CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis

Ruixiang Feng[1] **Shen Gao**[1 *] **Xiuying Chen**[2] **Lisi Chen**[1] **Shuo Shang**[1 *]
[1]University of Electronic Science and Technology of China,
[2]MBZUAI
`{202421081331, shengao, chenlisi}@uestc.edu.cn`
`xiuying.chen@mbzuai.ac.ae`, `jedi.shang@gmail.com`

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across various tasks, yet they often exhibit a specific cultural biases, neglecting the values and linguistic diversity of low-resource regions. This cultural bias not only undermines universal equality, but also risks reinforcing stereotypes and perpetuating discrimination. To address this, we propose CulFiT, a novel culturally-aware training paradigm that leverages multilingual data and fine-grained reward modeling to enhance cultural sensitivity and inclusivity. Our approach synthesizes diverse cultural-related questions, constructs critique data in culturally relevant languages, and employs fine-grained rewards to decompose cultural texts into verifiable knowledge units for interpretable evaluation. We also introduce GlobalCultureQA, a multilingual open-ended question-answering dataset designed to evaluate culturally-aware responses in a global context. Extensive experiments on three existing benchmarks and our GlobalCultureQA demonstrate that CulFiT achieves state-of-the-art open-source model performance in cultural alignment and general reasoning. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, including reasoning (Ahn et al., 2024; Huang et al., 2023), natural language understanding (Yuan et al., 2024; Bi et al., 2024), and daily communication. Owing to their advanced functionalities, LLMs have gained widespread popularity globally. However, these models often exhibit a Western-centric perspective (Wang et al., 2023b; Shen et al., 2024) and tend to neglect the values and differences of regions with low-resource languages (Naous et al., 2024). This cultural bias not only challenges the principle of universal equality but also poses significant risks, such as reinforcing stereotypes, perpetuating discrimination, and potentially inciting social conflicts. Consequently, there is an urgent need to develop models that are culturally sensitive and inclusive, ensuring they respect and reflect the diversity of global cultures. To address the issue of cultural bias, recent studies (Fung et al., 2024; Nguyen et al., 2023; Huang and Yang, 2023; Liu et al., 2024) use LLMs to generate cultural-related texts and filter data through cleaning pipelines or human annotations. Li et al. (2024b) fine-tune culture-specific LLMs using data obtained from multi-agent communication and employ the model to tackle hate-speech detection tasks across countries.

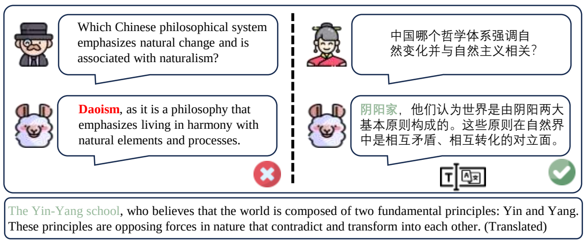

Which Chinese philosophical system emphasizes natural change and is associated with naturalism?
Daoism, as it is a philosophy that emphasizes living in harmony with natural elements and processes.
中国哪个哲学体系强调自然变化并与自然主义相关？
阴阳家，他们认为世界是由阴阳两大基本原则构成的。这些原则在自然界中是相互矛盾、相互转化的对立面。
The Yin-Yang school, who believes that the world is composed of two fundamental principles: Yin and Yang. These principles are opposing forces in nature that contradict and transform into each other. (Translated)

Figure 1: An example of language inconsistency. When asked cultural-specific questions, LLMs can generate correct answers in the local language but fail to provide appropriate responses in English.

Previous approaches have primarily relied on descriptive, monolingual text (*e.g.,* English) to train LLMs with cultural knowledge (Fung et al., 2024; Shi et al., 2024). However, understanding cultural queries often depends on the dialogue context, and such dialogue usually uses the culturally-relevant languages (*e.g.,* Malay for Singaporean culture, Chinese for Chinese culture). Therefore, learning cultural knowledge within culturally relevant linguistic contexts is crucial. As shown in Figure 1, when asked cultural-specific questions, LLMs can generate a correct answer in local language scenario but fail to provide appropriate re-

---

[*] Corresponding authors

[1] Code is available at https://github.com/MMadmax/CulFiT

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22413–22430
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

sponses in English, which shows a language inconsistency phenomenon. This phenomenon further underscores the need for culturally diverse and linguistically inclusive training approaches.

Additionally, existing evaluation methods in cultural domains are coarse-grained, often relying on metrics such as text overlap, binary classification, or multiple-choice questions (Chiu et al., 2024b; Fung et al., 2024; Shi et al., 2024). However, these methods usually fail to account for the inherent flexibility of cultural queries (Pawar et al., 2024), creating a gap between the evaluation and real-world cultural knowledge applications. These methods also lack interpretability in assessing cultural understanding, thereby undermining the reliability of the evaluation results.

In this paper, we propose Target-aware **Cul**tural Data Synthesis and **Fi**ne-grained **T**raining (CulFiT), a novel cultural-aware training paradigm that leverages target-aware multilingual data for model training and employs fine-grained rewards as training signals. Specifically, we first synthesize diverse cultural-related questions based on descriptive cultural knowledge texts. Next, we construct critique data from the content generated by the target model, which is then translated into multiple culturally relevant languages. This multilingual dataset is subsequently used to train the LLM. To provide fine-grained feedback for model training, we introduce fine-grained reward modeling, which decomposes culturally relevant texts into verifiable knowledge units, enabling a quantized interpretable evaluation of cultural alignment.

Based on the proposed data construction method, we introduce a new culturally-aware benchmark dataset **GlobalCultureQA**, which is designed for multilingual open-ended question-answering settings, focusing on evaluating the ability to generate culturally-aware answers in a global context. Extensive experiments conducted on three commonly used benchmarks and GlobalCultureQA show that our CulFiT achieves state-of-the-art performance on open-source models. We also explore the cultural alignment of our models based on Hofstede cultural dimensions and further investigate the effectiveness of how multi-lingual data increases robustness across various languages.

The main contributions of this work are as follows:

- We propose CulFiT, which employs multi-lingual critique data synthesis for fine-grained culturally-aware model training.
- We propose a target-aware data critique method to specifically address the cultural knowledge gaps in the target model, enhancing its robustness in multilingual scenarios.
- We introduce a fine-grained reward to quantitatively evaluate the cultural alignment.
- Experiments conducted on our newly proposed GlobalCultureQA and three benchmarks demonstrate the effectiveness of CulFiT in terms of cultural-aware metrics and general reasoning capabilities of LLM.

2 Related Work

Cultural Bias in LLM Numerous studies have revealed that LLMs exhibit an unequal representation of world values across different regions and countries (Li et al., 2024c; AlKhamissi et al., 2024). Specifically, they often reflect a Western-centric perspective (Wang et al., 2023b; Shen et al., 2024) and overlook values from regions with low-resource languages (Naous et al., 2024). To address this issue, a growing body of research has focused on enhancing the cultural awareness of LLMs. For instance, Choenni and Shutova, 2024; Tao et al., 2024 found that employing culturally-aware prompts can enhance model performance by leveraging the internal cultural knowledge of LLMs. Similarly, Li et al., 2024a utilized surveys such as the World Value Survey (Survey, 2022) as seed questions and augmented them semantically to fine-tune a more culturally-aware model.

Cultural Data Synthesis Significant progress has been made in the development of datasets related to cultural aspects. Huang and Yang, 2023; Lee et al., 2024 construct the cultural datasets through human annotation, which is labor-intensive and difficult to scale. Meanwhile, many works develop data cleaning pipelines from social media platforms such as TikTok, Reddit (Shi et al., 2024; Nguyen et al., 2023) and Wikimedia (Fung et al., 2024; Liu et al., 2024). Rao et al., 2024; Shum et al., 2023 synthesize their data from existing datasets and transfer cultural knowledge to specific domains such as norms or etiquette. However, most cultural datasets are predominantly composed in English, limiting their ability to effectively capture the context of real-world scenarios.

Cultural Benchmarks Extensive research has also focused on developing cultural benchmarks, which can be categorized into: culture-specific benchmarks and multicultural benchmarks.

Culture-specific benchmarks are designed to evaluate LLMs' cultural capacities in specific regions and countries, such as Southeast Asia (Wang et al., 2023a) and China (Sun et al., 2024). On the other hand, multicultural benchmarks aim to explore cultural diversity, constructed by human annotation (Chiu et al., 2024b; Myung et al., 2024), model generation (Putri et al., 2024), and human in the loop (Chiu et al., 2024a). However, these methods primarily rely on multiple-choice or Yes/No questions, which are prone to positional bias and lack fine-grained evaluation in open-ended scenarios.

3 CulFiT

3.1 Overview

As illustrated in Figure 2, our Target-aware **Cul**tural Data Synthesis and **Fi**ne-grained **T**raining (CulFiT) comprises three components: (1) **Target-aware critique generation** reflects the common errors for the target LLM (§3.2). (2) **Multilingual data synthesis** increases the generalization ability in real-world application scenarios by augmenting the cultural-aware data (§3.3). (3) **Fine-grained model training** provides interpretable evaluation protocols to optimize the LLM (§3.4).

3.2 Target-aware Data Critique

Data Synthesis We first construct cultural-aware QA pairs from three widely-used data sources: CANDLE (Nguyen et al., 2023), CultureAtlas (Fung et al., 2024), and CultureBank (Shi et al., 2024). However, these datasets primarily contain discrete, assertive statements that fail to reflect how cultural concepts naturally emerge when chatting with users. To address this limitation, we first aggregate related cultural statements by topic and synthesize them into coherent knowledge paragraphs K by employing a data generation model $\mathcal{G}$. We then employ prompting strategies to generate culturally-grounded questions Q based on the knowledge K, with automated verification by using $\mathcal{G}$ to ensure each question is answerable using K. Then we generate two type of answers with two different LLMs: (1) *Golden Answer* (A_g): Produced by data generation LLM $\mathcal{G}$ through knowledge-aware synthesis. (2) *Target-aware Answers* (A_t): Generated by the target model $\mathcal{M}$ using few-shot exemplars to control answer quality and instruction following, where the target model $\mathcal{M}$ denotes the model which we want to fine-tune.

Critique Generation Inspired by control theory in sociology (Carver and Scheier, 1982), which posits that self-regulation and discrepancy-reducing feedback contribute to the development of social identity and cultural cognition, we propose a critique-based data generation framework for targeted cultural knowledge acquisition. However, recent studies (Huang et al., 2023; Kamoi et al., 2024) reveal that conventional critique generation methods often fail to provide insightful feedback for improving cultural knowledge. Moreover, Gou et al. (2023) demonstrates that simply using direct-generated critique can degrade model performance by corrupting correct responses. To address these challenges, we propose to decompose the golden answer A_g and target-aware answers A_t into atomic cultural knowledge units. This decomposition yields two knowledge units sequences: $A_g^u = [A_g^1, A_g^2, \cdots, A_g^n]$ and $A_t^u = [A_t^1, A_t^2, \cdots, A_t^m]$, where n and m denote the sequence lengths representing distinct cultural knowledge units:

$$A_g^u = \mathcal{G}(A_g) = [A_g^1, A_g^2, \cdots, A_g^n], \quad (1)$$

$$A_t^u = \mathcal{G}(A_t) = [A_t^1, A_t^2, \cdots, A_t^m], \quad (2)$$

where $\mathcal{G}$ denotes the generation model same above.

After obtaining the knowledge units A_g^u and A_t^u, we construct a fine-grained critique set T by comparing each ground truth knowledge unit $A_g^i \in A_g^u$ with corresponding knowledge units $A_t^j \in A_t^u$ in the target-aware answer. To ensure critique quality, we generate the meta-critique C_r by the data generation model $\mathcal{G}$, and categorize meta-critique C_r into three types:

(1) *Semantic Equivalence*: Indicates A_g^i has an exact semantic match in A_t^u, suggesting no further training is required for this cultural knowledge unit.

(2) *Unaddressed Knowledge*: Occurs when A_g^i lacks any corresponding unit in A_t^u, necessitating explicit pointing out this cultural knowledge in C_r.

(3) *Contradictory Statement*: Identifies cases where A_g^i conflicts with statements in A_t^u, requiring corrective meta-critique C_r to align the target model M with appropriate cultural norms.

This critique method allows the model to compare the golden answer A_g with target-aware answers A_t, thereby generating nuanced and reliable targeted critiques. These critiques will direct subsequent supervised fine-tuning to prioritize knowledge domains where the target LLM is prone to errors. Each critique instance $T_i \in T$ is represented

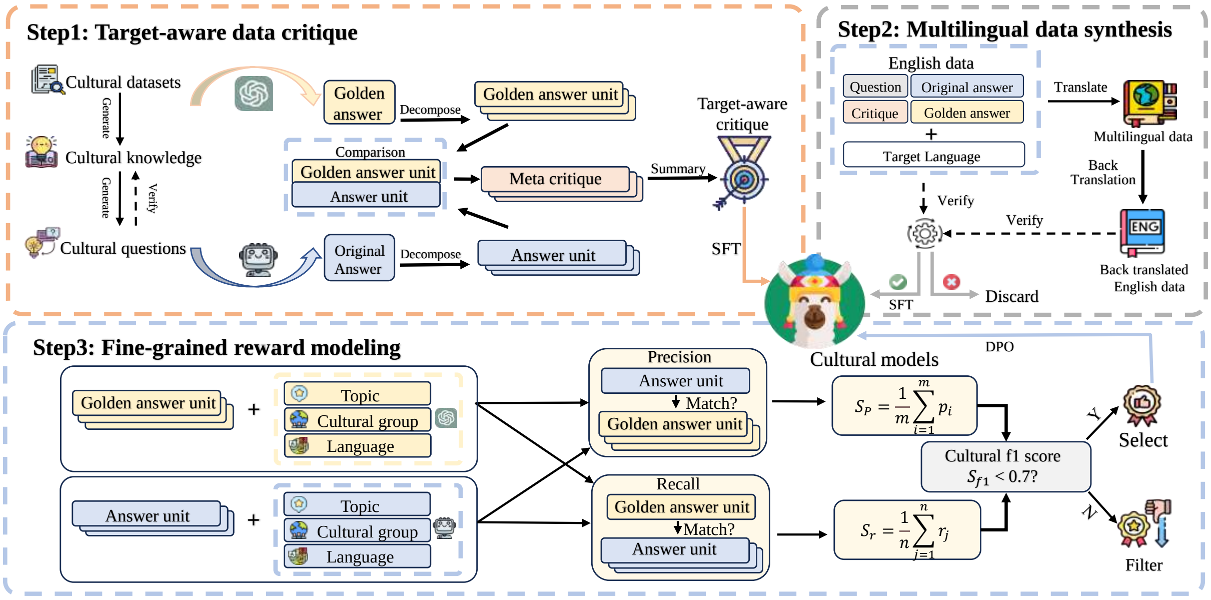


Figure 2: The overview of our proposed CulFiT.

as a triple:

$$T_i = \{A_g^i, A_t^j, C_r\}, \tag{3}$$

where C_r denotes the meta-critique described above. Finally, we summarize all meta critiques $(T_1, T_2, \cdots, T_k)$ for corresponding answer into a comprehensive critique C, and it will be used to serve as target-aware cultural error reminder in the supervised fine-tuning stage.

$$C = \text{LLM}(P, (T_1, T_2, \cdots, T_k)), \tag{4}$$

where P denotes the critique summary prompt and C denotes the final critique we obtain. All prompts can be seen in Appendix § 7.7

3.3 Multi-lingual Data Synthesis

In real-world scenarios, cultural-aware dialogue frequently occur in culturally-relevant languages (*e.g.,* Malay for Singaporean culture, Chinese for Chinese culture). To enhance model robustness in generating culturally appropriate knowledge across multilingual contexts, we propose a *Multi-lingual Data Synthesis* approach that generates answers in culturally relevant languages. After collecting critique-annotated cultural data $U = (Q, A_g, A_t, C)$, we first translate the data into target languages using our data generation model $\mathcal{G}$:

$$U_{target} = \mathcal{G}(U, L), \tag{5}$$

where U and U_{target} represent the source and target language cultural data, respectively, and L denotes the target language. To mitigate hallucination and ensure translation quality, we employ a back-translation verification mechanism. Specifically, we translate the target language text back to English using generation model $\mathcal{G}$:

$$U_{back} = \mathcal{G}(U_{target} \rightarrow U), \tag{6}$$

where U_{back} represents the back-translated text. We then perform semantic alignment between U_{back} and the original English text U, ensuring consistent semantic meaning of multilingual pairs.

3.4 Fine-grained Model Training

To train a cultural-aware model, we conduct a two-stage training method, which first uses supervised fine-tuning with target-aware multi-lingual critique data to equip the model with the ability to rectify areas prone to errors in the original answer and then leverage Direct Preference Optimization (DPO) (Rafailov et al., 2024) to further align the model. However, due to the challenge of intricacy and reliability in rewarding cultural-related texts, in this paper, we propose a *fine-grained cultural-aware reward modeling* approach that contains two sub-metrics: cultural precision and cultural recall to evaluate how culturally reliable the open-ended answer is.

Firstly, we enhance the evaluation framework by requiring the model to generate three additional contextual units for each question: (1) cultural group affiliation (A^c), (2) cultural topic (A^s), and (3) primary language(s) of the cultural group (A^l).

These units are appended to the original answer units, forming an extended answer representation:

$$A = [A^1, A^2, \cdots, A^k, A^c, A^s, A^l], \qquad (7)$$

where A^{1-k} denotes the atomic answer units obtained in 3.2. The inclusion of these contextual units serves two purposes. First, it captures the cultural contextual awareness, since as a culturally-aware model should accurately identify the background of the question. Second, these contextual units A^c, A^s, A^l provide precise, easily verifiable evaluation targets that reduce scoring variance due to their concise and factual nature.

3.4.1 Fine-grained Reward Modeling

During our training process, we begin by applying supervised fine-tuning that takes questions Q combined with original answer A_t and target-aware critique C as input and grounded answer A_g as output. To align the model with human cultural preferences, we also adopt Direct Preference Optimization. However, human preference is often subjective and context-dependent, which is hard to quantify in a reward function. To address the gap between the inherently subjective nature of cultural judgments and the objective metrics provided by standard reward functions, we design a fine-grained reward function that fully assesses the quality of cultural answers to select preference pairs automatically and robustly.

Cultural Precision Metric We introduce a *Cultural Precision Metric* $\mathbf{M_p}$ to evaluate the extent of cultural knowledge incorporation in model-generated answers. The intuition for this metric is that culturally-aware responses should precisely encompass relevant cultural knowledge. Following the methodology in Equation 1, we decompose both golden and target-aware answers into verifiable knowledge units, denoted as $A_g^u = [A_g^1, A_g^2, \cdots, A_g^n]$ and $A_t^u = [A_t^1, A_t^2, \cdots, A_t^m]$, respectively. The precision evaluation for each proposed answer unit $A_t^i \in A_t^u$ is formalized as:

$$p_i = \begin{cases} 1 & \text{if } \exists A_g^j \in A_g^u \text{ where } A_t^i \text{ matches } A_g^j, \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

The cultural precision S_p is then computed as:

$$S_p = \mathbf{M_p}(P) = \frac{1}{m}\sum_{i=1}^{m} p_i \qquad (9)$$

Cultural Recall Metric To complement the precision evaluation, we introduce a *Cultural Recall Metric* M_r that measures the coverage of golden answer knowledge units in the proposed answer. This metric is motivated by the principle that a comprehensive cultural response should encompass all relevant cultural knowledge points present in the golden answer, thereby achieving "culture-completeness". Following the same unit decomposition approach as in the precision task, we evaluate recall at the knowledge unit level through pairwise matching. The recall score for each golden answer unit $A_g^j \in A_g^u$ is computed as:

$$r_j = \begin{cases} 1 & \text{if } \exists A_t^i \in A_t^u \text{ where } A_g^j \text{ matches } A_t^i, \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

The cultural recall score S_r is then calculated as:

$$S_r = \mathbf{M_r}(R) = \frac{1}{n}\sum_{j=1}^{n} r_j \qquad (11)$$

Cultural F1 Metric To provide a comprehensive evaluation metric, we introduce the *Culture F1 Metric*, which combines precision and recall through their harmonic mean:

$$S_{f_1} = 2 \cdot \frac{S_p \cdot S_r}{S_p + S_r}. \qquad (12)$$

It is notable that we select our DPO training data using cultural F1 metric with $S_{f1} < 0.7$, which results in preference pairs maintaining both high-quality and relative larger learnable cultural gaps.

4 Experimental Setup

4.1 Datasets

We test our CulFiT and baselines on four datasets. GlobalCultureQA is our newly proposed multilingual benchmark that evaluates open-ended cultural knowledge question answering ability with 1104 questions, covering 400 specific topics and 23 languages. CANDLE500 (Nguyen et al., 2023) and CulturalBench (Chiu et al., 2024b) are multi-choice benchmarks that focus on evaluating cultural knowledge with 500 and 1224 samples. BLEnD (Myung et al., 2024) is a hand-crafted benchmark designed to evaluate LLM's cultural common knowledge across 16 countries and 13 different languages, comprising 52.6k question-answer pairs. Detailed distribution of topics and cultural groups across continents of GlobalCultureQA and examples are in Appendix § 7.2, § 7.5.

Model	Precision	Recall	F1
Close-source Models			
`4o`	72.34	**73.29**	72.81
`4o-mini`	72.89	72.47	72.68
Open-source Models			
`Mistral`	67.26	68.76	66.73
`SeaLLMs`	71.50	66.04	68.71
`Aya`	69.88	69.52	68.66
`Qwen2.5`	66.97	68.80	66.79
CulFiT (Qwen2.5)	72.16	67.56	68.81
-SFT	69.11	67.44	68.26
-DPO	70.95	66.57	67.60
`Llama3.1`	62.52	68.96	64.53
CulFiT (Llama3.1)	**74.73**	71.21	**72.94**
-SFT	71.33	70.55	70.94
-DPO	74.07	69.84	70.81

Table 1: Performance on our GlobalCultureQA.

Model	CANDLE500	CulturalBench
Close-source Models		
`4o`	91.2	84.1
`4o-mini`	87.0	82.1
Open-source Models		
`Mistral`	69.0	67.1
`Aya`	73.2	67.2
`SeaLLMs`	75.2	68.5
`CultureBank`	38.4	53.8
`Qwen2.5`	76.0	68.9
CulFiT (Qwen2.5)	79.6	72.9
-SFT	76.4	70.9
-DPO	78.2	71.0
`Llama3.1`	72.4	66.5
CulFiT (Llama3.1)	**81.2**	**73.1**
-SFT	78.6	69.1
-DPO	80.0	71.9

Table 2: Performance on CANDLE500 and CulturalBench.

4.2 Baselines

We employ several state-of-the-art LLM as baselines: close-source models including gpt-4o (`4o`) and gpt-4o-mini (`4o-mini`) (Hurst et al., 2024) and open-source models including Llama3.1-8B-Instruct (`Llama3.1`) (Dubey et al., 2024), Qwen2.5-7B-Instruct (`Qwen2.5`) (Yang et al., 2024), mistral-7B-Instruct-v0.3 (`Mistral`), aya-8B-expanse (`Aya`), SeaLLMs-v3-7B-Chat (`SeaLLMs`) and `CultureBank` (Shi et al., 2024). Training details can be found in § 7.1.

4.3 Evaluation Metric

For GlobalCultureQA benchmark, we evaluate cultural precision score S_p, cultural recall score S_r and then calculate cultural f1 score S_{f1} described in § 3.4.1. For CANDLE500 and CulturalBench, we report the precision of multi-choice questions. As for BLEnD, we first use corresponding lemmatizers and stemmers for model-generated answers and then compute the scores by marking whether the LLM's answer is included by the human annotator's answer.

5 Experimental Results

5.1 Overall Performance

GlobalCultureQA. In the open-ended question-answering task, as demonstrated in Table 1, CulFiT surpasses other open-source models and performs comparably to or even better than advanced closed-source models, achieving the highest precision score of 74.73 and cultural F1 score of 72.94 on GlobalCultureQA datasets. These results highlight the superior cultural awareness of our proposed CulFiT in addressing open-ended cultural knowledge questions, as well as its ability to effectively generate fine-grained cultural knowledge.

CANDLE500 and CulturalBench. In the cultural knowledge multiple-choice task, as shown in Table 2, CulFiT consistently outperforms base models, achieving improvements of up to 8.8% on CANDLE500 and 6.6% on CulturalBench, while also surpassing other open-source models by a large margin. However, it still lags behind SOTA models like `4o`, primarily due to differences in model scale. These results demonstrate that our CulFiT effectively enhances the model's cultural capability within the cultural knowledge domain.

BLEnD. When evaluating cultural understanding in local languages, as shown in Table 3, CulFiT outperforms open-source models such as `Aya` and `Mistral` in 12 out of 16 countries. Notably, the improvements are observed in low-resource language regions, such as Sundanese in West JAVA (increasing from 12.43 to 20.13) and Amharic in Ethiopia (increasing from 8.26 to 12.34). Additionally, an intriguing phenomenon emerges: the inherent cultural knowledge distribution within models is highly imbalanced. For example, `Qwen2.5` achieves a score of 60.33 on Chinese, while `Mistral` scores only 48.31. Similarly, `Aya` attains a score of 66.64 on Indonesian, whereas `Qwen2.5` scores 49.58. This phenomenon likely stems from

Models	**US**	**GB**	**CN**	**ES**	**MX**	**DZ**	**GR**	**KR**	**JB**	**IR**	**ID**	**AZ**	**KP**	**NG**	**AS**	**ET**
								Close-source Models								
`4o`	84.29	82.37	76.48	78.46	76.36	60.36	65.34	66.36	55.83	70.56	69.46	59.48	45.98	40.26	43.67	20.51
`4o-mini`	83.72	82.78	73.51	77.34	76.48	59.34	66.87	53.61	69.72	69.12	68.13	49.57	43.39	39.48	40.69	17.25
								Open-source Models								
`Mistral`	83.29	82.41	48.31	60.12	58.24	30.20	25.09	48.0	8.21	33.77	60.83	27.93	35.58	12.47	8.80	3.95
`SeaLLMs`	77.09	73.01	**66.97**	64.39	64.30	37.98	21.37	45.89	20.85	30.04	51.87	24.09	34.23	8.35	10.69	3.17
`Aya`	81.56	76.26	54.36	62.78	57.75	**47.95**	**47.33**	54.32	19.24	46.83	**66.64**	24.52	34.23	16.72	12.99	10.77
`Llama3.1`	82.46	76.48	56.54	61.62	64.09	40.73	41.52	50.94	12.43	48.46	58.75	39.87	36.26	20.42	15.09	8.26
CulFiT (Llama3.1)	**85.46**	**83.29**	57.38	**67.59**	**66.17**	45.76	42.17	49.16	20.13	**49.78**	64.87	**43.92**	**37.61**	21.03	**21.80**	**12.34**
`Qwen2.5`	78.52	72.52	60.33	64.60	58.24	38.15	21.05	52.19	21.17	36.18	49.58	38.59	25.67	24.60	19.25	11.41
CulFiT (Qwen2.5)	80.12	77.48	63.28	66.78	60.36	36.86	30.17	**56.42**	**24.21**	39.18	58.95	41.15	30.67	**25.68**	20.46	11.63

Table 3: Performance on BLEnD dataset. We use green to indicate that our CulFiT exceeds directly prompting the base LLM, and red shading indicates that they do not exceed the base LLM. We use ISO codes for each country, and the country and code mapping and experiment details can be seen in Appendix §7.4.

Model	**Precision**	**Recall**	**F1**
CulFiT	**74.73**	**71.21**	**72.94**
w/o Critique	69.54	67.63	68.57
w/o Multilingual	70.95	70.63	70.79

Table 4: Ablation study on GlobalCultureQA.

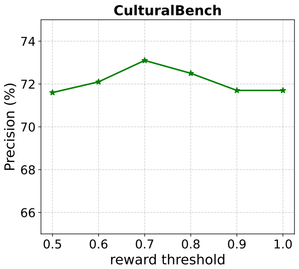


Figure 3: Results of the precision on different reward threshold S_{f1} from 0.5 to 1.0 with an interval of 0.1.

the differences in pre-training data across various models, highlighting that enhancing a model's cultural competence requires a careful consideration of its internal knowledge architecture and training data composition. In the contrast, CulFiT samples data uniformly across 100+ countries to ensure a balanced representation of diverse culture knowledge with multi-lingual augmentation, which effectively mitigate cultural biases in base models.

5.2 Ablation Study

We verify the effectiveness of our CulFiT by comparing it with two variant models: **(1) CulFiT w/o critique**: We remove model-generated answers and corresponding critiques, leaving only golden answers to train our model. **(2) CulFiT w/o multilingual**: We exclude the multi-lingual data synthesis stage in our method thus only using mono-English critique data. As shown in Table 4, these ablation models both achieve lower scores compared to CulFiT. Moreover, removing critique data performs worst in all metrics, which emphasizes the effectiveness of pointing out the weakness in cultural answers, and providing target-aware critique during fine-tuning is crucial for enhancing the cultural ability of models.

In Table 1 and Table 2, we also show the performance of the ablation models: *-DPO* which removes all fine-grained reward data in DPO and *-SFT* which deletes target-aware data in our training dataset. The performance of these models all decrease on three datasets, with a larger drop in *-SFT*, demonstrating the effectiveness of our training paradigm and the importance of high-quality target-aware SFT data.

5.3 Analysis of Reward Function

We analyze the impact of the reward function by varying the threshold S_{f1} (in Equation 12 S_{f1}) for selecting DPO data on the CulturalBench dataset. As illustrated in Figure 3, we observe that setting the threshold to 0.7 yields the best performance for our model, while incorporating higher-performing answers (*e.g.,* those rewarded with 0.9) degrades performance. A potential explanation for this phenomenon is that DPO benefits from preference pairs with larger differences, as pairs with small differences may hinder the model's ability to identify where errors are likely to occur. This finding further validates the effectiveness of our reward function in selecting lower-performing cultural answers, which enhances the model's learning process.

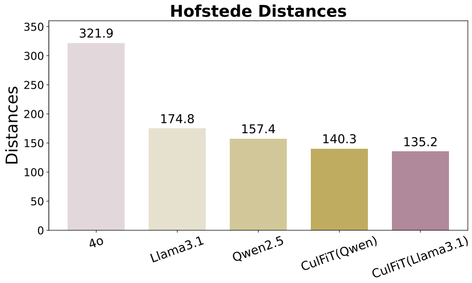


Figure 4: Comparison in terms of Hofstede distance.

Model	MMLU	MMMLU
`Llama3.1`	49.6	52.0
`Qwen2.5`	66.2	63.9
CulFiT (Llama3.1)	**50.1**	**62.5**
CulFiT (Qwen2.5)	**66.4**	**67.9**

Table 5: Precision scores on multi-lingual scenario.

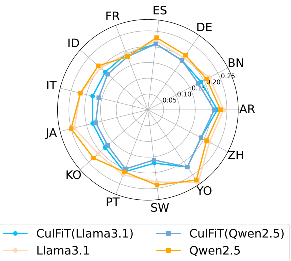


Figure 5: Results on multi-lingual inconsistent rate.

5.4 Analysis of Cultural Alignment

We further conduct a cultural alignment evaluation using Hofstede's cultural dimensions (Hofstede and Minkov, 2013), a well-established framework for quantifying cultural value differences across countries based on data collected from local residents. To assess the cultural alignment of LLMs, we prompt them to answer 24 questions from the VSM13 survey (Hofstede and Minkov, 2013), which measures local attitudes toward specific cultural questions. We then compute the Euclidean distance across six cultural dimensions between the LLM's responses and human responses. Details about Hofstede's cultural dimensions and experimental settings are provided in Appendix 7.4.

Figure 4 presents the results of cultural distances between `4o`, `Qwen2.5`, `Llama3.1`, and our CulFiT, which is based on `Qwen2.5` and `Llama3.1`. Our findings reveal two key insights: **(1)** Our CulFiT outperforms both `4o` and its base LLM, reducing the cultural distance for `Llama3.1` from 174.83 to 135.24 and for `Qwen2.5` from 157.41 to 140.32. This demonstrates that CulFiT achieves better cultural value alignment and exhibits superior cultural reasoning capabilities. **(2)** Fundamental model abilities, such as math and coding, do not correlate with cultural alignment performance. While SOTA LLMs like `4o` excel in fundamental tasks, they underperform in cultural value alignment compared to smaller models. This discrepancy may stem from the unbalanced cultural knowledge in the training data of SOTA LLM, which can skew their value systems. These results highlight the importance of reducing cultural bias and developing models that ensure equitable cultural representation.

5.5 Analysis of Multilingual Data

To investigate the effectiveness of our CulFiT in multilingual settings, we conduct experiments on the MMLU (Hendrycks et al., 2020) and MMMLU (Wang et al., 2024) dataset which translates MMLU into 14 languages. We randomly select 150 English questions related to cultural domains such as world religions, human sexuality, and sociology. Table 5 reports the precision of our CulFiT and its base LLM `Llama3.1`, where our model outperforms `Llama3.1` by a large margin (10.5% in MMMLU) when answering the same cultural questions while maintaining comparable or even superior performance in MMLU.

To validate the robustness of our CulFiT in the multilingual scenario, we count the inconsistent responses between two models on two datasets (*a.k.a.,* MMLU and MMMLU) and group the results by language. Figure 5 illustrates that our model exhibits a lower inconsistent error rate than base models across all 14 languages, demonstrating excellent robustness.

5.6 Discussion of General Capability

To evaluate the generalization ability of our method and mitigate the risk of catastrophic forgetting, we conduct experiments on commonsense and reasoning datasets, including CSQA (Talmor et al., 2018), Hellaswag (Zellers et al., 2019), and MMLU-pro (Wang et al., 2024). As shown in Table 6,

Model	CSQA	Hellaswag	MMLU-pro
`Llama3.1`	70.1	71.5	36.8
CulFiT (Llama3.1)	73.1(+3.0)	72.7(+1.2)	38.7(+1.9)
`Qwen2.5`	80.3	75.9	47.0
CulFiT (Qwen2.5)	**80.6(+0.3)**	**77.5(+1.6)**	**47.1(+0.1)**

Table 6: Comparison of general abilities of between CulFiT and base LLM.

our models consistently outperform the original LLM across all three tasks, demonstrating that integrating culture-related knowledge by using our proposed CulFiT not only enhances culture-related knowledge but also improves general reasoning capabilities and prevents catastrophic forgetting.

6 Conclusion

In this paper, we introduced Target-aware **Cul**tural Data Synthesis and **Fi**ne-grained **T**raining (CulFiT), a novel cultural-aware training paradigm that addresses cultural bias in large language models (LLMs) through target-aware multilingual data synthesis and fine-grained reward modeling. Our approach enhances cultural sensitivity and robustness across diverse linguistic and cultural contexts. Experiments on our newly proposed GlobalCultureQA benchmark and three cultural knowledge benchmarks show that CulFiT outperforms existing open-source models and competes with state-of-the-art closed-source models. Analysis using Hofstede's cultural dimensions reveals that CulFiT achieves better cultural value alignment than base models and advanced LLMs like GPT-4o.

Limitations

While CulFiT shows significant improvements, it still faces challenges in fully capturing the fine-grained cultural knowledge of low-resource languages due to limited training data. Another minor limitation is the computational cost associated with generating and processing multilingual critique data, which could be a bottleneck for smaller research teams.

Ethical Considerations

Despite ongoing efforts to reduce cultural bias, large language models (LLMs) can still unintentionally reinforce stereotypes or present inaccurate portrayals of certain cultures. This often stems from biases embedded in the data they are trained on, which may reflect dominant cultural narratives or historical inequalities. As a result, the outputs generated by these models may marginalize underrepresented voices or misrepresent diverse communities. Addressing these issues is a critical ethical responsibility to ensure fairness, inclusivity, and respectful representation in AI systems.

Acknowlegements

This work was supported by the National Key R&D Program of China 2024YFE0111800, the National Natural Science Foundation of China (62032001, 62432002, 62406061, and T2293773), and the Natural Science Foundation of Shandong Province (ZR2023QF159).

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Charles S Carver and Michael F Scheier. 1982. Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin*, 92(1):111.

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

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*.

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

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

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*.

Geert Hofstede and Michael Minkov. 2013. Vsm 2013. *Values survey module*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224.

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*.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024c. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina 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. *arXiv preprint arXiv:2406.09948*.

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)*.

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.

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*.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv preprint arXiv:2402.17302*.

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*.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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*.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations. *arXiv preprint arXiv:2403.14112*.

World Values Survey. 2022. World values survey. https://www.worldvaluessurvey.org/wvs.jsp.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

7 Appendix

7.1 Training and Implementation Details

In the supervised fine-tuning(SFT) stage, we use CANDLE and CultureAtlas as seed data to train LLMs and in direct preference optimization(DPO) stage we adopt CultureBank as source data. We arrange the data in a dataflow of Question->Original Answer->Critique->Golden answer, with Golden answer as output and others as output. We finally get 25344 QA pairs with critique in English and 20,140 pairs in other languages, spanning 3026 topics and 24 languages in supervised-finetuning stage and 16334 preference pairs in Direct preference Optimization stage.

We train our models on 8 NVIDIA L40s and train the model for 1000 steps of batch size 16 on every stage. We select a learning rate of 1e-5 in SFT stage and 5e-6 in DPO stage with a warmup ratio of 0.1. For parameter efficiency, all training process use LoRA with a rank of 16.

We use greedy decoding for multi-choice questions and tak temperature of 0.7 for other tasks. We use one-shot prompting on our GlobalCultureQA dataset during inference and adopt zero-shot prompting strategy for other tasks.

7.2 Details on GlobalCulture QA

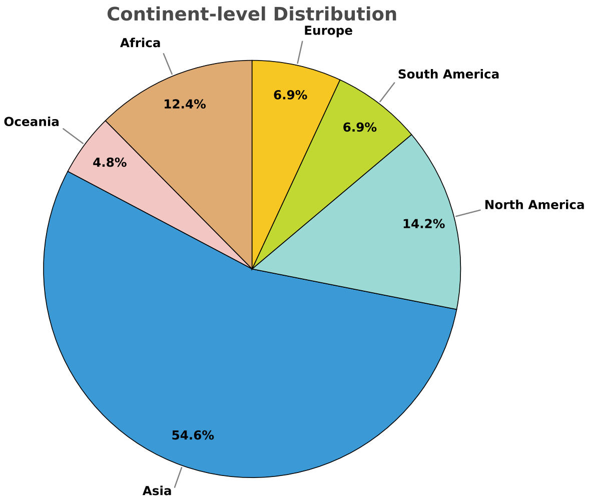


Figure 6: The continent level distribution on our GlobalCultureQA

We provide an example of our benchmark's question:

How do the enforcement and interpretation of hijab regulations differ between regions such as Gaza, Iran, Indonesia, and Saudi Arabia, and what social tensions do these differences reflect?

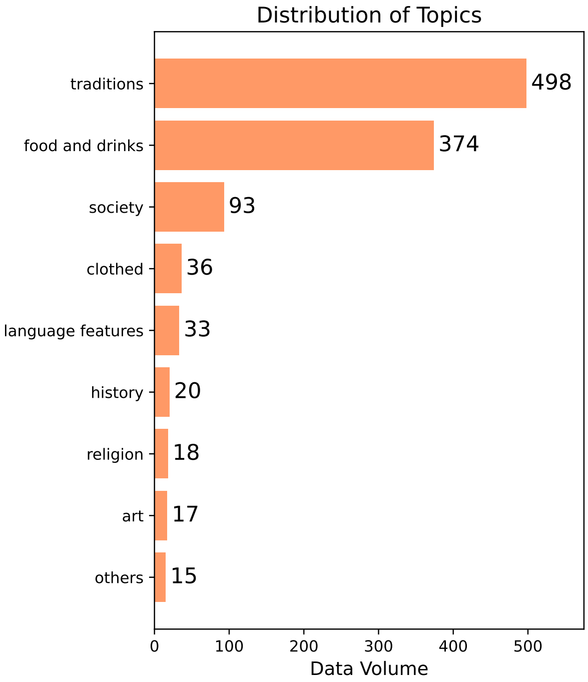


Figure 7: Distribution of overall topics. We classify the specific topics into 9 categories and report the volume of each category

and we provide an example of the corresponding golden answer:

The enforcement and interpretation of hijab regulations vary significantly across regions such as Gaza, Iran, Indonesia, and Saudi Arabia, reflecting complex social tensions. In Gaza, members of the Palestinian jihadist group Unified Leadership have rejected policies mandating the hijab, even targeting those who attempt to enforce such rules. In Iran, the shift from a ban on veils in 1936 to mandatory Islamic dress following the 1979 Islamic Revolution established a stringent legal framework, leading to severe penalties for women not adhering to hijab requirements. This has resulted in fluctuating enforcement practices, with women often challenging these restrictions. In Indonesia, particularly in the province of Aceh, local Sharia law mandates that Muslim women wear hijab in public, showcasing a more localized interpretation of dress codes. In Saudi Arabia, while the government requires women to cover their hair and wear full-body garments, enforcement has been inconsistent, leading to criticism of the religious police for their actions, notably hindering the rescue of schoolgirls in 2002 due to their attire. Overall, these differing practices highlight ongoing tensions and varied interpretations of religious and cultural norms across societies, reflecting broader social dynamics and conflicts over gender, identity, and religious expression.

7.3 Details On BLEnD

When conducting experiments on BLEnD, we adpot a system prompt `'You are a helpful {country} AI chatbot that know the culture of {country} very well. You task is to answer the question about {country} in {language}.'`

We also choose the results of 'pers-3' prompt described in the original paper: `'You are a person from {country} who is trying to explain your country's culture to a foreigner. Answer the following question, providing a single answer without any explanations.'`

Table 7 shows the mapping of country and ISO code, with the corresponding answers of each country.

Country/Region	Code	Language
United States United Kingdom	US GB	English
China	CN	Chinese
Spain Mexico	ES MX	Spanish
Indonesia	ID	Indonesian
South Korea North Korea	KR KP	Korean
Greece	GR	Greek
Iran	IR	Persian
Algeria	DZ	Arabic
Azerbaijan	AZ	Azerbaijani
West Java	JB	Sundanese
Assam	AS	Assamese
Northern Nigeria	NG	Hausa
Ethiopia	ET	Amharic

Table 7: The details of country and ISO code mapping with their corresponding languages

7.4 Details On Hofstede Cultural Dimentions

This survey identified six dimensions of national culture: Power Distance Index (PDI), Individualism vs. Collectivism (IDV), Masculinity vs. Femininity (MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation vs. Short-Term Orientation (LTO) and Indulgence vs. Restraint (IND). VSM 2013 is an authoritative and famous cultural questionnaire devised by Hofstede and is widely used. In this experiment, we evaluate the cultural alignment of our models on 9 cultures(Arabic, Bangladesh, Chinese Germany, Korean, Portuguese, Brazil, Argetina and Turkish) and calculate the average distances of all countries. To be specific, the VSM 2013 have 24 questions in total. The computation of six cultural dimensions is based on the following formulas:

$$PDI = 35(\mu_{Q7} - \mu_{Q2}) + 25(\mu_{Q20} - \mu_{Q23}) + C_{PDI} \tag{13}$$

$$IDV = 35(\mu_{Q4} - \mu_{Q1}) + 35(\mu_{Q9} - \mu_{Q6}) + C_{IDV} \tag{14}$$

$$MAS = 35(\mu_{Q5} - \mu_{Q3}) + 25(\mu_{Q8} - \mu_{Q10}) + C_{MAS} \tag{15}$$

$$UAI = 40(\mu_{Q18} - \mu_{Q15}) + 25(\mu_{Q21} - \mu_{Q24}) + C_{UAI} \tag{16}$$

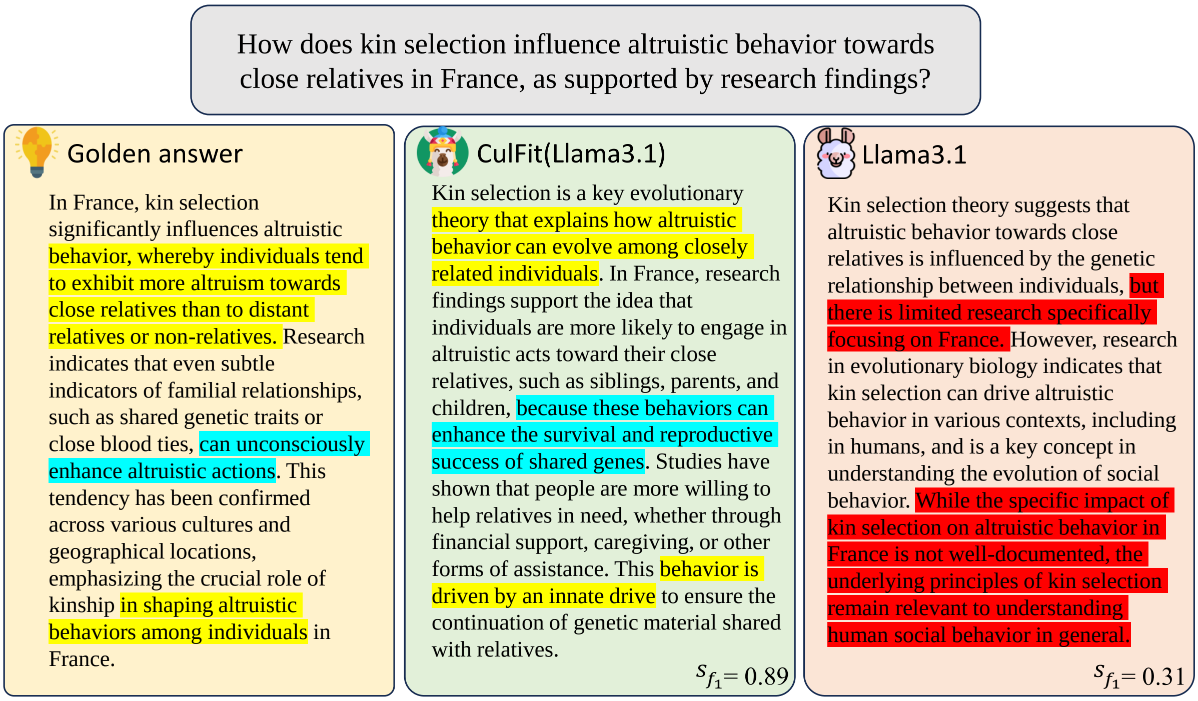
$$LTO = 40(\mu_{Q13} - \mu_{Q14}) + 25(\mu_{Q19} - \mu_{Q22}) + C_{LTO} \tag{17}$$

$$IVR = 35(\mu_{Q12} - \mu_{Q11}) + 40(\mu_{Q17} - \mu_{Q16}) + C_{IVR} \tag{18}$$

where μ means the average of all the answers to each question. C is constants that can be used to adjust to scores to fit a range between 0 and 100 or anchor new data to Hofstede's old dataset (Hofstede and Minkov, 2013). During experiment, we convert the questions into the multi-choice format, with a `'You are a {culture} chatbot that know {culture} very well. Now your task is to represent the people in culture and answer the following question. Please be sure that you should only consider the culture of {culture} when answering the question.'` system prompt.

We get the Euclidean distance of the gaps from six cultural dimensions as the metric $Distance$. Larger $Distance$ means weak cultural understanding ability of LLMs, and vice versa.

$$Distance = \sqrt{\sum (d_{model} - d_{hofstede})^2}, \tag{19}$$



How does kin selection influence altruistic behavior towards close relatives in France, as supported by research findings?

Golden answer

In France, kin selection significantly influences altruistic behavior, whereby individuals tend to exhibit more altruism towards close relatives than to distant relatives or non-relatives. Research indicates that even subtle indicators of familial relationships, such as shared genetic traits or close blood ties, can unconsciously enhance altruistic actions. This tendency has been confirmed across various cultures and geographical locations, emphasizing the crucial role of kinship in shaping altruistic behaviors among individuals in France.

CulFit(Llama3.1)

Kin selection is a key evolutionary theory that explains how altruistic behavior can evolve among closely related individuals. In France, research findings support the idea that individuals are more likely to engage in altruistic acts toward their close relatives, such as siblings, parents, and children, because these behaviors can enhance the survival and reproductive success of shared genes. Studies have shown that people are more willing to help relatives in need, whether through financial support, caregiving, or other forms of assistance. This behavior is driven by an innate drive to ensure the continuation of genetic material shared with relatives.

s_{f_1}= 0.89

Llama3.1

Kin selection theory suggests that altruistic behavior towards close relatives is influenced by the genetic relationship between individuals, but there is limited research specifically focusing on France. However, research in evolutionary biology indicates that kin selection can drive altruistic behavior in various contexts, including in humans, and is a key concept in understanding the evolution of social behavior. While the specific impact of kin selection on altruistic behavior in France is not well-documented, the underlying principles of kin selection remain relevant to understanding human social behavior in general.

s_{f_1}= 0.31

Figure 8: A case study on the results of GlobalCultureQA. We use yellow to indicate the the part that corresponds with golden answer, blue to show the extensive content compared to golden answer and red to highlight vague and uncultural answers.

7.5 Case Study

As shown in Figure 8, we compare the answer of our proposed CulFiT and `Llama3.1`. We use yellow to indicate the part that corresponds with golden answer, and We use blue to highlight content that is more extensive compared to the golden answer. Red highlights indicate responses that are vague compared to the golden answer and fail to provide a corresponding answer. In our CulFiT's answer, we have parts that precisely reflect the golden answer(`theory that explains how altruistic behavior can evolve among closely related individuals`) and have contents that extend the cultural knowledge to a more nuanced extent(`because these behaviors can enhance the survival and reproductive success of shared genes`). On the contrast, the original `Llama3.1` just gives vague and incorrect answers like `but there is limited research specifically focusing on France.`, which hinders the cultural nuances in these sentences. We attribute this to the target-aware data training because we force model to capture the 'target' in the question and thus avoid generating bad answers like `While the specific impact of kin selection on altruistic behavior in France is not well-documented`. Additionally, our CulFiT get a cultural f1 score S_{f1} of 0.89, while `Llama3.1` only obtain 0.31, which is correlated with the analysis above, demonstrating the stability and fairness of our evaluation metric.

Model	Win	Tie	Lose
CulFiT (Llama3.1) vs `Llama3.1`	38	46	16
CulFiT (Llama3.1) vs `gpt-4o`	25	50	24

Table 8: Human study of our model vs Llama3.1 and gpt-4o

7.6 Human Evaluation

We performed a model comparison study involving 100 randomly sampled questions. Four native speakers (1 PhD and 3 Master's students proficient in both Chinese and English) participated in this single-blind evaluation. The annotators achieved an impressive inter-annotator agreement (Kappa score = 0.86), demonstrating high consistency in assessing cultural awareness, readability, and question relevance. Results in Table8 show our CulFit (Llama3.1) model significantly outperforms the base model, aligning better with human preferences across all evaluation dimensions.

7.7 Prompts For Data Synthesis

The prompt for cultural question generation based on cultural knowledge:

You are a helpful expert in generating cultural-aware quetions through cultural knowledge. You are privided with a piece of cultural knowledge and the background of the cultural knowledge. Your task is to generate a single question based on the cultural knowledge that is given to you. The input form is encoded as JSON format, and below is its JSON fields: {"cultural_group": "", "topic": "", "source": "", "cultural_knowledge": ""}
the detailed explanation of the fields are as follows:
-cultural_group: the country or the cultural group where the cultural knowledge is from
-topic: the topic of the cultural knowledge
-source: the source of the cultural knowledge
-cultural_knowledge: the cultural knowledge that is provided to you, which should pay most attention

Please strictly follow the following rules: 1. Factuality: Your question should only stems from the cultural knowledge that is provided to you and you shouldn't add other knowledge to your generated question. 2. Specificity: Your question should cover the main idea of the cultural knowledge and should be comprehensive, but not too broad. Try to specific the question with the cultural knowledge and do not ask too general questions. 3. Coverage: You should carefully understand the cultural knowledge and extract the cultural knowledge points as much as possible. And use these cultural knowledge ponints to formulate your question.

The prompt for answer generation process:

You are a helpful consultant for a cultural knowledge question answering scenario. You are given the following question and its cultural knowledge. Your task is to generate a culturally-aware answer to the question based on the cultural knowledge. Remember, your answer should be encoded in JSON format. The detailed explanation of the fields is as follows:
`{"answer": "", "cultural_group": "", "language": "", "topic": ""}`
answer: your answer to the question
cultural_group: the country or the cultural group your answer points to
language: the language that the cultural group mainly speaks
topic: the main topic of your answer

Notably, the question stems from the cultural knowledge, so your answer should also be based on the provided cultural knowledge. You should always follow the instructions and directly answer the questions that are provided to you.
<example_start>
...
<example_end>
Remember, your answer should correlate with the cultural knowledge . You should only return the answer.
Your Answer:

The prompt for target-aware critique generation:

You are an expert reviewer for a cultural knowledge question answering system. You have plenty of cultural knowledge in {}.
You are given a JSON object and the detailed explanation of the fields are as follows:
{"question":"","grounded_answer":"",
"answer_to_critique":"",
"grounded_answer_knowledge_points":"",
“knowledge_points_to_critique”: ""}
-question: the cultural question that is given to you
-grounded_answer: the grounded answer to the question, which is the reference answer
-answer_to_critique: the answer that you should critique
-grounded_answer_knowledge_points: the knowledge points extracted from the grounded answer, each knowledge point is a single sentence and is seperated with a comma in a list
-knowledge_points_to_critique: the knowledge points extracted from the answer_to_critique, each knowledge point is a single sentence and is seperated with a comma in a list
You should compare the grounded_answer_knowledge_points and the answer_knowledge_points and provide a detailed critique based on the comparison. And your critique should based on the principles below:
1. Correctness: Be sure to point out any factual inaccuracies or errors in the answer_to_critique and provide corrections based on the

grounded_answer_knowledge_points. 2. Comprehensiveness: The answer_to_critique should cover the main points of the grounded_answer and should not miss any key information, if the answer_to_critique miss the cultural knowledge points, you should say "not addressed clearly" between the comparison. 3. Stability: If the grounded_answer_knowledge_points and the knowledge_points_to_critique are mainly the same, you should say "Roughly the same" in your critique. 4. Point by point: You should compare the grounded_answer_knowledge_points and knowledge_points_to_critique point by point and provide your critique based on the comparison. Between the comparison, you should choose the most relevant knowledge_points_to_critique from the list while comparing the grounded_answer_knowledge_points. You should always follow the instructions and carefully compare the grounded_answer_knowledge_points and the answer_knowledge_points point by point and provide your critique. Remember, you should directly compare the grounded_knowledge_points and knowledge_points_to_critique and point out the flaws made by knowledge_points_to_critique. And your critique should be encoded in JSON format, with each unit be a small JSON format , which will be like:
{
{
"grounded_answer_knowledge_points": "",
"knowledge_points_to_critique":""
"Critique": ""
},
{
"grounded_answer_knowledge_points": "",
"knowledge_points_to_critique": ""
"Critique": ""
},
.....
}
You answer:

7.8 Examples Of Critique And Answer

We also elaborate an example of meta critique:

{
"grounded_answer_knowledge_points": "In Chinese culture, eggs symbolize fertility and new life.",
"knowledge_points_to_critique": "In Chinese culture, eggs symbolize fertility and new life due to their ability to hatch.",
"Critique": "Roughly the same"
},
{
"grounded_answer_knowledge_points": "Eggs represent blessings and auspiciousness in Chinese culture.",
"knowledge_points_to_critique": "Not addressed clearly.",
"Critique": "The answer_to_critique does not mention the aspect of blessings and auspiciousness associated with eggs, which is an important cultural significance. Including this detail would enhance the comprehensiveness of the response."
},
{
"grounded_answer_knowledge_points": "Eggs play an integral role in many aspects of Chinese life.",
"knowledge_points_to_critique": "Contradictory.",
"Critique": "The answer_to_critique does not encapsulate the integral role that eggs play in various aspects of Chinese life, which is a key point. Acknowledging this would provide a holistic view of the cultural significance of eggs."
}

We also elaborate an example of the Critique C:

The critique identifies several key aspects of Chinese cultural significance related to eggs that are either missing or inadequately addressed in the answers. While some points align closely, many important cultural dimensions are overlooked. Specifically, the aspects of blessings and auspiciousness associated with eggs, the gifting of eggs during Chinese New Year, and their involvement in rituals and culinary practices are not clearly articulated. Additionally, the use of eggs in fortune-telling, their culinary significance in traditional dishes like mooncakes, and the symbolic representation of the yolk are omitted. Unique customs, such as boiling eggs in urine and using eggshells for folk art, as well as the broader role of eggs in Chinese life, also lack sufficient discussion. Including these details would enhance the depth and richness of the response, providing a more comprehensive understanding of the cultural significance of eggs in Chinese traditions.

Additionally, here is an example of our question:

How does the integration of high-quality materials, such as English wool blends, and traditional Chinese techniques in the production of hand-tufted rugs like Whisper reflect the rich textile heritage of China?

And here is an example of the corresponding answer:

The integration of high-quality materials, such as English wool blends, and traditional Chinese techniques in the production of hand-tufted rugs like Whisper reflects the rich textile heritage of China by showcasing the country's exceptional craftsmanship and adaptability. The use of fine English wool enhances the quality of the rugs, while traditional Chinese techniques ensure that the artistry and cultural significance of the textile production are preserved. This blending of premium materials and age-old methods illustrates how China not only honors its textile heritage but also embraces global influences to create products that stand out in the market, reflecting both tradition and innovation."

We also exhibt the example of answer unit:

{
"knowledge_points":
"The production of hand-tufted rugs like Whisper integrates high-quality materials such as English wool blends with traditional Chinese techniques.",
"This integration reflects the rich textile heritage of China.",
"The combination of modern materials and ancient techniques showcases the mastery of Chinese artisans.",
"The use of English wool blends contributes softness, durability, and stain resistance to the rugs.",
"Traditional Chinese techniques like hand-tufting and natural dyeing maintain the rugs' cultural and aesthetic value.",
"The fusion of old and new in rug production demonstrates China's long history of textile innovation.",
"Chinese textile production adapts to changing times while remaining true to cultural roots."
}

7.9 Evaluation Example

We first display the prompt for our fine-grained evaluation process:

You are an expert evaluator for a cultural knowledge question answering system. You are given a piece of cultural knowledge point and a list of reference cultural knowledge. Your task is to evaluate whether the given cultural knowledge point satisfies one of the reference cultural knowledge points and give a concise explanation.
Here are some examples and explanations:
</example>
<example/>

Remember, Your output should first generate 'Yes' or 'No', and give a concise explanation of your evaluation.
If your answer is "Yes", your explanation should specifically incorporate the given cultural knowledge point satisfies which reference cultural knowledge point.
cultural knowledge points:
{ }
reference cultural knowledge points:
{ }
Your output:

We then display the 'Yes' case of the evaluation process with explanation:

cultural knowledge points:
"The centers aim to improve literacy rates among Afghan citizens, particularly women and children."

reference cultural knowledge points:
"Lincoln learning centers in Afghanistan improve literacy rates among Afghan citizens.",
"These centers were established in response to low literacy rates in Afghanistan.",
"Lincoln learning centers serve as educational hubs providing English language classes, library facilities, Internet connectivity, and counseling services.",
"The initiative aims to reach at least 4,000 Afghan citizens each month at each location.",
"Literacy courses are mandatory for the military and national police forces in Afghanistan.",
"The initiative reflects a broader commitment to enhancing literacy levels across Afghanistan.",

"Educational programs at the centers promote an understanding of American culture.",
"The primary languages spoken in the Lincoln learning centers are Dari and Pashto."

Your output:
Yes

explanation: the cultural knowledge point "The centers aim to improve literacy rates among Afghan citizens, particularly women and children." is similar to the reference cultural knowledge point "These centers were established in response to low literacy rates in Afghanistan.", so the output is Yes.

And a 'No' case for the evaluation with explanation:

cultural knowledge points:
"Basketball is gaining popularity in Afghanistan and is enjoyed by both men and women."

reference cultural knowledge points:
"The Afghan Sports Federation was established in 1922.",
"The Afghan Sports Federation promotes sports like football and basketball in Afghanistan.",
"The federation is responsible for developing, organizing, and overseeing various sports in Afghanistan.",
"Afghanistan's national football team qualified for the 2014 FIFA World Cup.",
"The qualification for the 2014 FIFA World Cup was a significant milestone for Afghan football.",
"The Afghan Sports Federation faces challenges such as financial constraints and infrastructure limitations.",
"Ongoing conflicts in Afghanistan have impacted the development of sports.",
"The history of the national football team reflects Afghanistan's turbulent past.",
"Many players and coaches of the national football team have fled Afghanistan due to conflict or persecution.",
"The Taliban banned sports during their rule from 1996 to 2001.",
"The ban on sports during Taliban rule hindered the progress of the national football team.",
"The national football team has shown resilience despite numerous challenges.",
"Players like Zohib Islam Amiri and Faisal Hamidi have represented Afghanistan in international competitions.", "The 2021 Taliban takeover has raised concerns about the future of sports in Afghanistan."

Your output:
No

explanation: the cultural knowledge point "Basketball is gaining popularity in Afghanistan and is enjoyed by both men and women" is not addressed clear in the reference cultural knowledge points, so the output is No.