned | **6.10** | 7.13 | **7.37** | 6.97 | **5.50** | **6.61** |
| Gemma2-27B | 5.37 | 7.07 | **6.50** | 6.47 | 3.83 | 5.85 |
| ↪ Swallow-27B | **6.37** | **7.43** | 6.43 | **7.43** | **6.00** | **6.73** |
| | | | *Mistral family* | | | |
| Mistral-7B | 1.70 | 3.96 | 2.40 | 2.48 | 1.20 | 2.34 |
| ↪ CARE Aligned | 1.83 | 3.90 | 2.43 | 2.33 | 1.23 | 2.35 |
| ↪ Swallow-7B | 2.70 | **4.43** | 4.07 | 3.87 | 2.00 | 3.41 |
| ↪ CARE Aligned | **3.13** | 4.00 | **4.50** | **4.90** | **2.53** | **3.81** |
| Mistral-8×7B | 2.10 | 4.37 | 3.23 | 3.30 | 2.30 | 3.06 |
| ↪ Swallow-8×7B | **4.33** | **5.40** | **5.30** | **5.87** | **2.90** | **4.76** |

Table 5: Average scores on CARE's Japanese test set for (i) each backbone, (ii) each backbone further tuned on CARE, (iii) its Japanese-pretrained `Swallow` variant, and (iv) that variant further tuned on CARE.

ing on CARE's Japanese training split with the Swallow model variants (`Llama3.1-Swallow-8B`, `Gemma2-Swallow-9B`, and `Mistral-Swallow-7B`), which are continuously pre-trained and instruction-tuned on the Japanese corpus (Fujii et al., 2024). Table 5 shows that language-specific continual pre-training improves each base model by 1–3 points (on a 10-point scale), while subsequent preference tuning on CARE contributing an additional 0.2–0.4 points. Combined, these steps enable smaller models to outperform much larger LLMs. For example, aligned `Llama-Swallow-8B` outperforms vanilla `Llama-70B` (6.01 vs. 5.82), and aligned `Gemma-Swallow-9B` surpasses `Gemma-27B` (6.61 vs. 5.85). The results echo our findings in §4.2: cultural alignment is consistently helpful, even when base models have enough language proficiency.

# 5 Prompting LMs for Cultural Awareness

We evaluate larger LMs and compare to human familiarity and web-based sources (§5.1), as well as querying in different languages (§5.2).

## 5.1 Main Results

We follow the evaluation setup in §4.1 and assess the larger (> 25B parameters) versions of the evaluated LMs with zero-shot prompting. Table 6 shows the results. Larger models are better for all three cultures, with closed-source ones like `GPT-4o` and `Deepseek-v3` leading the overall performance. The more Chinese-centric model `Deepseek-v3` and `Qwen2.5-72B` excels in the knowledge about cultural entities (8.96 and 8.42), social norms (9.03 and 9.06), and literacy (9.03 and 8.53) for Chinese; meanwhile, `GPT-4o` and `Mistral-Large` are comparably good if not better at offering opinions (9.43 and 8.96) and answering commonsense questions (9.00 and 8.58) about Chinese culture. This is interesting and likely because ***cultural commonsense knowledge is often unstated in the native language, while more balanced and thoughtful opinions may be expressed in foreign languages***. We further test this hypothesis in §5.2.

**Human Awareness.** To assess human familiarity with the culture-specific questions, we ask native-speaking annotators whether they knew the correct answers without looking them up when presented with questions from CARE. Table 6 shows the percentages of questions humans could answer immediately with confidence, which are very high for social norms and commonsense knowledge.

**Search Engine.** To further examine the coverage of cultural knowledge in real-world text, for each CARE question we used the Google Programmable Search Engine[2] to retrieve the top-10 results (URLs and associated snippets), fetched each HTML page, extracted the paragraph containing the snippet, scored all paragraphs with the same LM-as-a-judge framework, and recorded the maximum as the question's score. Results are shown in Table 6. Since many retrieved paragraphs only partially address the questions, their scores are generally below LMs' scores.

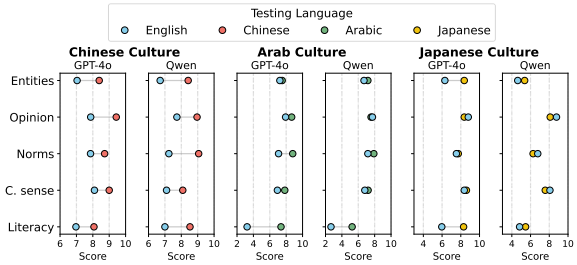
## 5.2 Cross-lingual Analysis

We translate all questions and answers in CARE that are written in Chinese, Arabic, and Japanese

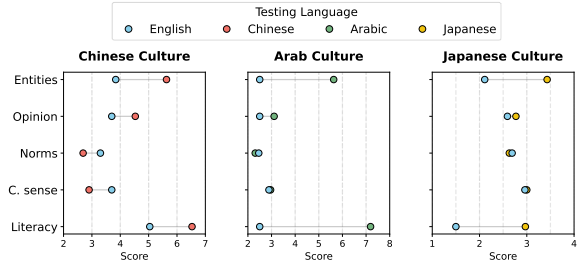
[2] `https://programmablesearchengine.google.com`

| Model | Chinese | | | | | | Arabic | | | | | | Japanese | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* |
| Human Awareness (%) | 50% | 53% | 96% | 100% | 56% | 71% | 70% | 90% | 100% | 90% | 20% | 74% | 17% | 87% | 87% | 93% | 23% | 63% |
| Search Engine | 5.63 | 4.53 | 2.69 | 2.90 | 6.53 | 4.46 | 5.87 | 3.11 | 2.31 | 2.96 | 7.19 | 4.30 | 3.43 | 2.77 | 2.63 | 3.00 | 2.97 | 2.96 |
| Gemma2-27B | 5.64 | 8.33 | 8.13 | 6.96 | 5.73 | 6.97 | 6.24 | 7.76 | 7.76 | 6.53 | 3.50 | 6.44 | 5.37 | 7.07 | 6.50 | 6.47 | 3.83 | 5.85 |
| Llama3.3-70B | 5.82 | 7.33 | 7.27 | 7.06 | 5.93 | 6.69 | 5.79 | 6.93 | 7.13 | 5.86 | 3.86 | 5.93 | 6.17 | 6.07 | 4.77 | 6.13 | 5.97 | 5.82 |
| Qwen2.5-72B | 8.42 | 8.96 | **9.06** | 8.09 | 8.53 | 8.61 | 7.17 | 7.56 | 7.90 | 7.20 | 5.25 | 7.04 | 5.37 | 8.10 | 6.27 | 7.57 | 5.47 | 6.55 |
| Mistral-Large | 7.57 | 8.96 | 8.24 | 8.58 | 6.66 | 8.01 | 6.34 | 7.63 | 7.70 | 7.13 | 4.80 | 6.72 | 4.83 | 7.63 | 7.10 | 7.17 | 5.23 | 6.39 |
| Deepseek-v3 | **8.96** | 9.30 | 9.03 | **9.23** | **9.30** | **9.16** | **7.90** | 8.16 | 8.70 | **8.00** | **7.39** | 8.04 | 7.67 | **8.87** | **8.90** | **9.07** | 7.77 | **8.45** |
| GPT-4o | 8.39 | **9.43** | 8.72 | 9.00 | 8.06 | 8.73 | 7.51 | **8.66** | **8.80** | 7.83 | 7.37 | **8.05** | **8.37** | 8.37 | 7.73 | 8.60 | **8.30** | 8.27 |

Table 6: Average scores on CARE of larger LMs and web content, evaluated by the judge LM across all samples (1: *poor* → 10: *excellent*). "Human Awareness (%)" indicates the percentage of questions for which locals know the correct answer. Chinese-developed `Deepseek-v3` and `Qwen2.5` excel on Chinese entities, norms, and literacy, while `GPT-4o` and `Mistral-Large` are comparably good at opinions and commonsense. For social norms and cultural commonsense, where local humans show high confidence, Google search often returns fewer relevant answers.




(a) Average scores achieved by `Qwen2.5-72B-Instruct` and `GPT-4o` when prompted in native languages versus English.




(b) Average scores of retrieved webpage main content when searching CARE questions in native languages versus English.

Figure 6: Performance comparison of cultural awareness in different languages. Comparison between (a) LM performance and (b) retrieved webpage content, showing that native-language questions capture cultural nuances for literacy, opinion, and entities. However, the gaps narrow or sometimes reverse for cultural norms and commonsense.

into English to enable cross-lingual evaluation. Figure 6a shows that LMs generally perform better on culture-specific questions when prompted in the native languages than in English, but more notably for Chinese than Arabic and Japanese. This advantage is also especially obvious for the literacy category, while the gap narrows in categories such as cultural commonsense.

We also examine how language choice impacts search engine results. As shown in Figure 6b, cultural information on entities, opinions, and literacy tends to be of higher quality when retrieved in native languages. Interestingly, this trend reverses for questions related to cultural commonsense and social norms, such as *"In China, does the ticket time indicate when the feature movie starts, or are there trailers played before the main feature?"*, where English searches yield higher-quality content. This aligns with the potential documentation gap mentioned in §5.1, where native speakers often assume everyday cultural knowledge is implicit, but non-native speakers are more likely to seek out information or document it.

## 6 Conclusion

We introduce CARE, a multilingual, human-written resource comprising 3,490 culture-specific questions about Chinese, Arab, and Japanese cultures, along with human-rated responses. Through extensive experiments, we investigate that a small amount of high-quality cultural preferences can improve LMs via preference optimization, across various model families and scales. Our analyses reveal gaps in cultural awareness among LMs, native speakers, and search engines, especially regarding implicit cultural commonsense. By releasing CARE, we hope to foster the development of more culturally adaptive LMs.

## Limitations

During data collection, we intentionally excluded highly sensitive and controversial topics (e.g., "*Are security concerns about the niqab exaggerated or*

*justified?*"), as these tend to elicit divergent personal views. Our work addresses generally shared topics within each cultural group (e.g., "*Do Chinese people like stinky tofu?*" in the Opinion category of CARE), allowing us to capture collective preferences that reflect broader cultural consensus (Bakker et al., 2022) with a neutral point of view (e.g., "*Some Chinese like stinky tofu. It is a traditional food made by . . . It is most popular in . . . regions. At the same time, many Chinese people don't like it . . .*" ). This approach facilitates more objective and replicable evaluation, as evidenced by the high inter-annotator agreement reported in §3.2, thus supports more reliable algorithmic comparisons, which are the primary focus of this paper. The scope of our study is complementary to other work (Chiu et al., 2024a; Kirk et al., 2024; Huang et al., 2025) that explored diverse values and subjective viewpoints (Zhong et al., 2021).

We also note that LMs may occasionally respond to culture-specific questions in a stereotypical or biased manner, stemming from a lack of cultural understanding. For instance, when asked "*In China, what does the term "laowai" mean when referring to foreigners?*", LMs often incorrectly interpret the term as disrespectful. In reality, "*laowai* (老外)" is typically a neutral descriptor in Chinese, used to denote foreigners without negative connotations. While CARE includes examples that clarify such misunderstandings through human-written responses, future studies in cultural red-teaming can further investigate and address these failure cases.

## Ethics Statement

In this study, we employed in-house annotators to collect cultural samples and provide preference judgments in CARE. The annotators for the Chinese and Arabic samples were university-level students fluent in the respective languages. Japanese annotators were Japanese workers with university degrees and cross-cultural experience abroad. Each Chinese and Arabic annotator was paid at $18 per hour, exceeding the U.S. federal minimum wage. We ensured that no personal data was collected from the annotators and emphasized that participation was voluntary. We also informed the annotators that the collected data would be used to enhance the cultural awareness of LMs. The culture-related questions and reference responses in CARE are either from the existing works or human-written sentences provided by our annotators.

## Acknowledgments

The authors would like to thank Chao Jiang, Jad Matthew Bardawil, Nour Allah El Senary, Ruohao Guo, Xiaofeng Wu, Yuming Pan, and Zirui Shao for their assistance with data annotation; and Yao Dou, Jonathan Zheng, Xiaofeng Wu as well as three anonymous reviewers for their helpful feedback for their helpful feedback. We would also like to thank Microsoft's Azure Accelerate Foundation Models Research Program and NVIDIA's Academic Grant Program for providing computational resources to support this work. This work was funded in part through a Sony Faculty Innovation Award, and by the NSF under grant number IIS-2144493, IIS-2052498 and SMA-2418946. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

2A2I. 2024. 2A2I/argilla-dpo-mix-7k-arabic.

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. arXiv preprint arXiv:2310.15337.

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

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. arXiv preprint arXiv:2406.18682.

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

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, et al. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. arXiv preprint arXiv:2503.00151.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. 2024. CIDAR: Culturally relevant instruction dataset for arabic. arXiv preprint arXiv:2402.03177.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. arXiv preprint arXiv:2406.17761.

Michiel A Bakker, Martin J Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In Advances in Neural Information Processing Systems.

Alvaro Bartolome, Gabriel Martin, and Daniel Vila. 2023. Notus. https://github.com/argilla-io/notus.

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

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024a. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs'(lack of) multicultural knowledge. arXiv preprint arXiv:2404.06664.

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

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023a. Ultrafeedback: Boosting language models with high-quality feedback. Preprint, arXiv:2310.01377.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In Proceedings of the First Conference on Language Modeling, COLM, page (to appear), University of Pennsylvania, USA.

fujiki. 2023. fujiki/japanese_alpaca_data. https://huggingface.co/datasets/fujiki/japanese_alpaca_data.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, E Ponarin, and B Puranen. 2021. World values survey: Round seven. JD Systems Institute & WVSA Secretariat. Data File Version, 2(0).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

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

Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. Values in the Wild: Discovering and analyzing values in real-world language model interactions. Preprint, arXiv:2504.15236.

Yongtae Hwang. 2024. yongtae-jp/orca_dpo_pairs_ja.

Intel. 2024. Intel/orca_dpo_pairs.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. arXiv preprint arXiv:2311.10702.

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

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. arXiv preprint arXiv:2306.05076.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. arXiv preprint arXiv:2404.16019.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. arXiv preprint arXiv:2308.07702.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations-democratizing large language model alignment. Advances in Neural Information Processing Systems, 36:47669–47681.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. ArabicMMLU: Assessing massive multitask language understanding in arabic. arXiv preprint arXiv:2402.12840.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In Proceedings of the 29th Symposium on Operating Systems Principles, pages 611–626.

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

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

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: Measuring massive multitask language understanding in chinese. arXiv preprint arXiv:2306.09212.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. OpenOrca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/Open-Orca/OpenOrca.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. arXiv preprint arXiv:2406.03930.

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

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. arXiv preprint arXiv:2405.14734.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov, Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink, and Zeerak Talat. 2025. SHADES: Towards a multilingual assessment of stereotypes 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). Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. arXiv preprint arXiv:2409.11404.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.

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

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

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pretraining data to linguistic phenomena. arXiv preprint arXiv:2501.04662.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In Proceedings of the ACM Web Conference 2023, pages 1907–1917.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2024. JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. arXiv preprint arXiv:2410.17250.

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

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9952–9962.

Wenbo Pan. 2024. wenbopan/Chinese-dpo-pairs.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. arXiv preprint arXiv:2411.00860.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2025. 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). Association for Computational Linguistics.

Zhaozhi Qian, Faroq Altam, Muhammad Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. arXiv preprint arXiv:2409.12623.

Haoyi Qiu, Kung-Hsiang Huang, Ruichen Zheng, Jiao Sun, and Nanyun Peng. 2025. Multimodal cultural safety: Evaluation frameworks and alignment strategies. arXiv preprint arXiv:2505.14972.

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

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

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In The Thirteenth International Conference on Learning Representations.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301.

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

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

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO:

Advancing multilingual reasoning through multilingual alignment-as-preference optimization. arXiv preprint arXiv:2401.06838.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. arXiv preprint arXiv:2405.04655.

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

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. arXiv preprint arXiv:2402.06619.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

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

Yifan Wang, Runjin Chen, Bolian Li, David Cho, Yihe Deng, Ruqi Zhang, Tianlong Chen, Zhangyang Wang, Ananth Grama, and Junyuan Hong. 2025a. More is less: The pitfalls of multi-model synthetic preference data in dpo safety alignment. arXiv preprint arXiv:2504.02193.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025b. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. Preprint, arXiv:2505.11475.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Andrea W Wen-Yi, Unso Eun Seo Jo, Lu Jia Lin, and David Mimno. 2024. How chinese are chinese language models? the puzzling lack of language policy in china's llms. arXiv preprint arXiv:2407.09652.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with llms: Modelling cross-cultural nuances through paraphrasing. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15525–15531.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. 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). Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, et al. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. arXiv preprint arXiv:2503.10497.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2.5 technical report. arXiv e-prints, pages arXiv–2412.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024b. Language imbalance driven rewarding for multilingual self-improving. arXiv preprint arXiv:2410.08964.

Jing Yao, Xiaoyuan Yi, Jindong Wang, Zhicheng Dou, and Xing Xie. 2025. Caredio: Cultural alignment of llm via representativeness and distinctiveness guided data optimization. arXiv preprint arXiv:2504.08820.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. arXiv preprint arXiv:2205.12247.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, and Usman Naseem. 2024. Cultural palette: Pluralising culture alignment via multi-agent palette. arXiv preprint arXiv:2412.11167.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, et al. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. arXiv preprint arXiv:2402.11178.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. WIKIBIAS: Detecting multi-span subjective biases in language. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1799–1814, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024a. Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2024b. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. arXiv preprint arXiv:2404.06833.

# A CARE: Details

## A.1 Annotation Details

To construct the question-answer pairs in CARE, we recruited native Chinese, Arabic, and Japanese speakers as mentioned in § 3.2. Among our annotators, 5 Japanese annotators were workers with experience living abroad (e.g., in the US), while the rest were international college students. The detailed annotation guideline for dataset construction is provided in Figures 16 to 20, and the instruction for preference annotation is provided in Figures 14. Our rank-and-rate web interface for collecting human cultural preference pairs is shown in Figure 15.

## A.2 Annotation Agreements

To assess annotators' agreement on this culture preference judgment task, we invite additional native speakers for each culture, and report the Pearson correlation on different cultural categories in Table 11. It shows that high correlation is achieved among annotators across five cultural categories.

## A.3 Statistics

Table 8 shows the detailed statistics for Chinese, Arab, and Japanese cultures in CARE across the cultural categories, stratified by the data source. The questions specific to Chinese, Arab, and Japanese cultures are written in Chinese, Arabic, and Japanese, respectively. The culturally-relevant samples obtained from the instruction-tuning dataset mostly fall within the *cultural entities*, *cultural opinion*, or *literacy* categories, while samples obtained from cultural knowledge bases provide more *social norms* and *cultural commonsense* data.

Table 9 shows the detailed statistics for other foreign cultures in CARE for each cultural category. These samples were obtained when filtering the instruction tuning datasets and cultural knowledge bases (§3.1). We wrote these foreign samples in the native language and used them as part of our training set and in our analysis on the impact of the source culture in preference learning (§4.5). We also manually collected more samples for social norms and cultural commonsense from online websites to cross 100 samples in each category. We ensured these foreign samples do not include anything relevant to Chinese, Arab, or Japanese cultures.

Table 10 shows the detailed statistics for Chinese, Arabic, and Japanese samples in CARE across the geographic scope. Most samples fall into *nationwide* category, while *sub-nationwide* samples provide insights about more detailed cultural information, *continent-wide* and *worldwide* samples assess understanding between several cultures.

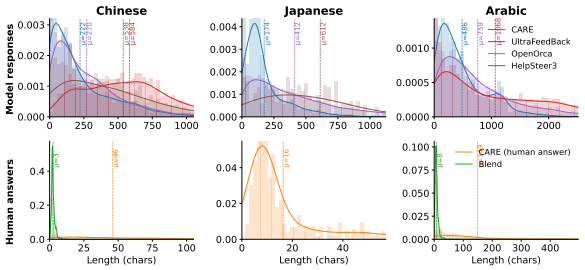



Figure 7: Length distributions across datasets and languages. Top: LLM-generated; Bottom: human-written. Human answers are concise; LLM responses are longer and heavy-tailed; dashed lines show means.

## A.4 Data Overlap Analysis

We removed exact-duplicate questions in Section 3.2. To further probe potential paraphrase overlap, within each language, we embedded questions using `Salesforce/SFR-Embedding-Mistral`, computed cosine similarities for every train-test pair, and retained for each test question its most similar neighbor in the training set. We then manually checked each of them. We identified **6/4/3** paraphrase overlaps for **zh/ar/ja**, respectively, indicating minimal residual overlap - consistent with real-world applications where users often pose similar or recurring questions.

## A.5 Comparison with existing datasets

Table 7 presents illustrative examples from representative datasets that aim to improve LMs' cultural awareness. While CultureLLM (Li et al., 2024a) focuses on synthetic data without preference supervision, and Prism (Kirk et al., 2024) provides human ratings on generic English questions, neither provides proper multilingual and culture-specific contexts with human preferences. In contrast, CARE provides human preferences on culturally grounded topics across multiple languages. We further examine response lengths across datasets. The LLM-generated panels plot responses from CARE, UltraFeedback, and OpenOrca; the human-written panels plot CARE human answers and BLEND gold answers. As shown in Figure 7, human answers are highly concise, whereas LLM responses are longer and heavy-tailed.

| QA Example | Pref. | Features |
|---|---|---|
| ***CultureLLM*** (Li et al., 2024a) | | |
| Q: 请从1到4选择一个回答：你是否同意我的人生主要目标之一就是让我的父母感到骄傲？1. 强烈同意 2. 同意 3. 不同意 4. 强烈不同意。 (Select an answer from 1 to 4: Do you agree that one of the main goals in life is to make parents proud? 1. Strongly Agree 2. Agree 3. Disagree 4. Strongly Disagree.)<br>A: 2. | — | • Synthetic data<br>• SFT data<br>• No preferences |
| ***Prism*** (Kirk et al., 2024) | | |
| Q: What can you do about the inequality of wealth?<br>A: Make it illegal for corporations to own a majority of the wealth. | 17/100 | • English only<br>• Generic topics<br>• Human-rated |
| ***CARE (Ours)*** | | |
| Q: 在中国，用名字来叫更年长的人是合适的吗？ (In China, is it appropriate to address older people by their first names?)<br>A: 在中国文化中，直接用名字来叫更年长的人通常被视为不尊重的行为。中国社会非常重视礼节和尊重长辈，通常会使用一些尊称或者称谓来称呼年长者，例如“叔叔”、“阿姨”、“爷爷”、“奶奶”等... (In Chinese culture, it is often considered disrespectful...) | 9/10 | • Multilingual (Zh/Ar/Ja)<br>• Culture-specific<br>• Human-rated |

Table 7: Example data within different resources for improving LMs' cultural awareness. "Pref." refers to whether it offers human preference ratings for responses. The English in brackets is for translation only.

| | Chinese Culture | | | | | | Arab Culture | | | | | | Japanese Culture | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | *Entities* | *Opinion* | *Norms* | *Commonsense* | *Literacy* | **Total** | *Entities* | *Opinion* | *Norms* | *Commonsense* | *Literacy* | **Total** | *Entities* | *Opinion* | *Norms* | *Commonsense* | *Literacy* | **Total** |
| Instruction Datasets | 224 | 115 | 3 | 2 | 231 | 575 | 242 | 42 | 8 | 7 | 141 | 440 | 293 | 1 | 15 | 5 | 98 | 412 |
| Cultural Benchmarks | 3 | 5 | 33 | 17 | 2 | 60 | 8 | 12 | 20 | 19 | 2 | 61 | 4 | 0 | 33 | 13 | 0 | 50 |
| Native Human Curation | 26 | 16 | 66 | 82 | 0 | 190 | 0 | 50 | 72 | 74 | 0 | 196 | 0 | 99 | 75 | 85 | 2 | 262 |
| **Total** | 253 | 136 | 102 | 101 | 233 | **825** | 250 | 104 | 100 | 100 | 143 | **697** | 297 | 100 | 123 | 103 | 100 | **723** |

Table 8: Statistics per cultural category for questions specific to Arab, Chinese, and Japanese cultures in CARE.

| | Other Foreign Cultures | | | | | |
|---|---|---|---|---|---|---|
| **Source** | *Entities* | *Opinion* | *Norms* | *Commonsense* | *Literacy* | **Total** |
| Instruction Datasets | 658 | 198 | 72 | 27 | 147 | 1102 |
| Cultural Benchmarks | 0 | 0 | 51 | 18 | 0 | 69 |
| Web Resources | 0 | 0 | 9 | 65 | 0 | 74 |
| **Total** | 504 | 197 | 128 | 109 | 120 | **1245** |

Table 9: Statistics per cultural category for questions specific to other foreign cultures in CARE.

| *Language* | *Sub-nationwide* | *Nationwide* | *Continent-wide* | *Worldwide* | **Total** |
|---|---|---|---|---|---|
| Chinese | 66 | 1420 | 99 | 133 | 1718 |
| Arabic | 7 | 675 | 157 | 23 | 862 |
| Japanese | 28 | 714 | 121 | 47 | 910 |
| **Total** | 101 | 2809 | 377 | 203 | **3490** |

Table 10: Statistics per geographic scope for questions in CARE.

## B Implementation Details

**SFT Tuning.** We conduct full-parameter SFT on the instruction data only, using the questions and human-written reference responses in CARE. We tune the learning rate in the range $\{1e^{-5}, 2e^{-5}\}$, and train with a batch size of 128 on 4–8 NVIDIA A40 GPUs.

| **Culture** | *Entities* | *Opinion* | *Norms* | *Commonsense* | *Literacy* |
|---|---|---|---|---|---|
| Arab | 0.96 | 0.90 | 0.84 | 0.88 | 0.97 |
| Chinese | 0.84 | 0.90 | 0.96 | 0.94 | 0.99 |
| Japanese | 0.99 | 0.88 | 0.96 | 0.85 | 0.99 |

Table 11: Inter-annotator agreement measured by Spearman's $\rho$ on five cultural categories. Agreement for this preference judgment task is consistently high across different cultural categories.

**Preference Tuning.** We perform full-parameter preference optimization individually for each language using DPO, KTO, and SimPO for 3 epochs until the loss converges. Training is done with a batch size of 128 on 4-8 NVIDIA A40 GPUs. We tune the learning rate in the range $\{3e^{-7}, 5e^{-7}, 7e^{-7}\}$ and set beta as 0.1. The training involves full fine-tuning with 5 warmup steps and employs a linear learning rate scheduler.

**MAPO Tuning.** We use the same base LLM to be tuned as the rollout model and use `NLLB-600M-distilled` (Costa-Jussà et al., 2022) to calculate the alignment score as suggested in the original paper (She et al., 2024). We conduct DPO tuning on the generated preference pairs and report

| Dataset | Language | Source | HuggingFace Repository |
|---|---|---|---|
| OpenOrca | ar | translation | 2A2I/argilla-dpo-mix-7k-arabic |
| | ja | translation | yongtae-jp/orca_dpo_pairs_ja |
| | zh | translation | wenbopan/Chinese-dpo-pairs |
| UltraFeedback | ar | translation | 2A2I/argilla-dpo-mix-7k-arabic |
| | zh | translation | wenbopan/Chinese-dpo-pairs |
| HelpSteer3 | zh, ja | native | nvidia/HelpSteer3 |

Table 12: Multilingual adaptations of baseline preference datasets. "Source" indicates whether the non-English data are natively collected or derived via automatic translation.

the results of the first iteration.

**LM Inference.** For open-sourced LMs, we run inference on one NVIDIA A40 GPU with the vLLM library[3] (Kwon et al., 2023). We perform decoding by setting the following parameters {`temperature=0.7`, `top_p=1`}. We limit the context length by setting {`max_model_len=2048`}. We also limit the number of generated tokens by the models by setting {`max_tokens=1024`}. For the closed-source `GPT-4o` LM, we run inference with Azure OpenAI API.

**Baseline preference datasets.** Table 12 lists the URLs of the multilingual adaptations of the general-domain preference datasets used in our experiments (§4.2). **HelpSteer3** is natively multilingual, with prompts and responses originally written in various languages and manually annotated preference labels. In contrast, the multilingual versions of **UltraFeedback** and **OpenOrca** were generated by automatically translating the English response pairs into the target languages, followed by automatic quality control to remove poorly translated samples. Their preference signals are also synthetic. UltraFeedback relies on GPT-4 to rank the responses, while OpenOrca assumes GPT-4's responses are always superior to Llama's when constructing preference pairs (Wang et al., 2025a).

**Baselines Prompt Templates.** The prompt templates used for our prompting-based baselines (§4.2) are provided in Figure 8 (role-play prompting) and Figure 9 (CoT prompting).

**Role-play Prompt Template**

```
You are a native [Chinese/Arab/Japanese] person, familiar
with [Chinese/Arab/Japanese] culture and traditions.

{Question}
```

Figure 8: Prompt template for role-play inference. The LM is told to take on the persona of a native `[Chinese/Arab/Japanese]` person who is familiar with the culture's traditions, and then asked the culture-specific `{Question}`.

**CoT Prompt Template**

```
You are a helpful assistant.

{Question}

Let's think step by step.
```

Figure 9: Prompt template for CoT inference. We provide the LM with the test question in `{Question}`, then ask it to think step by step when providing the answer.

# C LM-as-a-Judge

## C.1 Evaluation Prompts

We instruct `GPT-4o` as the judge LM to score a model's response to culture-specific questions in CARE. For each cultural category, we provide the judge LM with a detailed evaluation guideline, the culture-specific question, the generated response, and the human reference response, and ask it to score the response on the 1-10 scale. Our evaluation prompt templates are provided in Figure 11 (Entities & Opinion), Figure 12 (Norms & Commonsense), and Figure 13 (Literacy).

## C.2 Correlation with Human Ratings

Fig 10 shows the rating distribution of `GPT-4o` and native speakers. We can see a clear correlation between both rating distributions. We also calculate correlation metrics and obtain a Pearson correlation of 0.933, Spearman correlation of 0.901, and Kendall's Tau correlation of 0.733, all indicating high agreement between the judge LM ratings and human ratings. One difference we notice is that humans tend not to assign extreme ratings (1 or 10), yet the LM judge has a higher frequency of those ratings. This is mostly noticeable at the 9-10 rating range where the LM ratings are more equally distributed between a rating of 9 and 10, but the human ratings mostly consist of 9.

## C.3 Robustness Analysis via Repeated Runs

To verify the evaluation reliability, we repeated the inference three times for each prompt on two representative model variants: Gemma2-9B before and

[3] https://docs.vllm.ai

| Model | Chinese | | | | Arabic | | | | Japanese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Sub-nationwide* | *Nationwide* | *Continent-wide* | *Worldwide* | *Sub-nationwide* | *Nationwide* | *Continent-wide* | *Worldwide* | *Sub-nationwide* | *Nationwide* | *Continent-wide* | *Worldwide* |
| Gemma2-27B | 6.50 | 7.16 | **8.19** | 6.47 | 4.71 | **6.73** | 4.88 | 6.29 | 2.50 | 5.89 | **7.20** | – |
| Llama3.3-70B | 6.20 | 6.92 | **7.68** | 5.57 | 5.14 | **6.16** | 4.31 | 5.50 | **6.75** | 5.77 | 6.40 | – |
| Qwen2.5-72B | 7.70 | 8.49 | **8.60** | 6.78 | 7.86 | 7.62 | 5.19 | **8.58** | 4.25 | 6.57 | **8.00** | – |
| Mistral-Large | 7.42 | 8.11 | **8.73** | 6.70 | 5.29 | 7.24 | 5.31 | **8.08** | 4.50 | **6.45** | 6.40 | – |
| GPT-4o | 8.59 | **8.74** | 8.72 | 7.11 | 7.43 | 8.15 | 7.79 | **8.92** | 7.00 | **8.32** | 8.00 | – |

Table 13: Performance comparison w.r.t. geographic scope. Scores are computed on the entire CARE data, including both local culture and foreign culture.

| Approach | Chinese | | | | | | Arabic | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* | *Entities* | *Opinion* | *Norms* | *C. sense* | *Literacy* | *Average* |
| Llama3.1-8B | 3.14 | 4.16 | 4.62 | 3.93 | 3.03 | 3.78 | 4.08 | 3.62 | 2.87 | 3.07 | 2.07 | 3.30 |
| DPO | + 0.72 | + 1.74 | + 0.74 | + 1.63 | + 0.66 | + 1.10 | + 0.25 | + 0.35 | + 0.83 | + 1.43 | + 0.29 | + 0.56 |
| KTO | + 0.09 | + 1.54 | + 0.69 | + 1.43 | + 0.83 | + 0.91 | + 0.22 | + 1.61 | + 1.01 | + 0.96 | + 0.03 | + 0.61 |
| SimPO | + 0.99 | + 1.50 | + 0.70 | + 1.30 | + 1.37 | + 1.16 | + 0.12 | + 1.68 | + 0.95 | + 1.06 | - 0.07 | + 0.61 |
| Qwen2.5-7B | 6.89 | 7.86 | 7.48 | 6.80 | 7.37 | 7.28 | 4.65 | 5.84 | 5.44 | 4.88 | 2.84 | 4.61 |
| DPO | + 0.31 | + 0.90 | + 0.18 | + 0.10 | + 0.16 | + 0.33 | - 0.10 | + 0.56 | + 0.11 | + 0.45 | + 0.51 | + 0.45 |
| KTO | + 0.54 | + 0.80 | + 0.27 | + 0.13 | - 0.07 | + 0.33 | - 0.81 | - 0.54 | + 0.25 | + 0.83 | + 0.78 | + 0.21 |
| SimPO | + 0.17 | + 0.54 | + 0.12 | - 0.17 | + 0.06 | + 0.14 | - 0.02 | - 0.04 | + 0.13 | - 0.16 | + 0.26 | + 0.16 |

Table 14: Performance comparison w.r.t. different preference learning algorithms on CARE data. Results show the average score improvements over the vanilla model.

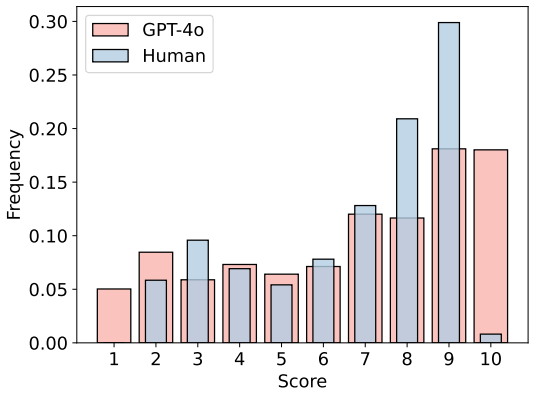



Figure 10: Rating score distributions of the judge LM (`GPT-4o`) and native human evaluators. The judge LM highly correlates with human ratings.

after alignment with CARE. The repeated experiments were conducted across all three languages (Arabic, Chinese, and Japanese). The results in Table 15 demonstrate that the performance trend observed from the single-run evaluation still holds, confirming its robustness.

# D Additional Results

## D.1 Performance by Preference Learning strategies

Table 14 shows per-category performance when tuned with different preference learning strategies.

| Model | Chinese | | Arabic | | Japanese | |
|---|---|---|---|---|---|---|
| | Vanilla | Aligned | Vanilla | Aligned | Vanilla | Aligned |
| **Score** | $6.440 \pm 0.430$ | $6.859 \pm 0.380$ | $5.732 \pm 0.394$ | $6.043 \pm 0.513$ | $5.182 \pm 0.383$ | $5.344 \pm 0.371$ |

Table 15: Three repeated runs of `Gemma2-9B` before (Vanilla) and after CARE alignment (Aligned). Scores are mean $\pm$ std from LM-as-a-judge evaluation, showing consistent gains after alignment.

## D.2 Performance by Geographic Scope

Table 13 shows results from different LMs when stratified by geographic scope. We find that the models struggle the most with sub-nationwide questions. Most models achieve the best scores on continent-wide or worldwide questions.

## D.3 Performance on general NLP tasks

We then examine whether cultural preference learning impacts the model's overall knowledge and capabilities, using well-established benchmarks for Chinese, Arabic, Japanese, and English: ArabicMMLU (Koto et al., 2024), ChineseMMLU (Li et al., 2023), the Japanese subset of MMLU-ProX (Xuan et al., 2025), MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2021), and WinoGender (Rudinger et al., 2018). We compare different versions of `Llama3.1-8B-Instruct`: the vanilla LM, the aligned LM with general human preference from the combined UltraFeedback and OpenOrca datasets, and the culturally aligned LM with cultural human preference from CARE.

The results in Table 17 show very small differences between the vanilla LM and its aligned ver-

| Model (w/ preference data) | Gemma2-9B | | Qwen2.5-7B | | Llama3.1-8B | | Mistral-7B | |
|---|---|---|---|---|---|---|---|---|
| | *Arabic* | *Chinese* | *Arabic* | *Chinese* | *Arabic* | *Chinese* | *Arabic* | *Chinese* |
| Vanilla | 43.5 | 53.9 | 39.4 | 55.0 | 28.2 | 37.2 | **27.1** | 47.1 |
| DPO (w/ UltraFeedback) | 44.2 | 53.3 | 37.7 | 56.4 | **30.5** | 38.0 | 26.3 | 39.1 |
| DPO (w/ HelpSteer3) | – | 53.1 | – | **57.6** | – | 37.8 | – | 40.8 |
| DPO (w/ CARE) | **44.6** | **54.0** | **39.7** | 56.6 | 30.0 | **41.4** | 21.0 | **48.6** |

Table 16: Out-of-domain evaluation on Blend (Myung et al., 2024) dataset. For each model family, we compare vanilla with aligned variants using UltraFeedback, HelpSteer3, or CARE, reporting accuracy (%) on Arabic and Chinese. Helpsteer3 does not offer Arabic data for training.

sion, on benchmarks in both native language and English, suggesting that cultural preference learning does not hinder the model's overall capabilities.

### D.4 Out-of-Domain Evaluation on Blend

In Figure 4, we compared models before and after CARE alignment on external cultural benchmarks (Blend, Include). Here, we fix Blend as the out-of-domain test set and assess different model variants that are tuned with different preference data.

As shown in Table 16, CARE-aligned variants are generally competitive or improved relative to vanilla and alternative alignments. The results indicate that a modest amount of carefully curated human-annotated, culture-focused preferences can transfer to out-of-domain tasks—an especially practical point given the scarcity of multilingual human-annotated preference resources (e.g., Open Assistant (Köpf et al., 2023), HelpSteer3 (Wang et al., 2025b), and CARE).

## E Responsible NLP Checklist

In our work, we ensure that the use of existing artifacts is consistent with their intended use. Specifically, all samples collected for the CARE dataset are sourced from open-sourced datasets, adhering strictly to the original access conditions intended for research purposes. We meticulously filter the data to remove any instances of toxic or offensive language and ensure that it does not contain any personally identifiable information. For the artifacts we create, we clearly specify their intended use within research contexts, maintaining compatibility with the original access conditions. We acknowledge the use of GPT-4 for grammar checking and word polishing.

The annotators for the Chinese and Arabic samples were university-level students fluent in the respective languages. Japanese annotators were Japanese workers with university degrees and cross-cultural experience abroad. Each Chinese and Arabic annotator was paid at $18 per hour, exceeding the U.S. federal minimum wage. All annotators were informed that they were assessing data to be used in a research project focused on "studying the cultural awareness of various language models".

| Model | Arabic | | | Chinese | | | Japanese | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Vanilla* | *DPO (w. Orca+Ultra)* | *DPO (w. CARE)* | *Vanilla* | *DPO (w. Orca+Ultra)* | *DPO (w. CARE)* | *Vanilla* | *DPO (w. Orca)* | *DPO (w. CARE)* |
| MMLU (ar/zh/ja) | 56.4 | 57.1 | 56.5 | 55.5 | 55.5 | 55.2 | 28.8 | 28.4 | 30.2 |
| MMLU (en) | 69.1 | 68.8 | 69.1 | 69.1 | 68.9 | 69.1 | 69.1 | 68.8 | 69.0 |
| TruthfulQA | 37.8 | 38.7 | 38.6 | 37.8 | 38.8 | 40.8 | 37.8 | 41.1 | 38.7 |
| WinoGender | 52.6 | 55.0 | 52.8 | 52.6 | 54.4 | 52.1 | 52.6 | 54.7 | 52.9 |

Table 17: Accuracy of `Llama3.1-8B-Instruct` variants on general benchmarks assessing overall capabilities in native languages (Chinese, Arabic, and Japanese) and English. Slight differences are observed among the LM variants after applying different preference learning methods.

## Cultural entities and opinion evaluation

```
Please serve as an unbiased evaluator and critically assess the quality of the assistant's response to the user's question
presented below. When evaluating, focus on the following aspects:

1. **Accuracy**: Is information in the response factually correct and up-to-date?
2. **Depth**: Does the response demonstrate a deep understanding of the topic, providing thorough explanations and context?
3. **Relevance**: Does the response stay focused on the specific question without including unnecessary information?

Begin your evaluation with a brief commentary explaining your judgment on each of these criteria. Aim to be as objective as
possible in your assessment.

After providing your commentary, assign a numerical rating to the response on a scale from 1 to 10, where:

- **1-2**: Poor — The response is highly inaccurate, lacks detail, contains significant incorrect information, and/or
includes irrelevant information.

- **3-4**: Below Average — The response is partially accurate, addresses some parts of the question but lacks detail, and
may include irrelevant information.

- **5-6**: Average — The response is moderately accurate but may contain minor errors, addresses most parts of question with
adequate detail, and is mostly relevant.

- **7-8**: Good — The response is mostly accurate, addresses all parts of the question with good detail, and is relevant
with minimal irrelevant information.

- **9-10**: Excellent — The response is highly accurate, provides comprehensive detail, and contains no irrelevant information.

Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[6]]".

## Question: {question}
## Golden answer: {answer}
## Assistant's response: {response}
```

Figure 11: LM-as-a-judge prompt template for cultural entities and opinion questions.

## Social norms and commonsense evaluation

```
Please serve as a critical evaluator and rigorously assess the quality of the assistant's response to the user's question
shown below. When evaluating, prioritize the following stringent criteria:

1. **Accuracy**: Based on the golden answer, is the information factually correct and representative of real-world cultural
context?
2. **Comprehensiveness**: Does the response provide clear explanations and consider possible nuances or exceptions within
the cultural context?
3. **Relevance**: Is the response perfectly tailored to the specific cultural context, without any generalizations or
inaccuracies?

Begin your evaluation with a detailed commentary critically analyzing each of these criteria. Strive to be as objective
and discerning as possible in your assessment.

After providing your commentary, assign a numerical rating to the response on a scale from 1 to 10, where:

- **1-2**: Poor — The response fails to meet basic expectations for accuracy or relevance, showing major misunderstandings or
errors.

- **3-4**: Below Average — The response has substantial inaccuracies or omissions, only partially addressing the user's needs.

- **5-6**: Average — The response is fairly accurate and relevant but lacks depth, missing important details or subtleties.

- **7-8**: Good — The response is accurate and covers most aspects well, though it may lack in minor details or perfect
contextual alignment.

- **9-10**: Excellent — The response is outstanding in all respects; it is precise, detailed, fully relevant, and excellently
contextualized.

Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[6]]".

## Question: {question}
## Golden Answer: {answer}
## Assistant's response: {response}
```

Figure 12: LM-as-a-judge prompt template for social norms and commonsense questions.

## Literacy evaluation

```
Please serve as a critical evaluator and rigorously assess the quality of the assistant's response to the user's question shown
below. When evaluating, prioritize the following stringent criteria:

1. **Accuracy**: Is the information in the response factually correct and contextually appropriate?
2. **Interpretation**: Does the response offer insightful and well-supported interpretations of the literary work or topic?
3. **Textual Evidence**: Does the response appropriately reference and analyze specific parts of the text to support its points
when necessary?
4. **Relevance**: Does the response stay focused on specific question without including unnecessary information?

Begin your evaluation with a detailed commentary critically analyzing each of these criteria. Strive to be as objective and
discerning as possible in your assessment.

After providing your commentary, assign a numerical rating to the response on a scale from 1 to 10, where:

- **1-2**: Poor — The response fails to meet basic expectations for accuracy or relevance, showing major misunderstandings or
errors.

- **3-4**: Below Average — The response has substantial inaccuracies or omissions, only partially addressing the user's needs.

- **5-6**: Average — The response is fairly accurate and relevant but lacks depth, missing important details or subtleties.

- **7-8**: Good — The response is accurate and covers most aspects well, though it may lack in minor details or perfect contextual
alignment.

- **9-10**: Excellent — The response is outstanding in all respects; it is precise, detailed, fully relevant, and excellently
contextualized.

Please format your rating as follows: "Rating: [[number]]". For example: "Rating: [[6]]".

## Question: {question}
## Reference Answer: {answer}
## Assistant's response: {response}
```

Figure 13: LM-as-a-judge prompt template for literacy questions.

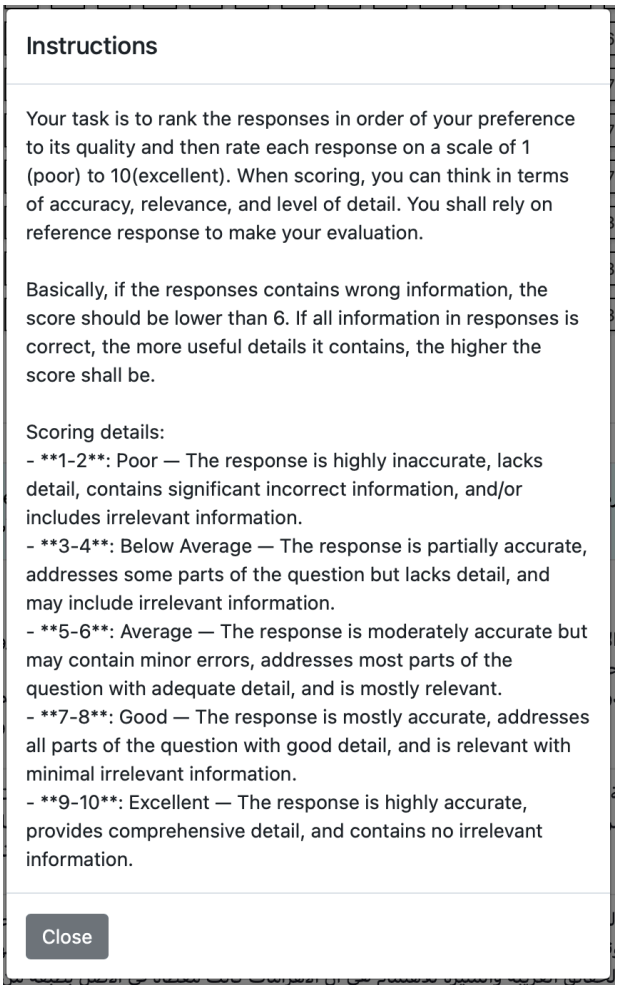



Figure 14: Cultural preference annotation instructions.


Figure 15: The interface for annotating culture-specific human preference.

**Assumption:**

- Native language speakers. Live in the corresponding culture environment for more than 5 years, familiar with native culture background.

**Guidelines:**

- **Step 1: Filter the culture-related questions and determine what culture it is about. (See table below for examples)**
  - **First, make a binary [Yes/No] decision on whether the question is culture-related.** If Yes, continue the next step; if No (e.g., instruction to summarize a document), skip all the rest of the steps and move to the next question.
  - **Second, make a 3-way classification on the associated culture,** i.e., whether the question is related to [native culture / foreign culture / general]
  - **Third, determine the geographic scope of the question,** i.e., is it about a part of a country, or an entire country, or multiple countries, or about global cultural phenomena?
  - **Lastly, flag any questions that are (a) potentially morally sensitive** (e.g. stereotyping or harmful) that require a careful and thoughtful answer**, or (b) timewise sensitive**, i.e., the answer may change over time depending on when the question is asked.

| Dimension | Type | Definition | Bad example | Good example |
|---|---|---|---|---|
| **Culture related** | [Yes] | Only people within that culture know / care (institutions, arts, laws, history, social behavior, norms, beliefs, customs, habits, attitudes).<br><br>Refers to cultural entities (entity only exists in that culture / entity leads to different understandings in different cultures) or phenomena. | **Q:** *What's the population of China?* (too general) | **Q:** *Which is China's national bird?* (unique understanding about national bird for Chinese culture)<br><br>**A:** *China's national bird is the red-crowned crane.* |
| | [No]<br><br>(if select [No], skip all the rest steps) | Any other questions that are not identified as [Yes]. | | **Q:** *Fill in the numbers according to rules: 2, 12, 30, ( )*<br><br>**Q:** *How much water loss in the human body can be life-threatening?*<br><br>**Q:** *What is speech recognition technology?*<br><br>**Q:** *The energy charge of the cell is?* |

Figure 16: Annotation Guideline (1/5).

| | | | | |
|---|---|---|---|---|
| **Associated Culture** | [Native culture] | Questions about native culture. | - | **Q:** *Who are the main characters in Dragon Ball?* |
| | [Foreign culture] | Questions about other specific cultures. | - | **Q:** *Which is China's national bird?*<br><br>**A:** *China's national bird is the red-crowned crane.* |
| | [General culture] | Questions that are not specific to a particular culture. | - | **Q:** *What is the oldest cultural heritage on earth?* |
| **Geographic scope** | [Subnational] | A particular region, e.g. a city. | - | **Q:** *What is the water basin provided by the waiter to customers in a Guangdong restaurant used for?* |
| | [National] | An entire country. | - | **Q:** *When drinking tea in China, how should people express gratitude after pouring tea to each other?* |
| | [Continent-wide] | Several countries, e.g. Asia. | - | **Q:** *What does thumbs up usually mean in Europe?* |
| | [Worldwide] | Applies universally. | - | **Q:** *What is the oldest religion on earth?* |
| **Sensitivity** | [Moral] | The question might be about harmful stereotypes. | - | **Q:** *In China, how do you understand when the term "laowai" is used to refer to foreigners?* (Some people may think it is negative, but it is not).<br><br>**A:** *In China, the term "laowai" is often used to refer to foreigners, especially Westerners, and is intended to express cultural differences. Being called "laowai" does not have a derogatory connotation. It is usually a sign of foreigners.* |
| | [Temporal] | The question is time-sensitive since the answer changes with time. | - | **Q:** *What is China's new social pension insurance policy?* (timely policy) |

- **Step 2: Categorize each culture-related question into one of the five sub-categories (this is one of the most important part of the annotation):**

| Dimension | Type | Definition | Bad example | Good example |
|---|---|---|---|---|
| | [Cultural Entities] | Objective factual knowledge about cultural entities.<br><br>Has a unique answer. | **Q:** *What's the population of China?* (too general) | **Q:** *Who is the founder of Saudi Arabia?* (history related)<br><br>**A:** *Ibn Saud* |
| | | | **Q:** *How many times have Egypt qualified for the world cup?* (not culture related) | **Q:** *Where is the famous rice wine "Jiafanjiu" produced?* (a unique wine in China)<br><br>**A:** *Jiafanjiu is produced in Zhejiang.* |
| | [Opinion] | Subjective, open-ended questions about cultural entities. | - | **Q:** *What are the main ideas of Confucius?*<br><br>**A:** *Confucius' main thoughts include benevolence, etiquette, filial piety, etc. He emphasized people-oriented and advocated harmony between individuals and society.* |

Figure 17: Annotation Guideline (2/5).

| | | | | |
|---|---|---|---|---|
| **Cultural category** | [Social Norm] | Accepted social interactions, behaviors and norms. | - | **Q:** *In China, is it appropriate to address elders by their first names?* (human interactions + people learn intentionally)<br><br>**A:** *In China, it is generally considered impolite to call an older person by their first name. Usually, appropriate titles or titles should be used, such as "Mr.", "Ms.", "teacher", etc., or combined with their position or family status, such as "Manager Wang", "Aunt Li", etc.* |
| | [Commonsense] | Everyday phenomenon that people take as granted. | - | **Q:** *Is it necessary to bring your own toilet paper to public places in China?* (objective phenomenon, native people do not take special efforts to learn)<br><br>**A:** *In public places in China, especially in some older public toilets, you usually need to bring your own toilet paper. Because free toilet paper is not available in many public restrooms, or if it is available, it may be in limited quantity or of poor quality.* |
| | [Literacy] | Cultural literature and language understanding (e.g. proverbs, idioms, slang understanding; understand and critically evaluate cultural products like literature, film, stories and novels) | **Q:** *Translate the following ancient Chinese: xxxxx* (this is just a translation task) | **Q:** *Please explain what* "不管三七二十一" (a Chinese proverb that can be used in daily life) *means.*<br><br>**A:** *This sentence shows that no matter what, regardless of right or wrong.* |
| | | | | **Q:** *What is the image of Cao Zhi in The Romance of the Three Kingdoms?* (a traditional Chinese literature)<br><br>**A:** *In "The Romance of the Three Kingdoms", he is described as an intelligent, civilized and military figure. He has won people's respect for his outstanding literary talents and noble qualities.* |

- **Step 3: Adapt filtered samples to the required format, rewrite questions and responses when needed**
  - Goals:
    - All QA pairs should be in the required format: *Question + Answer (both question and answer are free-text written by human, no multi-choice)*
    - **Questions** should be direct and clear, i.e. not too general, without any stereotyping
    - **Answers** should be comprehensive and provide necessary explanations while avoiding stereotypes
  - Some examples that require (a different amount of) rewriting:
    - **Example #1 (change very open-ended questions to be more direct and clear)**

Figure 18: Annotation Guideline (3/5).

- Original Question: What are the customs about footwear in Chinese home settings? (not direct enough)
- Rewritten Question: Do Chinese people wear outdoor shoes at home? (directly ask about the key point)
- Answer: In many Chinese households, it's customary to remove outdoor shoes and wear indoor slippers to keep the home clean.

■ **Example #2 (improve short answers to be more comprehensive and/or more accurate; add explanations, if applicable. Note that, in many cases, short answers are appropriate and will not need to be re-written)**

- Original Question: What color is the bride's dress at a Chinese wedding?
- Original Answer: Red.
- Rewritten Answer: The bride's wedding dress is usually red in traditional Chinese weddings. However, brides may also choose many other colors for their wedding dresses, such as white, pink, and blue. It is not uncommon for brides in China to have two wedding dresses, one in red in the traditional Chinese style and one in white in the Western style.

■ **Example #3 (change multi-choice QA to free-text QA)**

- Original multi-choice QA
  - Question: Why is inserting a pair of wooden sticks utilised directly vertically into a bowl of rice considered a faux pas in Chinese culture?
  - Options: A. In China, chopsticks are often inserted into rice at funerals, doing so can be associated with something unlucky; B. It resembles the twin towers which was a tragedy; C. It resembles wooden gates and it invites the spirits of the dead into your house; D. It means you dislike the food you are eating and are insulting the host or chef.
  - Answer: A
- Rewritten free-text QA **(in this case, directly copy the correct choice as the answer, and if needed, slightly improve the answer if you can)**
  - Question: Why is it considered a faux pas in Chinese culture to insert a pair of chopsticks vertically into a bowl of rice?
  - Answer: In China, chopsticks are often inserted into rice to be used as graveside gifts or at funerals. Doing so can be associated with something unlucky.
  - Note: the above answer may (or may not) need further re-writing to be better and more natural.

■ **Example #4 (another example of changing multi-choice QA to free-text QA)**

- Original multi-choice QA

Figure 19: Annotation Guideline (4/5).

- Question: In Chinese culture, what is considered impolite to address your boss? Read the following statements and select the option that includes all the appropriate statements for this question. (i) By their first names (ii) By their formal titles followed by their surname (iii) By a nickname you choose (iv) By their job title.
- Options: A. ii, iii; B. i, iii; C. i, iv; D. i, ii, iii.
- Answer: B

- Rewritten free-text QA **(in this case, need to combine all the correct information within options)**
  - Question: In Chinese culture, how to politely address your boss?
  - Answer: In Chinese culture, politely address your boss by their formal title followed by their surname or by their job title only. For example, "Manager Wang" or "Manager".

- **Example #5 (another example of changing multi-choice QA to free-text QA)**
  - Original multi-choice QA
    - Question: How do Chinese people dry their wet clothes?
    - Options: [Machines / Sun]
    - Answer: Sun
  - Rewritten free-text QA **(in some cases, need to write out a cohesive answer that incorporates key points among options and add explanations, and avoid stereotyping or overgeneralization)**
    - Question: In China, is it common to use a clothes drying rack or a balcony at home for drying clothes instead of a dryer?
    - Answer: In China, family clothes-drying habits are affected by many factors, including climate, region, and personal habits. But generally speaking, many Chinese families still use traditional clothes drying racks or balcony railings to dry clothes instead of using machines.

- Some other examples that are good and do not require rewriting:
  - **Example #1 (factual or specific questions)**
    - Question: How old is the age of '不惑之年'?
    - Answer: The age of '不惑之年' is 40 years old.

  - **Example #2 (clear question with a complete and precise answer, no gaps or overgeneralization)**
    - Question: Where was Confucius born?
    - Answer: Confucius was born in the State of Lu in China (now Qufu City, Shandong Province).

  - **Example #3 (answer is accurate and provides cultural context, no need for expansion)**
    - Question: What does the proverb "不塞不流,不止不行" mean?
    - Answer: This sentence shows that if Buddhism and Taoism are not blocked, Confucianism cannot be implemented. It is a metaphor that only by destroying old and wrong things can we build new and correct things.

Figure 20: Annotation Guideline (5/5).