CURE: Cultural Value-based Reasoning for Enhancing the Cultural Adaptability of Large Language Models

Mirae Han and **Harksoo Kim***
Konkuk University
Seoul, Republic of Korea
{future26,nlpdrkim}@konkuk.ac.kr

Abstract

While large language models (LLMs) are widely used across cultures, they often generate culturally inappropriate responses in unfamiliar cultural contexts due to biases embedded in their training data. Existing approaches primarily rely on expanding static cultural knowledge, which fails to capture the inherently relative and context-dependent nature of culture. In this paper, we propose a **Cu**ltural value-based **Re**asoning (**CURE**) framework that interprets behaviors through underlying cultural value systems. In addition, we integrate CURE into LLMs via Chain-of-Thought (CoT) distillation, referred to as **CURE-distillation**, to internalize culturally grounded reasoning. Experimental results show that models trained with CURE-distillation improve cultural adaptability, enabling them to produce culturally aligned ethical judgments across diverse cultural scenarios. These results suggest that strengthening sociocultural reasoning capabilities can substantially improve the cultural adaptability of LLMs. The code is available at https://github.com/KUNLP/CURE.

1 Introduction

Large language models (LLMs) are increasingly used by users across diverse cultural contexts. However, LLMs are biased toward cultures that are more prevalent in their training data, leading to culturally inappropriate or even offensive responses in underrepresented contexts (Johnson et al., 2022; Yu et al., 2025; Venkit et al., 2023; Deshpande et al., 2023; Sheng et al., 2021).

This challenge arises because the social acceptability and meaning of specific expressions or behaviors vary across cultures (Ricoeur, 1973). For example, when asked *"How should one greet a teacher?"*, an LLM lacking cultural adaptability might respond, *"Wave your hand and greet them cheerfully."* While this may be appropriate in Western contexts, it is likely to be perceived as rude in several Asian cultures, where bowing is the customary form of greeting toward superiors or elders. This example illustrates that LLM responses require alignment with the user's cultural context rather than mere plausibility, underscoring the need to improve cultural adaptability.

Cultural adaptability is essential for real-world scenarios regardless of model scale. In resource-constrained environments, the use of small models is often unavoidable, yet cultural adaptability remains critical. While prompting strategies may offer a potential solution, their reliance on inference-time constraints increases computational overhead and sensitivity to prompt design. These limitations motivate approaches that embed cultural adaptability directly into model parameters.

Previous studies have primarily sought to improve this ability by training models to learn cultural ethical norms (Pham et al., 2024; Rao et al., 2025; Shi et al., 2024; Dwivedi et al., 2023, 2025). In these studies, *culture* is conceptualized as observable expressions of values that members of a particular group consider important, as reflected in language and behavior. Accordingly, culture is represented through semantic proxies such as social etiquette and norm (Adilazuarda et al., 2024). However, cultural ethics are highly context-dependent, with standards of judgment varying across cultures and situations (Benedict, 2019; Tilley, 2000; Österman, 2021). This relativity indicates that treating cultural ethics as static knowledge results in superficial cultural understanding and limited contextual generalization.

According to cultural anthropology, culture is not merely a collection of isolated facts but an interpretive system shaped by a society's underlying

* Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2145–2162, 2026. https://doi.org/10.1162/TACL.a.795
Action Editor: Zeljko Agic. Submission batch: 10/2025; Revision batch: 3/2026; Published 8/2026.

values (Geertz, 2008; Schwartz, 1992; Hills, 2002; Roccas and Sagiv, 2010). Cultural differences arise from how these value systems are structured and prioritized, and the meaning of a behavior is defined not by its surface expression but by the cultural values that give it significance. Therefore, accurately interpreting behavior requires reasoning over a culture's value system.

Inspired by this, we propose a **Cu**ltural value-based **Re**asoning (**CURE**), a framework grounded in cultural anthropology that formalizes how behaviors are interpreted through cultural value systems. CURE progressively integrates information at each reasoning step, allowing earlier inferences to be refined or reinterpreted in subsequent steps. This design enables nuanced cultural reasoning in highly context-dependent environments. To internalize this reasoning pattern, we distill teacher-generated rationales into student models through **CURE-distillation**. Extensive experiments demonstrate the logical coherence of CURE and show that CURE-distillation improves cultural adaptability compared to existing methods. These results suggest that equipping models with cultural reasoning capabilities can substantially enhance the cultural adaptability. The main contributions of our paper are as follows:

- We propose CURE, a cultural value-based reasoning framework for effective reasoning in culturally relative contexts.
- We introduce CURE-distillation, a method that transfers cultural reasoning abilities to student models, thereby improving cultural adaptability.
- We demonstrate that the proposed method improves LLM performance on cultural ethics tasks, indicating that strengthening cultural reasoning capabilities significantly enhances cultural adaptability.

2 Related Work

2.1 Cultural Bias in LLMs

Previous studies have consistently shown that LLMs exhibit a strong bias toward Western cultures. Naous et al. (2024) and Wang et al. (2024) analyzed cases in which, despite being provided with explicit cultural cues, LLMs failed to reflect the target cultural context and generated responses based on Western perspectives. Durmus et al. (2023) and AlKhamissi et al. (2024) quantitatively compared LLM outputs with human survey results across different countries, revealing a close alignment between LLMs and Western value systems. In addition, recent benchmarks evaluating cultural adaptability show that current models struggle to produce responses aligned with the ethical standards of diverse cultures (Shi et al., 2024; Pham et al., 2024; Dwivedi et al., 2023; Rao et al., 2025).

To mitigate cultural biases, many studies have pretrained culturally specialized models on large-scale corpora representing specific cultures (Chan et al., 2024; Nguyen et al., 2024; Pipatanakul et al., 2023; Abbasi et al., 2023; Lin and Chen, 2023). Other studies have constructed datasets reflecting cultural norms for fine-tuning (Shi et al., 2024; Pham et al., 2024; Dwivedi et al., 2023, 2025). However, these approaches focus on the quantitative expansion of cultural knowledge. We therefore propose a training method that enables models to internalize reasoning mechanisms effective for interpreting cultural contexts.

2.2 CoT Distillation

CoT distillation has been proposed to transfer reasoning capabilities from a teacher to a student model. However, Chen et al. (2025) pointed out that, due to the student model's limited representational capacity, rationales may introduce excessive information that hinders learning. Wadhwa et al. (2024) and Feng et al. (2024) further suggested that the effectiveness of CoT distillation may not reflect improved reasoning ability, but rather arise from enriched label representations or overfitting to spurious input-label correlations.

Accordingly, Xi et al. (2024) and Shridhar et al. (2023) decomposed problems into simpler subproblems or separated reasoning steps into objective and interpretive information, enabling models to learn the structural characteristics of rationales. Ho et al. (2023) generated diverse forms of rationales to reduce overreliance on fixed reasoning paths, while Feng et al. (2024) used counterfactual examples to encourage reasoning from multiple perspectives and improve generalizability. In addition, Dai et al. (2024) modified the training sequence of rationales and labels to prevent models from learning superficial input-label correlations. However, the effectiveness of these methods largely depends on the teacher model's inherent knowledge. Therefore, we propose a CoT distil-

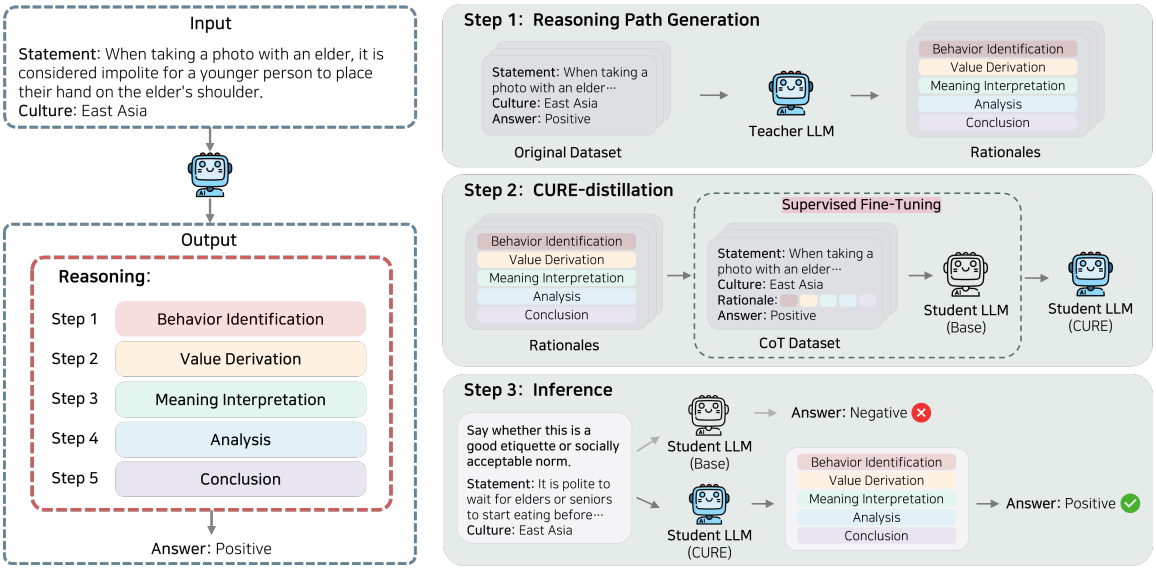

Figure 1: The CURE framework and the CURE-distillation process.

lation method that strengthens the student model's reasoning ability even when the teacher model exhibits limited performance.

3 Methodology

3.1 Problem Statement

We evaluate the cultural adaptability using the cultural ethics judgment task proposed by Dwivedi et al. (2023). Given a statement s that describes a social behavior and a culture c in which the behavior occurs, the model predicts a binary answer $a \in \{positive, negative\}$, indicating whether the behavior is socially acceptable within that culture.

3.2 CURE

We propose CURE, a reasoning framework grounded in cultural value theory. According to Schwartz (1992), culture is not merely a collection of individual facts but a system shaped by underlying value structures. Cultural differences arise from variations in how these value systems are composed and prioritized. The interpretation of behavior is determined by the values emphasized within a particular culture. In other words, evaluating the social acceptability of a behavior requires understanding the cultural values that guide how a society interprets and assigns meaning to that behavior. CURE formalizes this insight into a structured reasoning process. It interprets and analyzes social behaviors through the lens of cultural values, enabling logical inference about whether a given behavior is acceptable in a given culture.

As illustrated in Figure 1, CURE consists of five reasoning steps: Behavior identification (B), Value derivation (V), Meaning interpretation (M), Analysis (A), and Conclusion (C).

The reasoning process is as follows. First, in step B, the behavior to be evaluated is extracted from the statement. In this example, *"not placing a hand on an elder's shoulder when taking a photo"* is identified as potentially socially sensitive within the culture. Second, in step V, the core cultural values relevant to the behavior are identified. In East Asia, respect for hierarchical relationships based on age and social status serves as a key cultural norm guiding interpersonal interactions. Third, in step M, the cultural implications of the behavior are interpreted in light of these values. Within a value system emphasizing respect for age and authority, the behavior is understood as a younger person's way of showing politeness to an elder. Fourth, in step A, the model assesses whether this interpretation aligns with or contradicts the cultural value system. In this case, the behavior aligns with the social values of East Asian cultures. Finally, in step C, the model determines whether the behavior is socially acceptable based on the preceding reasoning. Thus, CURE functions as a cumulative reasoning framework in which information derived at each step is carried forward to progressively build richer and more contextually grounded reasoning.

3.3 CURE-distillation

We propose CURE-distillation, a method for transferring cultural reasoning abilities. In standard CoT distillation (Magister et al., 2023; Zhao et al., 2024), a teacher model generates a rationale and a corresponding answer for a given question, and a student model is trained to imitate this reasoning process. However, in domains such as cultural ethics, teacher-generated rationales may contain biased interpretations, which can be propagated to student models through distillation.

To address this, we adapt a label-conditioned rationale generation strategy from Liu et al. (2025). While this approach conditions rationale generation on the gold answer, we further constrain the teacher model to treat it as a fixed premise rather than independently interpreting it. The prompt explicitly presents the cultural judgment implied by the answer and instructs the model to explain why the described behavior is considered acceptable or unacceptable within the given cultural context. This guides the teacher model to interpret the statement from the perspective of the specified culture rather than relying on its own cultural priors.

In Figure 1, the green panels on the right illustrate the CURE-distillation process. In the first step, the teacher model sequentially generates CURE reasoning steps from the dataset by conditioning on the gold answer and following the prompt described above. These reasoning steps are then aggregated into structured rationales.

Specifically, when the gold answer is *positive*, the statement is treated as describing behavior that is socially acceptable within the given culture, and the teacher model generates rationales explaining why the behavior aligns with the cultural value system. Conversely, when the gold answer is *negative*, the statement is treated as describing inappropriate behavior, and the model generates rationales explaining how the behavior violates cultural norms and conflicts with cultural values. The model is further prevented from considering counterfactual or alternative interpretations beyond the provided premise, ensuring that reasoning remains consistent with the given cultural context. The detailed prompts are provided in Appendix C.

In addition, we apply a validation process to filter rationales that are inconsistent with the gold answer. Only samples for which the answer inferred from the generated rationale matches the gold answer are retained for training, while the rest are discarded. This procedure ensures that the retained rationales remain logically consistent with the given premise, thereby reducing the risk of distilling illogical or erroneous reasoning paths. The filtering rates and their distribution are reported in Appendix A.

In the second step, we construct a CoT dataset and use it to train the student model. Given an input (s, c), the student model is trained to first generate a rationale r following the CURE and then produce an answer a conditioned on r. The rationale and answer are treated as a single continuous output sequence and generated in an autoregressive manner. During training, the model is optimized to maximize the conditional likelihood $P(r, a|s, c)$ using a language modeling loss, without distinguishing between the rationale and answer or assigning different weights to them. This process is expressed as follows.

$$\mathcal{L} = -\,\mathbb{E}_{(s,c,r,a)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log P(y_t \mid s, c, y_{<t})\right]$$

Here, y denotes the output sequence that concatenates r and a, and T represents its total length. Through this training, the student model internalizes a structured cultural reasoning ability grounded in cultural value systems.

4 Experimental Setup

Dataset We used the EtiCor++ dataset (Dwivedi et al., 2025), which assesses the social acceptability of statements across five cultural regions: East Asia (EA), India (IN), Latin America (LA), Middle East and Africa (MEA), and North America and Europe (NE). We removed 1,549 samples that were likely mislabeled due to a surface-level labeling heuristic in the original dataset. These statements began with "Do not" and were labeled as *negative*, although this expression does not necessarily indicate socially unacceptable behavior. To evaluate generalizability, we ensured that highly similar statement pairs were assigned to the same data split. We used a Jaccard similarity threshold of 0.75 and a cosine similarity threshold of 0.90 computed using all-MiniLM-L6-v2 (Wang et al., 2020). Both thresholds were determined empirically through manual inspection. The remaining statements were sorted according to their average cosine similarity to all other statements and assigned to the test, validation, training sets in as-

Dataset	EA	IN	LA	MEA	NE	Total
Training Set	6,793	2,638	4,928	8,926	8,107	31,392
Validation Set	969	376	703	1,286	1,158	4,492
Test Set	1,939	752	1,407	2,553	2,316	8,967
Total	9,701	3,766	7,038	12,765	11,581	44,851

Table 1: Statistics of the dataset.

Method	Acc.	F1
zero-shot	59.06	64.27
standard CoT	58.72	61.70
standard CoT w/ gold rationales	<u>75.72</u>	<u>80.18</u>
CURE w/ silver rationales	58.39	57.73
CURE	**96.22**	**97.19**

Table 2: Performance of the teacher model. "standard CoT w/ gold rationales" denotes the use of gold rationales generated by label-conditioned standard CoT, while "CURE w/ silver rationales" denotes the use of silver rationales generated by CURE without label conditioning. The highest performance is shown in bold, and the second-highest is underlined.

cending order to control the difficulty and diversity of the test set. The final dataset was split into training, validation, and test sets at a 70:10:20 ratio, with statistics reported in Table 1.

We also used the NormAd dataset (Rao et al., 2025) for out-of-distribution (OOD) evaluation. NormAd differs from EtiCor++ in cultural coverage, input format, and judgment setting. It covers 75 countries and presents normative behaviors in concrete scenarios, requiring models to judge whether a character's behavior is socially and culturally appropriate in context. These differences introduce distribution shifts, making NormAd suitable for evaluating the OOD generalizability of acquired cultural reasoning abilities.

Models We used GPT-4o-mini as the teacher model, with the temperature set to 0.7. As student models, we selected Llama3.2-1B-Instruct, Llama3.2-3B-Instruct (Dubey et al., 2024), Qwen2.5-3B-Instruct, and Qwen3-4B-Instruct-2507 (Yang et al., 2025). After evaluating the performance of all student models, we conducted detailed experiments using Llama3.2-3B-Instruct.

Baselines We evaluate our method against four baselines: zero-shot, standard CoT, Answer-SFT, and standard CoT-distillation. For standard CoT, we append the prompt *"Let's think step by step."* (Kojima et al., 2022) to induce reasoning. Answer-SFT trains models to directly predict answers without reasoning. In standard CoT-distillation, gold rationales are generated using standard CoT under a label-conditioned setting, and models are fine-tuned on these rationales.

Environment The models were fine-tuned using LoRA (Hu et al., 2021), with hyperparameter settings provided in Appendix B. Training and evaluation were conducted on a single RTX A6000 GPU, and vLLM (Kwon et al., 2023) was used during inference. All experiments were implemented using HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019).

Metrics We evaluated model performance using accuracy and F1-score. Accuracy measures overall prediction correctness, and the F1-score mitigates potential overestimation of accuracy when predictions are biased toward a particular class by balancing precision and recall.

Checkpoint Selection Checkpoints were saved and evaluated every 400 steps for each cultural region and every 1,000 steps for the multi-cultural setting. The checkpoint with the highest accuracy was selected as the final model.

5 Evaluation

5.1 Effect of CURE

As shown in Table 2, the teacher model exhibits low zero-shot performance, which further degrades when standard CoT is applied. These findings are consistent with prior work indicating that CoT prompting is effective for logical tasks but provides little or no improvement for tasks requiring knowledge or common sense (Sprague et al., 2025). This suggests that reasoning patterns acquired during general post-training may be insufficient for cultural reasoning. CURE with silver rationales also fails to improve performance, as the model remains dependent on its own subjective interpretations.

In contrast, standard CoT with gold rationales significantly improves performance by guiding the model toward more plausible reasoning. Nevertheless, these rationales are still influenced by the model's subjective interpretations and do not consistently lead to correct predictions. However, CURE achieves significant performance gains by constraining the reasoning space according to the internal logic of the target culture. This indi-

Method	Teacher		Llama3.2-1B		Llama3.2-3B		Qwen2.5-3B		Qwen3-4B	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
zero-shot	59.06	64.27	55.98	54.82	41.53	49.17	72.43	72.38	68.17	69.48
standard CoT	58.72	61.70	35.93	40.41	43.89	45.17	67.30	66.74	51.59	59.61
Answer-SFT	-	-	72.79	70.83	72.54	72.48	82.39	81.93	86.47	86.26
standard CoT-distillation	-	-	59.77	56.76	67.71	65.95	63.32	59.53	64.05	60.55
CURE-distillation	-	-	**82.48**	**82.74**	**88.52**	**88.57**	**90.74**	**90.85**	**92.41**	**92.47**

Table 3: Effects of CURE-distillation.

Method	EA		IN		LA		MEA		NE		All Regions	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Teacher	49.56	49.56	70.61	70.23	54.02	58.53	57.54	63.82	57.25	65.02	59.06	64.27
zero-shot	43.01	48.13	53.06	55.57	38.59	46.01	39.01	48.25	38.86	47.34	41.53	49.17
standard CoT	44.66	44.53	52.93	49.08	41.01	41.97	41.95	44.03	44.43	44.51	43.89	45.17
Answer-SFT	84.68	84.39	79.65	79.04	83.94	84.79	78.30	77.34	82.30	82.00	72.54	72.48
standard CoT-distillation	69.21	68.47	64.49	61.52	64.46	62.08	67.76	66.17	66.97	64.48	67.71	65.95
CURE-distillation	**87.73**	**87.76**	**84.97**	**84.97**	**87.92**	**88.01**	**88.09**	**88.08**	**89.46**	**89.54**	**88.52**	**88.57**

Table 4: Effects of CURE-distillation across various cultural regions.

cates that effective cultural reasoning depends not merely on answer alignment or increased output length, but on structured reasoning paths that center on cultural values while mitigating inherent model biases. A qualitative comparison of the gold rationales generated by standard CoT and CURE is provided in Section 6.1.

5.2 Effect of CURE-distillation

As shown in Table 3, CURE-distillation achieves performance gains exceeding 20%p across all student models relative to the teacher model, indicating that student models can acquire cultural reasoning capabilities despite the teacher model's limited reasoning ability. CURE-distillation also surpasses all baselines, whereas standard CoT-distillation yields only modest gains confined to Llama3.2-1B and Llama3.2-3B. This may be because its rationales resemble narrative descriptions of cultural norms rather than systematic reasoning that explicitly grounds the final judgment in logical justification. A qualitative analysis of this limitation is provided in Section 6.2.

In contrast, CURE provides uniform reasoning structures grounded in the internal logic of the target culture, framing cultural judgments as systematic reasoning tasks. This suggests that robust cultural reasoning requires culturally grounded reasoning structures rather than increased knowledge or standard CoT transfer. The performance difference between Answer-SFT and CURE-distillation is statistically significant for all student models according to McNemar's test ($p < 0.001$).

5.3 Effect of CURE-distillation across Diverse Cultural Regions

Table 4 presents the results of CURE-distillation across cultural regions. Student models trained with CURE-distillation consistently outperform the teacher model, with an average accuracy improvement of 29.84%p across all regions. This indicates that the effectiveness of CURE-distillation is not confined to a specific culture but generalizes well across diverse cultural contexts. However, the performance gain in IN is relatively modest, possibly due to limited training data availability.

In mixed-culture settings, performance declines for other methods, whereas CURE-distillation maintains its performance. This robustness may stem from the model's exposure to diverse reasoning patterns, which amplifies the effectiveness of distillation. The model's ability to make culturally aligned judgments even in mixed cultural contexts demonstrates the level of cultural adaptability required for real-world applications. McNemar's test indicates statistically significant performance differences between Answer-SFT and CURE-distillation across all regions ($p < 0.001$).

5.4 Cultural Adaptability on OOD Data

As reported in Table 5, CURE-distillation achieves the best overall performance and demonstrates strong generalization under distribution shift. Answer-SFT shows a substantial performance drop, reflecting a bias-variance trade-off. Standard CoT-distillation exhibits relatively stable per-

Method	EtiCor++		NormAd	
	Acc.	F1	Acc.	F1
zero-shot	41.53	49.17	51.43	34.06
standard CoT	43.89	45.17	44.06	51.49
Answer-SFT	72.54	72.48	50.55	38.93
standard CoT-distillation	67.71	65.95	71.40	71.33
CURE-distillation	**88.52**	**88.57**	**81.35**	**81.45**

Table 5: Cultural adaptability on NormAd dataset.

Held-Out Culture	Answer-SFT		standard CoT -distillation		CURE -distillation	
	Acc.	F1	Acc.	F1	Acc.	F1
EA	72.87	70.97	66.06	64.18	**76.95**	**77.22**
IN	73.54	72.42	68.09	64.33	**76.86**	**76.50**
LA	74.56	73.18	62.90	60.25	**75.20**	**75.41**
MEA	68.43	64.99	63.65	61.62	**76.73**	**77.60**
NE	**81.09**	**80.89**	66.54	63.80	77.76	77.88

Table 6: Cross-cultural generalizability.

formance under distribution shift, likely due to the implicit regularization induced by reasoning-based training. However, this regularization may oversmooth the learned representations, leading to less discriminative decision boundaries and limiting performance gains even in in-distribution settings. Consequently, standard CoT-distillation appears less suitable for tasks that require judgments based on fine-grained contextual cues.

In contrast, reasoning supervision in CURE-distillation provides a strong learning signal that promotes cultural reasoning. This enables the model to achieve high in-distribution performance while maintaining robustness under distribution shift, with only limited performance degradation. These results further suggest that CURE-distillation facilitates the acquisition of fine-grained culture-specific normative criteria and distribution-invariant feature representations. Consequently, it achieves a more desirable bias-variance balance, highlighting its effectiveness in acquiring cultural adaptability.

5.5 Cross-Cultural Generalization

We conducted cross-validation, excluding each culture from training and using it for evaluation. To control for dataset size, we randomly sampled 2,000 instances per culture. As shown in Table 6, CURE-distillation achieves more robust generalization than Answer-SFT, likely due to its distinct generalization strategy. It enables more reliable transfer by abstracting structured reasoning processes, while Answer-SFT tends to rely on spurious input-label correlations. This contrast is more pronounced in EA, IN, and MEA, suggesting that shortcut-based predictions may be less effective in contexts requiring greater cultural interpretation.

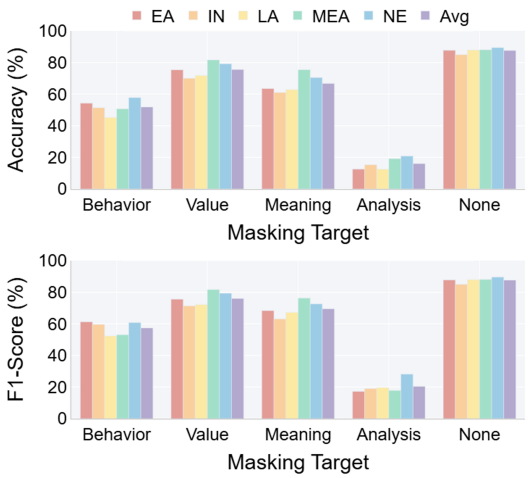

Figure 2: Step-wise ablation of CURE.

The strong performance of Answer-SFT in NE may reflect its closer alignment with a global decision boundary shaped by training on diverse cultural contexts. Such exposure encourages the model to learn more generalized decision criteria, which can be applied more stably in cultures with less extreme distributional bias. Consequently, shortcuts based on surface-level statistical patterns may remain relatively effective in such contexts.

5.6 Contribution Analysis of CURE's Reasoning Steps

To evaluate the contribution of each reasoning step in CURE, we constructed ablated versions of the rationales by individually removing each step. The removed text was replaced with $[MASK]$ tokens of the same length, and the student models were trained on each ablated rationale. Figure 2 shows that removing the Analysis step causes the largest performance drop, as it integrates information from preceding steps into a logical path toward the final judgment. Masking the Meaning interpretation and Behavior identification steps also leads to notable performance declines, suggesting that understanding how behaviors are perceived within a culture is an important prerequisite for culturally grounded judgments. These findings show that CURE accumulates and synthesizes information in a structured, hierarchical manner. McNemar's test indicates that removing the Value derivation step results in a statistically

Statement: A woman typically requires the consent of her male guardian, often the father or husband, to remarry after a divorce or the death of her spouse.
Culture: Middle East and Africa

Method	Rationale	Prediction
standard CoT	To evaluate the statement regarding the requirement for a woman's consent to remarry in Middle East and Africa, we can consider several factors: 1. Cultural Context: In many cultures, traditional practices often dictate familial roles and responsibilities. The expectation for a woman to seek consent from a male guardian may be rooted in historical norms that emphasize patriarchal structures. 2. Legal Framework: This practice may be enshrined in local laws, reflecting societal values and beliefs about marriage and gender roles. The legal requirements may be seen as a way to protect women and ensure family stability, although interpretations of protection can vary significantly. 3. Social Dynamics: While some may view this requirement as a protective measure, others may perceive it as restrictive or outdated. The impact on women's autonomy and individual rights can lead to differing opinions on whether it is socially acceptable. 4. Shift in Perspectives: In recent years, there has been a push for greater gender equality and women's rights in many parts of the world, which can influence perceptions of such practices. The acceptance of these norms may vary among different generations and social groups. 5. Comparative Analysis: When comparing this practice to norms in other regions, it may be viewed differently based on cultural, legal, and social frameworks. What is considered acceptable in one context may be viewed as problematic in another. Considering these points, the societal acceptance of the requirement for a woman's guardian's consent to remarry is complex and layered, reflecting ongoing debates about gender roles, rights, and cultural traditions.	Negative
CURE	The specific behavior mentioned in the given statement is 'a woman requiring the consent of her male guardian to remarry after a divorce or the death of her spouse.' In many Middle Eastern and African cultures, there is a strong emphasis on family structure, patriarchal authority, and the preservation of social order. These societies often prioritize familial consent and male guardianship in matters of marriage and relationships. The requirement for a woman to obtain her male guardian's consent before remarrying is interpreted as a reflection of respect for family hierarchy and the role of men as protectors and decision-makers. This practice is seen as a way to uphold traditional values and maintain social stability, ensuring that family interests are considered in personal decisions. The behavior of requiring male guardian consent to remarry aligns with the cultural value of patriarchal authority and familial integrity. This norm reinforces the traditional roles assigned to men and women, highlighting the importance of male oversight in significant life decisions. By adhering to this practice, individuals contribute to the preservation of cultural identity and social cohesion, reflecting an understanding of the roles that family dynamics play in personal relationships.	Positive

Table 7: Examples of teacher-generated gold rationales.

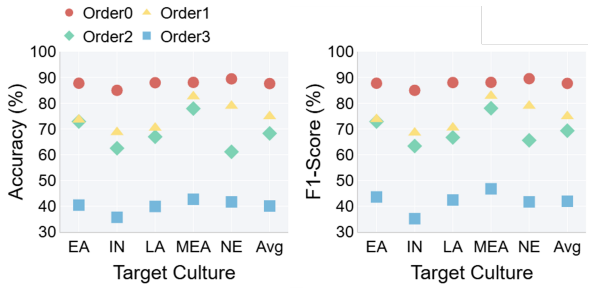

Figure 3: Performance changes according to variations in the logical structure of CURE.

significant performance decrease compared to the CURE ($p < 0.001$), suggesting that each step contributes meaningfully to the reasoning process.

5.7 Effect of CURE's Reasoning Structure

In tasks with strict logical dependencies between steps, such as mathematics, each step directly builds on the preceding one, making performance highly sensitive to perturbations in reasoning order. By contrast, cultural ethics judgment is better characterized as a contextual understanding task that requires the joint interpretation of a statement and its cultural context. Moreover, because LLMs can leverage contextual information, strong performance alone cannot distinguish between genuine use of the intended reasoning process and reliance on contextual cues, such as cultural descriptions and evaluative expressions.

To examine whether CURE functions merely as a collection of contextual cues or as a structured reasoning framework, we shuffled the teacher-generated reasoning steps, trained student models on the shuffled rationales, and analyzed the resulting performance changes. If CURE is used only as a bag of cues, performance degradation should be limited, as the informational content remains unchanged. In contrast, a systematic performance decline under increasing structural disruption would suggest that the model benefits not only from individual cues but also from CURE's sequential organization and logical coherence. We defined four step-ordering variants: Order 0 corresponds to the original CURE, while Orders 1–3 progressively disrupt inter-step dependencies and logical coherence, with Order 3 representing the strongest structural disruption.

In Figure 3, performance consistently declines across all cultures as the logical structure is increasingly disrupted. Order 1 shows a 12.76%p decrease in accuracy relative to Order 0, suggesting that even partial disruption of the reasoning structure reduces performance. In Order 2, accuracy drops by an average of 19.37%p, indicating that distortion of inter-step dependencies severely

Statement: In Iran, a married woman is expected to obtain her husband's permission before traveling abroad.
Culture: Middle East and Africa

Method	Rationale	Prediction
standard CoT-distillation	Let's analyze the statement step by step: 1. Cultural Context: In many cultures, including some in the Middle East, there are traditional norms regarding marriage and family. These norms can vary widely, and they often dictate how individuals, especially women, are expected to behave in certain situations. 2. Marriage and Travel: The expectation for a married woman to obtain her husband's permission before traveling abroad reflects a traditional view that emphasizes the role of the husband in decision-making, particularly in matters related to family and personal life. 3. Social Norms: In some societies, this practice is seen as a way to maintain family values and respect for the husband's authority. It can also be viewed as a means to ensure the safety and well-being of the woman while traveling. 4. Individual Freedom: However, this norm can also be seen as restrictive, limiting the autonomy of women and their ability to make choices about their own lives, including their personal travel plans. 5. Changing Trends: In recent years, there has been a shift towards greater gender equality and individual rights in many parts of the world, including Iran. This could lead to a reevaluation of traditional norms, with some individuals advocating for more independence in personal decisions. Considering these points, the expectation for a married woman to obtain her husband's permission before traveling abroad can be viewed as a reflection of cultural norms that may not align with contemporary values of personal freedom and equality.	Negative
CURE-distillation	The specific behavior mentioned in the given statement is 'a married woman obtaining her husband's permission before traveling abroad.' In many Middle Eastern and African cultures, particularly in Iran, there is a strong emphasis on family honor, marital respect, and the role of men as guardians and protectors of their families. In Iranian culture, a married woman obtaining her husband's permission before traveling abroad is perceived as a demonstration of respect for her husband's authority and the family's social standing. It reflects an understanding of the traditional roles that men and women are expected to adhere to within the family structure, where the husband is seen as the head of the household. The behavior of a married woman seeking her husband's permission before traveling aligns with the cultural values of respect for authority and familial hierarchy prevalent in Iranian society. This practice reinforces the notion of gender roles and the importance of maintaining family honor. It signifies a commitment to the social norms that govern personal relationships and travel, ensuring that decisions made by women are aligned with the expectations of their husbands and the broader community. This behavior serves to uphold the cultural values of respect and propriety, reflecting the significance of family and societal expectations in decision-making processes.	Positive

Table 8: An example of a standard CoT-distillation error and the corresponding successful reasoning result from CURE-distillation. Red highlights mark the segments where erroneous reasoning occurs, while yellow and green highlights denote the Value derivation and Meaning interpretation steps, respectively.

hinders valid reasoning. In Order 3, both accuracy and F1-score drop by more than half across all cultures, demonstrating the severe consequences of logical collapse. These results suggest that logical organization plays an important role in reasoning, and the model relies on this structure for reasoning.

6 Analysis

6.1 Qualitative Analysis of Rationales

Table 7 shows that the teacher model produces an incorrect prediction under standard CoT, but correctly predicts the answer when using CURE. Under standard CoT, the rationale reflects an interpretation that is not grounded in the target cultural perspective. Although the cultural context is described, the statement is evaluated according to contemporary normative standards such as *gender equality*. The rationale also employs explicitly negative terms, including *restrictive*, *outdated*, and *problematic*, thereby framing the statement unfavorably and consequently leading the model to judge it as unacceptable.

In contrast, CURE interprets the statement within the target culture, treating it as an established cultural convention and explaining its acceptability through culturally grounded norms such as *family hierarchy* and *patriarchal authority*. The rationale coherently connects cultural context, underlying values, and the final judgment. This structural consistency suggests that CURE provides more learnable reasoning patterns that facilitate the acquisition of cultural reasoning.

6.2 Error Analysis

Table 8 presents a representative error case of a model trained with standard CoT-distillation. Despite accurately describing the target cultural context, the model still produces incorrect judgments grounded in contemporary normative values. This pattern suggests that, although cultural knowledge is acquired at a descriptive level, the model fails to internalize the reasoning processes required to evaluate cultural acceptability within the target cultural perspective. This limitation may result from teacher model's standard CoT rationales that describe cultural context without explicitly linking such descriptions to the criteria guiding the final judgment. As a result, the student model does not sufficiently acquire culturally grounded judgment strategies and instead continues to rely on value standards implicitly reinforced during pretraining.

Statement: Smoking is a predominantly male behavior in India. **Culture:** India		
CURE-distillation	The specific behavior mentioned in the given statement is 'smoking being predominantly a male behavior in India'. Indian culture values gender roles and social norms that often emphasize traditional gender distinctions, including differences in behavior and social participation. There is a strong emphasis on respect for cultural traditions and the roles assigned to men and women in various contexts. In India, the assertion that smoking is predominantly a male behavior can be perceived as reinforcing outdated stereotypes and gender biases. It may suggest a lack of awareness of the evolving social landscape where women are increasingly participating in various activities, including smoking, and may overlook the complexities of individual preferences and choices. The statement conflicts with the contemporary understanding of gender roles in India, which is moving towards greater equality and inclusivity. By suggesting that smoking is exclusively a male behavior, it undermines the progress made in challenging traditional gender norms and can be seen as perpetuating stereotypes that are no longer relevant in modern Indian society. This misalignment with the evolving social dynamics can lead to social disapproval and a perception of ignorance regarding the changing cultural landscape.	Negative
Statement: It's considered polite to finish your meal. **Culture:** East Asia		
CURE-distillation	The specific behavior mentioned in the given statement is 'finishing your meal completely'. In East Asian cultures, the value of moderation and respect for food is deeply rooted in social etiquette. There is an emphasis on sharing food, showing appreciation for the meal, and maintaining a balance between consumption and generosity. In East Asian cultures, finishing a meal completely can be perceived as wasteful or overly greedy. It may suggest a lack of appreciation for the food and the effort that went into preparing it, as well as a disregard for the communal aspect of dining. The act of leaving some food on the plate is often seen as a sign of respect for the host and the meal, as it indicates that the guest is satisfied but also acknowledges the abundance of the food. Finishing a meal completely contradicts the cultural value of moderation and respect for food in East Asian societies. This behavior can be interpreted as a failure to honor the communal dining experience and the effort of the host. It undermines the cultural norm of sharing and can lead to social disapproval, as it may be seen as a lack of gratitude and an excessive focus on personal consumption rather than the social aspect of the meal.	Negative

Table 9: Representative error cases of a student model trained with CURE-distillation. Yellow highlights indicate the Value derivation step, while red highlights mark the segments where erroneous reasoning occurs.

In addition, Table 9 presents two representative error cases of a model trained with CURE-distillation. The first error type occurs when the model follows CURE's reasoning structure but fails to carry the target culture's acceptability criteria through to the final judgment. Although the model identifies *traditional gender roles* and *gendered social norms* as relevant cultural values, it evaluates the statement from contemporary normative perspectives, interpreting it as *reinforcing outdated stereotypes and gender bias*. Thus, the identified values are not retained as the basis for judgment. Rather than applying the target culture's acceptability criteria, the model adopts contemporary normative standards and predicts *negative*, even though the statement is aligned with the cultural values. This suggests that the student model can identify culture-specific criteria, but has not fully internalized how to consistently carry those criteria through to the final acceptability judgment.

The second error type occurs when the model identifies a mixture of relevant and context-dependent cultural values and misapplies them when interpreting the behavior. In this case, the model derives a relevant value such as *respect for food*, along with more context-dependent values such as *moderation*, *sharing food*, and *balance between consumption and generosity*. As a result, the model misinterprets *"finishing a meal completely"* as *wasteful* or *greedy*, rather than as an expression of *waste avoidance* or *satisfaction*. This error appears to arise from overgeneralizing dining etiquette knowledge from specific hosting contexts, where leaving some food may signal that one has been sufficiently served, to a general dining context. It therefore reflects the model's failure to identify the appropriate interpretive context for the statement, which then propagates to the judgment of behavior-culture alignment.

Taken together, these errors suggest that CURE-distillation's limitations lie less in acquiring the reasoning structure itself than in consistently applying identified cultural values across subsequent reasoning steps. In particular, pretrained biases may affect value-based interpretation, and errors at this stage can propagate to alignment judgments and final predictions. This indicates that CURE-distillation remains limited in reliably controlling value-based interpretation and inter-step error propagation.

6.3 Case Study: Cultural Adaptability

To assess the student model's sensitivity to cultural context, we fixed the input statement and varied only the cultural context. As shown in Table 10,

Method	Original Context		Shifted Context	
	Acc.	F1	Acc.	F1
Answer-SFT	72.54	72.48	58.40	52.09
CURE-distillation	**88.52**	**88.57**	**87.13**	**87.13**

Table 10: Cultural adaptability under shifted cultural contexts.

Answer-SFT exhibits a 14.14%p accuracy drop under the shifted cultural context, whereas CURE-distillation decreases by only 1.39%p. These results suggest that CURE-distillation is more robust to cultural context shifts and better aligns its judgments with the target culture.

Table 11 presents examples in which CURE-distillation changes its judgments in response to cultural context, whereas Answer-SFT produces invariant judgments. This contrast indicates that Answer-SFT fails to incorporate cultural information into its predictions, whereas CURE-distillation flexibly adapts its evaluative criteria based on contextual cues. In particular, the Value derivation and Meaning interpretation steps, highlighted in yellow and green, represent the interpretation of culture-specific values and the cultural significance of the behavior, illustrating how such information is translated into the final judgment. These examples suggest that the context-dependent judgments produced by CURE-distillation are not arbitrary but aligned with the value systems of the respective cultures, providing qualitative evidence of the model's cultural adaptability.

7 Conclusion

We proposed a structured reasoning framework and a training method to improve the cultural adaptability of LLMs. Our approach trains models to acquire reasoning abilities even when the teacher model fails to generate high-quality reasoning paths on its own. Experimental results show that our method outperforms existing approaches in cultural ethics judgment and generalizes effectively to novel scenarios. These findings indicate that internalizing reasoning structure can be more effective than simply expanding cultural knowledge for improving cultural fairness and acceptability of LLMs.

8 Ethical Considerations

This study addresses culturally sensitive aspects, which necessitate consideration of ethical implications and potential risks. First, since we abstract culture in terms of the values and norms of specific groups, it may overgeneralize cultural judgments and fail to capture intra-cultural diversity or individual differences. Second, our approach relies on a specific dataset and the cultural expressions it contains, which limits the scope and interpretation of the results. Consequently, the framework may not fully reflect all cultural contexts and may risk reinforcing stereotypes or biases associated with particular cultures. Finally, caution is required when applying the model's cultural judgments directly to real-world decision-making, as inaccurate or biased outputs may lead to unintended harm to certain groups. Accordingly, we emphasize that this framework is intended to support human decision-making rather than replace it.

Statement: People usually use chopsticks when having a meal.	
EA	**Rationale:** The specific behavior mentioned in the given statement is 'using chopsticks when having a meal.' In East Asian cultures, the use of chopsticks is deeply rooted in tradition, symbolism, and the cultural significance of dining etiquette. It reflects values of respect for food, the dining experience, and the communal aspect of sharing meals. In East Asian cultures, using chopsticks is not only a practical tool for eating but also a sign of respect for the food and the dining experience. It signifies attentiveness to the meal, the company, and the cultural practices that govern dining etiquette. Using chopsticks is perceived as a way to honor the food and the people sharing the meal, as it is a traditional method that conveys respect and mindfulness. The behavior of using chopsticks aligns with the cultural values of respect, tradition, and communal harmony in East Asia. This practice reflects an understanding of the importance of dining etiquette and the significance of food as a cultural artifact. By adhering to this norm, individuals demonstrate their respect for the meal, the people involved, and the cultural heritage that shapes their dining practices. This behavior fosters a sense of belonging and connection within the community, reinforcing social bonds and cultural identity. Therefore, the answer is: positive.
IN	**Rationale:** The specific behavior mentioned in the given statement is 'using chopsticks when having a meal.' In Indian culture, the use of chopsticks is generally not common, especially in most traditional and everyday meals. Instead, the preferred utensils for eating are hands, fingers, or specific flatware such as spoons and forks, which are more culturally accepted and appropriate for various types of food. In India, using chopsticks for eating is often perceived as a foreign or unconventional practice. It may be seen as a lack of familiarity with local customs and can be interpreted as a disregard for the traditional dining practices that are deeply rooted in the culture. Using chopsticks in a meal conflicts with the cultural norms in India that emphasize the use of hands, fingers, or specific utensils that are more aligned with the local culinary practices. This behavior can be viewed as an affront to cultural identity and may lead to social disapproval, as it disrupts the expected etiquette surrounding dining and communal sharing of food. Therefore, the answer is: negative.
Statement: Maintain clear communication and direct eye contact during interactions.	
NE	**Rationale:** The specific behavior mentioned in the given statement is 'maintaining clear communication and direct eye contact during interactions.' In North American and European cultures, effective communication is highly valued, and direct eye contact is seen as a sign of confidence, sincerity, and engagement in conversations. In these cultures, maintaining clear communication and direct eye contact is interpreted as a demonstration of respect and attentiveness. It conveys that the speaker is fully present and interested in the conversation, fostering a sense of connection and mutual understanding. This behavior is often associated with building trust and rapport. Maintaining clear communication and direct eye contact aligns with the cultural values of assertiveness and openness prevalent in North America and Europe. This behavior reflects a commitment to effective dialogue and mutual respect, which are essential for successful interpersonal interactions. By engaging in direct eye contact, individuals demonstrate their willingness to listen actively and express themselves clearly, reinforcing social bonds and enhancing the overall quality of communication. Therefore, the answer is: positive.
EA	**Rationale:** The specific behavior mentioned in the given statement is 'maintaining clear communication and direct eye contact during interactions.' East Asian cultures often value indirect communication, subtlety, and maintaining harmony in social interactions. The emphasis is placed on non-verbal cues, such as body language and facial expressions, which can convey respect and attentiveness without the need for direct confrontation or assertiveness. In East Asian cultures, maintaining direct eye contact can be perceived as aggressive or confrontational. Instead, individuals typically prefer to communicate through more subtle means, such as nodding, using gestures, or maintaining a gentle gaze. This approach reflects a cultural preference for maintaining harmony and avoiding direct confrontation, which is often seen as a sign of respect and consideration for others' feelings. The behavior of maintaining clear communication and direct eye contact contradicts the cultural value of subtlety and indirectness in East Asian interactions. This action may be interpreted as overly assertive or confrontational, which can disrupt the harmony that is highly valued in these cultures. Such behavior can lead to discomfort or offense, as it disregards the importance of maintaining a respectful and non-confrontational demeanor in social exchanges. Therefore, the answer is: negative.
Statement: Women often wear traditional attire that includes a large, body-covering cloth (moulafa or wraparound skirt) and a head covering.	
MEA	**Rationale:** The specific behavior mentioned in the given statement is 'women wearing traditional attire that includes a large, body-covering cloth (moulafa or wraparound skirt) and a head covering.' In many Middle Eastern and African cultures, traditional attire is deeply rooted in cultural identity, religious practices, and social norms that emphasize modesty, respect for tradition, and adherence to cultural values. In these cultures, wearing traditional attire such as a moulafa or wraparound skirt and a head covering is perceived as a sign of respect for cultural heritage and religious beliefs. It signifies a commitment to maintaining modesty and adhering to societal expectations regarding gender roles and appearance. Wearing traditional attire aligns with the cultural values of modesty, respect for tradition, and cultural identity in the Middle East and Africa. This behavior reflects an understanding of the significance of clothing as a means of expressing cultural identity and adherence to social norms. It fosters a sense of community and shared values, reinforcing the importance of cultural heritage and the role of women within it. By donning such attire, individuals demonstrate their connection to their cultural roots and contribute to the preservation of traditional practices that are valued within their communities. Therefore, the answer is: positive.
NE	**Rationale:** The specific behavior mentioned in the given statement is 'women wearing traditional attire that includes a large, body-covering cloth (moulafa or wraparound skirt) and a head covering.' In many North American and European cultures, there is a strong emphasis on individual expression, personal choice, and the freedom to dress as one pleases, regardless of cultural or religious affiliations. In these regions, the act of wearing traditional attire that includes a large, body-covering cloth and a head covering may be perceived as restrictive or oppressive, especially if it is seen as limiting personal expression or autonomy. It can be interpreted as a rejection of modern values that prioritize individualism and the ability to choose one's attire freely. The behavior of wearing traditional attire that includes a large, body-covering cloth and a head covering conflicts with the cultural values of individualism and personal choice prevalent in North America and Europe. This behavior may be viewed as an imposition of cultural norms that restrict personal freedom and autonomy, leading to social disapproval or discomfort among those who value the ability to express oneself through fashion choices. Such attire can be seen as a challenge to the cultural narrative of equality and individuality that is highly regarded in these regions. Therefore, the answer is: negative.

Table 11: Rationales generated by the student model trained via CURE-distillation. Yellow highlights denote the Value derivation step, and green highlights denote the Meaning interpretation step, where the behavior is interpreted based on the identified values. Purple highlights denote the final judgment resulting from the reasoning process.

Acknowledgments

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Develop AI agent technology to connect knowledge through public cultural facility-based discussion and communication, Project Number: RS-2026-25520645, Contribution Rate: 50%, Project Name: Development of an AI Agent Integrating Korean Language Knowledge for Personalized Language Consultation Services, Project Number: RS-2026-25506607, Contribution Rate: 50%)

References

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

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

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)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Ruth Benedict. 2019. *Patterns of culture*. Routledge.

Alex J. Chan, José Luis Redondo García, Fabrizio Silvestri, Colm O'Donnell, and Konstantina Palla. 2024. Enhancing content moderation with culturally-aware models.

Xinghao Chen, Zhijing Sun, Guo Wenjin, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. 2025. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15094–15119, Vienna, Austria. Association for Computational Linguistics.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Improve student's reasoning generalizability through cascading decomposed cots distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15643.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270.

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*.

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

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. Eticor: Corpus for analyzing llms for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931.

Ashutosh Dwivedi, Siddhant Shivdutt Singh, and Ashutosh Modi. 2025. EtiCor++: Towards understanding etiquettical bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9355–9376, Vienna, Austria. Association for Computational Linguistics.

Tao Feng, Yicheng Li, Li Chenglin, Hao Chen, Fei Yu, and Yin Zhang. 2024. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842.

Clifford Geertz. 2008. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge.

Michael D Hills. 2002. Kluckhohn and strodtbeck's values orientation theory. *Online readings in psychology and culture*, 4(4):3.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Yufei He, Jun Xia, Zhengyu Hu, Yulin Chen, Xihong Yang, Jiaheng Zhang, Stan Z. Li, Hui Xiong, and Bryan Hooi. 2025. Guardreasoner: Towards reasoning-based llm safeguards.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781.

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.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Viet Pham, Shilin Qu, Farhad Moghimifar, Suraj Sharma, Yuan-Fang Li, Weiqing Wang, and Reza Haf. 2024. Multi-cultural norm base: Frame-based norm discovery in multi-cultural settings. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 24–35.

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

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of 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)*, pages 2373–2403.

Paul Ricoeur. 1973. Ethics and culture. *Philosophy Today*, 17(2):153–165.

Sonia Roccas and Lilach Sagiv. 2010. Personal values and behavior: Taking the cultural context into account. *Social and Personality Psychology Compass*, 4(1):30–41.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning.

John J Tilley. 2000. Cultural relativism. *Human rights quarterly*, 22(2):501–547.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2024. Investigating mysteries of cot-augmented distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6086.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Nuwa Xi, Yuhan Chen, Sendong Zhao, Haochun Wang, GongZhang GongZhang, Bing Qin, and Ting Liu. 2024. As-es learning: Towards efficient cot learning in small models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10686–10697.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and

Isabelle Augenstein. 2025. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *arXiv preprint arXiv:2508.08879*.

Jiachen Zhao, Zonghai Yao, Zhichao Yang, and Hong Yu. 2024. Large language models are in-context teachers for knowledge reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16470–16486, Miami, Florida, USA. Association for Computational Linguistics.

Tove Österman. 2021. Cultural relativism and understanding difference. *Language & Communication*, 80:124–135.

A Distributional Effects of Filtering

As shown in Figure 4, approximately 1–2% of the samples were filtered across different cultural and topical categories, with only minor variation in filtering rates. Although categories such as LA and Business exhibit slightly higher filtering rates, these differences are negligible in absolute terms. This suggests that the teacher model's reasoning failures are not disproportionately concentrated in specific categories and that the filtering process does not meaningfully distort the overall training data distribution.

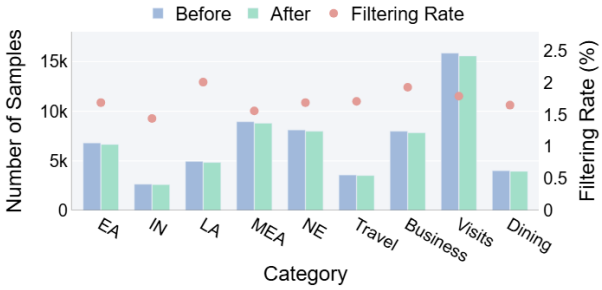

Figure 4: Distributional changes before and after filtering

B Hyperparameters

Table 12 presents the hyperparameter settings used for model training.

Hyperparameter	Llama3.2-1B	Llama3.2-3B	Qwen2.5-3B	Qwen3-4B
Grad. Accum. Steps	1	1	1	1
Micro Batch Size	8	16	32	8
Learning Rate	4e-6	5e-6	1e-5	4e-6
Optimizer	Paged AdamW	Paged AdamW	Paged AdamW	Paged AdamW
Scheduler	Cosine	Cosine	Cosine	Cosine
Epochs	3	4	2	3
Weight Decay	0.001	0.001	0.001	0.001
Warmup Ratio	0.05	0.05	0.05	0.05
LoRA Rank	8	8	8	8
LoRA Alpha	16	16	16	16
LoRA Dropout	0.05	0.05	0.05	0.05
Target Modules	Attention, MLP	Attention, MLP	Attention, MLP	Attention, MLP

Table 12: Hyperparameter settings.

C Prompts

The prompts were designed to guide the teacher model in generating CURE-based rationales consistent with the attributes of the gold answers. Each prompt instructs the model to produce a reasoning path that justifies the given answer, as shown in Tables 13 and 14, and is accompanied by three human-curated few-shot examples.

System Prompt	In this task, you are given a social norm related to etiquette, along with a region that provides the context in which the norm should be interpreted. Do not include your own opinions or personal judgments. Assume that the norm is generally accepted within the specified culture. Do not discuss counterfactual or opposite situations. Only explain why the given norm is culturally accepted within the provided context. Your task is to generate a rationale that explains why this norm is considered culturally appropriate. Your task is to generate: - Behavior: The specific action described in the social norm - Value: The cultural value associated with the norm - Meaning: Explain how the behavior is perceived and interpreted within the culture - Analysis: Analyze how the behavior aligns with or reflects the identified cultural value ## Definition of a "Behavior": - An explicit action that may be sensitive from a cultural or normative perspective. - Clearly and concisely describes the specific act to which the social norm applies. - Identifies the concrete behavior evaluated by social norms within the cultural context. ## Definition of a "Value": - A culturally grounded principle that helps individuals determine what is considered "right" or "wrong" within a given society. - A value is a generalized and abstract form of a social norm, reflecting what is traditionally regarded as important or virtuous in that cultural context. - It explains what is considered meaningful, desirable, or appropriate in the region's traditions, customs, and social expectations. ## Definition of a "Meaning": - It explains how the given behavior is perceived and interpreted by people within the cultural context. - It interprets what social meaning the specific behavior holds in light of the culture's values, customs, and expectations. ## Definition of a "Analysis": - Based on its alignment with cultural values, this explains logically why the given norm is socially acceptable within that culture. - It describes how the behavior is connected to the values considered important in the cultural context, and how it serves to realize or uphold those values.
User Prompt	### Region: {CULTURE} ### Statement: {STATEMENT}

Table 13: Prompt used to generate rationales when the gold answer is *positive*.

System Prompt	In this task, you are given a misleading or incorrect statement related to etiquette, along with a region that provides the cultural context in which the norm should be interpreted. However, the statement contradicts or misrepresents the actual etiquette norms of the specified culture. Assume that the behavior described is not generally accepted within that culture. Do not discuss counterfactual or opposite situations. Only explain why the given statement is culturally inappropriate in the provided context. Your task is to generate a rationale that explains why this statement is considered culturally inappropriate. Your task is to generate: - Behavior: The specific action described in the statement - Value: The cultural value that is contradicted by the behavior - Meaning: Explain how the behavior is perceived and interpreted within the culture - Analysis: Analyze how the behavior violates or conflicts with the identified cultural value ## Definition of a "Behavior": - An explicit action that may be sensitive from a cultural or normative perspective. - Clearly and concisely describes the specific act to which the social norm applies. - Identifies the concrete behavior evaluated by social norms within the cultural context. ## Definition of a "Value": - A culturally grounded principle that helps individuals determine what is considered "right" or "wrong" within a given society. - A value is a generalized and abstract form of a social norm, reflecting what is traditionally regarded as important or virtuous in that cultural context. - It explains what is considered meaningful, desirable, or appropriate in the region's traditions, customs, and social expectations. ## Definition of a "Meaning": - It explains how the given behavior is perceived and interpreted by people within the cultural context. - It interprets what social meaning the specific behavior holds in light of the culture's values, customs, and expectations. ## Definition of a "Analysis": - Based on its contradiction with cultural values, this explains logically why the statement is socially unacceptable within that culture. - It describes how the behavior violates, undermines, or disregards values considered important in the cultural context, and why such misalignment leads to social disapproval.
User Prompt	### Region: {CULTURE} ### Statement: {STATEMENT}

Table 14: Prompt used to generate rationales when the gold answer is *negative*.