# Can LLMs Grasp Implicit Cultural Values? Benchmarking LLMs' Metacognitive Cultural Intelligence with CQ-Bench

Anonymous ACL submission

## Abstract

Cultural Intelligence (CQ) refers to the ability to understand unfamiliar cultural contexts—a crucial skill for large language models (LLMs) to effectively engage with globally diverse users. While existing studies often focus on explicitly stated cultural norms, they fail to capture the subtle, implicit values that underlie real-world conversations. To address this gap, we introduce **CQ-Bench**, a benchmark specifically designed to assess LLMs' capability to infer implicit cultural values from natural conversational contexts. We generate multi-character conversation-based stories dataset using values from the *World Value Survey* and the *GlobalOpinions*, with topics including ethical, religious, social, and political. Our dataset construction pipeline includes rigorous validation procedures—incorporation, consistency, and implicitness checks—using GPT-4o, with 98.2% human-model agreement in the final validation. **CQ-Bench** consists of three tasks of increasing complexity: attitude detection, value selection, and value extraction. We find that while o1 and Deepseek-R1 models reach human-level performance in value selection (0.809 and 0.814), they still fall short in nuanced attitude detection, with F1 scores of 0.622 and 0.635, respectively. In the value extraction task, GPT-4o-mini and o3-mini score 0.602 and 0.598, highlighting the difficulty of open-ended cultural reasoning. Notably, fine-tuning smaller models (e.g., LLaMA-3.2-3B) on only 500 culturally-rich examples improves performance by over **10%**, even outperforming stronger baselines (o3-mini) in some cases. Using **CQ-Bench**[1], we provide insights into the current challenges in LLMs' CQ research and suggest practical pathways for enhancing LLMs' cross-cultural reasoning abilities.

[1]The code and dataset are available at https://anonymous.4open.science/r/CQ-Bench-508D.

## 1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in understanding and generating culturally relevant text (Li et al., 2024a,c,b; Putri et al., 2024). Prior research on cultural alignment in LLMs primarily focuses on modeling differences between national cultures or aligning model outputs with culturally specific norms (Pujari and Goldwasser, 2024; Kharchenko et al., 2024; Shi et al., 2024; Wang et al., 2024; Zhong et al., 2024; Rozen et al., 2024; Johnson et al., 2022). However, in real-world interactions, cultural values are not solely determined by demographic characteristics. Individuals within the same cultural group can hold diverse, nuanced beliefs (Fischer and Poortinga, 2012), and LLMs risk misinterpreting human perspectives if they rely solely on broad demographic generalizations. This limitation becomes particularly evident in human-AI interactions, where effective communication relies not only on an LLM's ability to recognize explicit cultural markers but also on its capacity to infer implicit cultural values. In human-AI interactions, personalization and customization are increasingly crucial. While LLMs can produce diverse responses when a persona and values are explicitly provided (Grassi et al., 2024), their effectiveness also depends on their ability to infer and align with human values even when such information is not directly given.

Cultural intelligence (CQ) refers to an outsider's ability to interpret unfamiliar and ambiguous cultural cues naturally (Blanchard and Mohammed, 2024; Earley and Ang, 2003). For LLMs, CQ is crucial for engaging in meaningful conversations with individuals from diverse backgrounds. Unlike traditional assessments of value understanding that focus on detecting explicit statements about cultural norms (Ren et al., 2024; Kiesel et al., 2023), real-life interactions require deeper contextual rea-

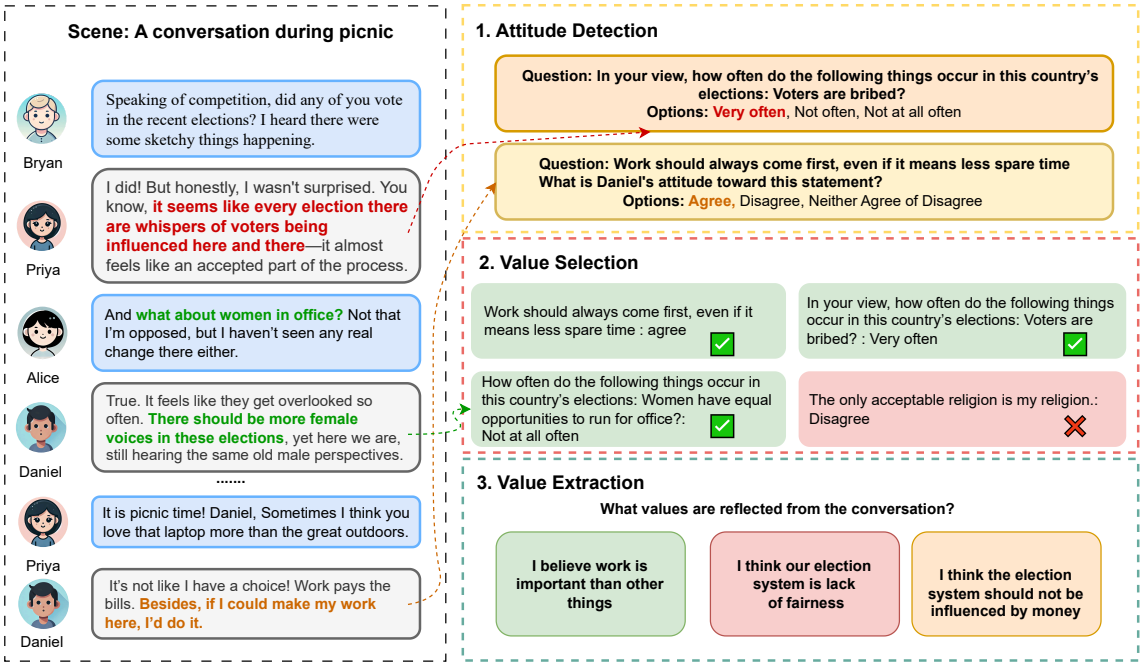


Figure 1: An illustration of **CQ-Bench**. We construct three distinct tasks based on conversation-style stories to assess the cultural intelligence of LLMs in **CQ-Bench**.

soning—understanding implicit beliefs embedded in everyday speech and actions. Humans do not typically express their values in a structured debate format; instead, values are subtly embedded into casual conversations and personal anecdotes. To enable LLMs to navigate these complexities, we need a more robust approach to evaluating and improving their ability to understand cultural values without relying on explicit cultural knowledge.

In this work, we introduce **CQ-Bench**, a benchmark designed to evaluate LLMs' ability to infer implicit cultural values within conversational contexts (Figure 1). We propose a structured methodology for generating multi-turn, multi-character conversation that naturally embeds cultural values. To ensure dataset quality, we perform automatic incorporation, consistency, and implicitness checks using GPT-4o. Human evaluation shows a 98.2% agreement with the model's validation. We evaluate various LLMs on three increasingly challenging tasks—attitude detection, value selection, and value extraction—to measure their ability to recognize, interpret, and extract cultural values from conversations. While larger models approach human-level performance, smaller models struggle with cultural reasoning without fine-tuning. Performance varies by category: models extract political values well (above 0.7 $F_1$ score overall) but worse in religious values (below 0.60). Notably, we find that fine-tuning on just 500 data points significantly enhances smaller models' ability to detect out-of-domain cultural values, demonstrating the efficiency to improve LLMs' capabilities with minimal data. Our contributions in this work can be summarized as:

- **Benchmark Design:** We introduce **CQ-Bench**, a novel benchmark for evaluating LLMs' cultural intelligence by assessing their ability to infer implicit cultural values from conversational contexts.
- **Comprehensive Evaluation:** We construct three escalating tasks—attitude detection, value selection, and value extraction—to systematically evaluate LLMs' cultural reasoning across multiple domains and model scales.
- **Data-Efficient Enhancement:** We demonstrate that fine-tuning with just 500 culturally rich examples significantly improves performance on smaller models, highlighting an efficient strategy for boosting cultural intelligence.

## 2 CQ-Bench Design and Implementation

### 2.1 Data Generation

Our goal is to generate conversation-based stories featuring 4–5 characters, with cultural values implicitly embedded in the narrative. We begin by out-

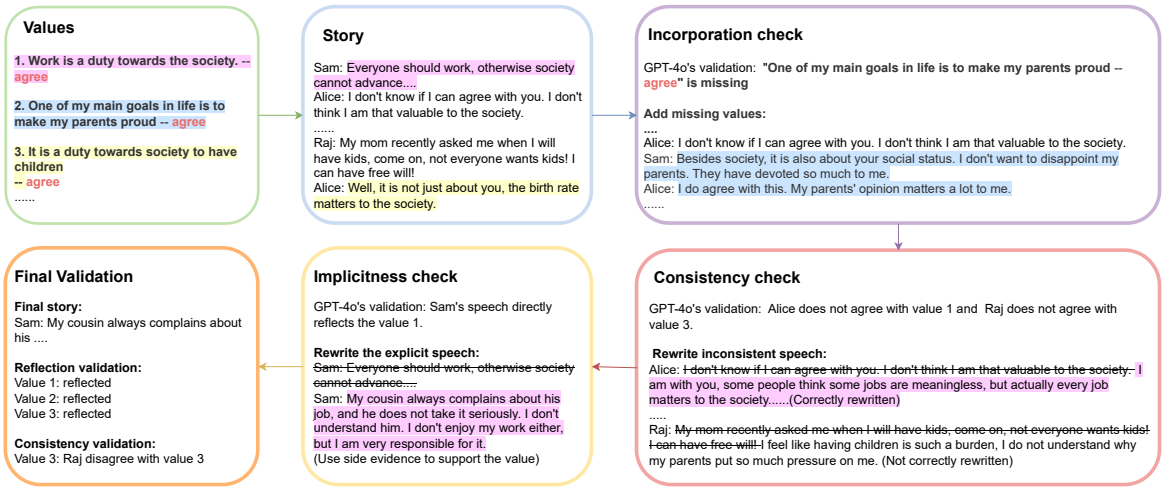


Figure 2: Dataset construction pipeline. We first create value sets, and then generate multi-character conversation style story. We conduct multiple checks and refinement to improve the quality of the story.

lining the process of value selection, followed by the story generation approach. Finally, we describe the validation steps taken to ensure the quality of the resulting dataset.

**Cultural Value Selection** Cultural values are defined as values inherently linked to culture and expressed through distinct attitudes. Each cultural value consists of two components: (1) **Statement**, which presents or solicits an opinion, and (2) **Attitude**, which signifies agreement or disagreement with the statement. As shown in Figure 1, *"The only acceptable religion is my religion"* is a statement while *"Disagree"* is an attitude. Different statements offer multiple attitude options, aligning with the settings in the original questionnaires.

The statements in this study are sourced from the World Values Survey (WVS) (Haerpfer et al., 2022) and the GlobalOpinion dataset (Durmus et al., 2023). The WVS is a global research project that explores individuals' values and beliefs, how their evolution over time, and their sociopolitical implications. We manually select values that either focus on personal beliefs or characterize societal and community attributes. The GlobalOpinion dataset contains a subset of survey questions about global issues and public opinions, adapted from the WVS and the Pew Global Attitudes Survey.

In **CQ-Bench**, each story incorporates five cultural values randomly selected from a list generated in three distinct settings:

- **Random Setting**: Aligning with previous research (Li et al., 2024a), which utilizes a subset of 50 values from the WVS, we randomly select 5 statements from this subset and assign each a random attitude to form cultural values.
- **Category-Specific Setting**: We expand the subset to include 23–28 statements per category: political, religious, social, and ethical. To assess whether domain-specific focus enhances cultural value comprehension, we select five statements from one category at a time and assign a random attitude. The detailed statistics of category-specific values are shown in Appendix A.2.
- **Multiple Attitude Setting**: The first two settings require consistency in attitude reflection throughout the story. In contrast, this setting assigns each character a distinct attitude toward the selected values, fostering diverse perspectives and enhancing complexity. For example, in the first two settings, one value could be *Work is a duty towards the society – agree*. In the multiple attitude setting, the values could be *Alice: Work is a duty towards the society – agree* and *Raj: Work is a duty towards the society – disagree*.

**Story Generation** Our goal is to create a dataset of conversation-based stories that reflect cultural values. Different from Li et al. (2024b), where characters sequentially present their ideas in a debate format, conversations in our **CQ-Bench** adopt a more natural and casual real-life style. Each story is generated based on five specified cultural values and a predefined scenario setting (Appendix A.2). The generated stories must adhere to the following guidelines:

- **Flexible Value Incorporation:** The cultural val-

ues may appear multiple times and do not need to follow a strict sequence, ensuring a natural conversational flow.

- **Character Value Consistency:** All characters must consistently adhere to their assigned values without contradiction.
- **Implicit Value Representation:** The cultural values should not be explicitly stated or directly rephrased within the dialogue. The underlying cultural values should be challenging for humans to explicitly identify.
- **Appropriate Story Length:** The story should be of sufficient length, incorporating multiple rounds of character interactions.

**Post-Generation Checking** All generated stories undergo a post-check to ensure quality. The checklist corresponds to the first three guidelines, while the fourth is verified directly using word count. Through preliminary experiments, we find that GPT-4o is better in verification while GPT-4o-mini can perform as well as GPT-4o in refining.

- **Incorporation Check:** We use GPT-4o to verify the inclusion of all assigned values in the story. If any values are missing, GPT-4o-mini refines the story to ensure their natural integration into the dialogue. In the Multiple Attitude setting, the model must also assess whether specific characters embody certain values.
- **Consistency Check:** In both the Random and Category-Specific settings, all characters in the story are expected to adhere to the assigned value. However, in some cases, certain characters may express opposing views, particularly when the assigned value contradicts prevailing social norms. For instance, if the given value is "If women go to work, the children will suffer—agree," a character might challenge this stance. GPT-4o is used to check consistency. When inconsistencies are detected, we revise the conflicting speech to align with the assigned value. Specifically, we provide GPT-4o with the full story, along with the original speech, and prompt it to generate only the revised version that conforms to the intended value. We do not use GPT-4o-mini for revision, as we found it is unable to effectively rewrite inconsistent speech and tends to follow the original stance.
- **Implicitness Check:** While the story is required to be implicit, the model generates explicit speech, as shown in Figure 2 and Table 5. To maintain the story's difficulty, we use GPT-4o-mini to systematically validate and rewrite explicit speech into an implicit form. We show the statistics of speech refinement in Table 3.

To enhance reliability, we perform three rounds of incorporation and consistency checks as final validation. Missing and contradictory values are addressed using a majority vote approach based on three evaluations. Specifically, missing values are removed from the ground truth, while contradictory values are documented along with the specific inconsistencies and the characters exhibiting inconsistent speech. Documented contradictory values are used in generating dataset. More details are shown in Appendix A.2.

In particular, resolving inconsistencies in speech rewriting remains challenging, even when the model successfully detects contradictions. However, we find that detecting these issues is generally easier for the model than correcting them. We conduct human validation for incorporation and consistency checks on 50 samples, comparing the results with model validation. Using Cohen's kappa, we calculate an agreement score of 0.658 for the incorporation check. To further analyze alignment, we remove instances where the model assigned a label of 0 (not incorporated), ensuring that corresponding values are also excluded from ground truth values. Among the remaining cases, where the model consistently identifies incorporation, 98.2% align with human judgment, demonstrating the faithfulness of the model's validation. Full agreement scores are provided in Appendix A.3.

### 2.2 Task Definition

As shown in Figure 1, **CQ-Bench** includes three tasks with increasing difficulties: attitude detection, value selection, and value extraction.

**Attitude Detection (AD)** Given a story $S$ and a statement $T$, the model identifies the attitude expressed within $S$. In the **Random** and **Category-Specific** settings, it determines the overall attitude including all characters, excluding cases where contradictions arise. If a story contains inconsistent mentions of certain values, those values will be removed when constructing the dataset. In the *Multiple Attitude Setting*, the model identifies the attitude of a specific character. The model selects an answer from a predefined set of options, making this a multiple-choice task.

**Value Selection (VS)** Given a story $S$ and a predefined set of 15 candidate values $\mathcal{V} =$

$\{v_1, v_2, \ldots, v_{15}\}$, the model is required to select exactly $X$ ground-truth values, denoted as $\mathcal{V}^* \subset \mathcal{V}$, where $|\mathcal{V}^*| = X$. The remaining $15 - X$ options are randomly sampled from non-ground truth values. This task is more challenging than attitude detection, as the model must first identify the relevant topics before selecting the correct values. Formally, the model must learn a function $f(S, \mathcal{V}) \rightarrow \mathcal{V}^*$, where $\mathcal{V}^* \subset \mathcal{V}$ and $|\mathcal{V}^*| = X$.

**Value Extraction (VE)** Given a story $S$, the model is asked to extract key cultural values on given topics without predefined answer choices (*e.g.*, social, ethical, political, *etc.*). The model is provided with examples illustrating the expected format of values, along with a set of topics $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ that cover all seed values in the prompt. For each topic $t_i \in \mathcal{T}$, the model generates a set of values $\mathcal{V}_{t_i}$, where:

$$\mathcal{V}_{t_i} = \begin{cases} \{v_1, v_2, \ldots, v_m\}, & \text{if relevant values exist} \\ \emptyset, & \text{otherwise} \end{cases}.$$

While the model is encouraged to provide comprehensive and detailed responses during the reasoning phase, we ask it to limit the answer size to 10 total values for easier performance comparison across models. Evaluation is based on recall, measuring the proportion of ground-truth values correctly identified.

## 3 Evaluating LLMs with CQ-Bench

### 3.1 Experimental Settings

**Models selection.** We select a diverse range of models, including both open-source and closed-source models, with varying sizes. For open-source models, we include Qwen 2.5 (7B, 14B, and 32B) (Bai et al., 2023), LLaMA 3.1 8B, and LLaMA 3.2 3B (Grattafiori et al., 2024), Deepseek-V3 and Deepseek-R1 (DeepSeek-AI, 2025). Additionally, we experiment with DeepSeek-Distill models (Qwen 2.5 1.5B, Qwen 2.5 7B, LLaMA 3.1 8B), which have been reported to achieve remarkable performance on math and coding tasks, even at smaller scales. Our goal is to evaluate if they can also perform well on culture-related reasoning tasks. Finally, for closed-source models, we select GPT-4o-mini, o3-mini, o1, o4-mini and o3 (Jaech et al., 2024). We also conduct analyses with four human participants who complete the same tasks on 25 stories, which took 5-6 hours in total [2].

[2]Hourly payment for annotators in this work is $16.5.

**Evaluation metrics.** For attitude detection and value selection, we use the F1 score to compare the predicted answers with the ground truth values. For value extraction, which is more open, we employ LLM-as a judge (Zheng et al., 2023) to evaluate the responses. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of ground truth values, and let $\hat{V} = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m\}$ be the set of predicted values. We use GPT-4o to access the output. We ask humans to conduct the same evaluation as GPT-4o on 25 stories. The agreement score is 0.864, which shows the reliability of LLM-as-a-judge.

For each ground truth value $v_i$, we define the score function as follows:

$$S(v_i, \hat{V}) = \begin{cases} 1, & \text{if } v_i \text{ is fully presented in } \hat{V} \\ 0.5, & \text{if } v_i \text{ is partially presented in } \hat{V} \\ 0, & \text{if } v_i \text{ is not mentioned in } \hat{V} \end{cases}.$$

We define "partially presented" as the output that mentions the same topic but is not detailed enough.

**Dataset.** We generate 500 stories for the random setting, 100 stories for each category-specific setting, and 100 stories for the multiple-attitude setting. For story generation, we use GPT-4o-mini, while GPT-4o is used for validation. In the multiple-attitude setting, we implement only the attitude detection task. The value selection/extraction task is not included because the ground truth values typically exceed ten, making it challenging for the model to choose accurately. We report the total datapoints for each task in Table 4.

**Summarize-then-analyse long CoT prompting.** For attitude detection and value selection, we experiment with two settings: no reasoning and CoT reasoning. In the no reasoning setting, we ask models to provide answers directly without explanation. In the reasoning setting, we provide a step-by-step reasoning guideline to guide the models' responses. For attitude detection (AD), we instruct models to first summarize speech relevant to the given statement and then analyze the attitude based on the retrieved speech. For value selection (VS), we employ a multi-step approach: (1) The model summarizes the topics mentioned in the story based on the provided options; (2) Selects values associated with the identified topics. (3) Reasons about which value best reflects the story. (4) Finally, outputs the selected value. For value extraction (VE), we ask the model to summarize the content for each topic and then predict relevant values based on the summarization.

| | | Qwen | | | Llama | | Deepseek-distill | | | GPT | | | | | Deepseek | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **7B** | **14B** | **32B** | **8B** | **3B** | **Q 1.5B** | **Q 7B** | **L 8B** | **4o-mini** | **o3-mini** | **o1** | **o4-mini** | **o3** | **V3** | **R1** |
| AD | W/O R | 0.529 | 0.553 | 0.572 | 0.527 | 0.455 | - | - | - | 0.604 | 0.622 | 0.622 | 0.60 | 0.689 | 0.642 | 0.595 |
| | W/ R | 0.620 | 0.616 | 0.624 | 0.506 | 0.372 | 0.381 | 0.484 | 0.556 | 0.639 | 0.661 | 0.622 | 0.595 | 0.622 | 0.660 | 0.635 |
| | Merged | 0.783 | 0.778 | 0.786 | 0.631 | 0.480 | 0.490 | 0.622 | 0.705 | 0.820 | 0.793 | 0.811 | 0.834 | 0.824 | 0.837 | 0.837 |
| | Multiple | 0.592 | 0.621 | 0.642 | 0.590 | 0.462 | 0.403 | 0.510 | 0.584 | 0.645 | 0.684 | - | 0.738 | - | 0.691 | - |
| VS | W/O R | 0.515 | 0.585 | 0.633 | 0.421 | 0.272 | - | - | - | 0.639 | 0.759 | 0.810 | 0.828 | 0.830 | 0.780 | 0.798 |
| | W/ R | 0.374 | 0.607 | 0.717 | 0.274 | 0.1 | 0.383 | 0.411 | 0.418 | 0.576 | 0.779 | 0.809 | 0.820 | 0.710 | 0.819 | 0.814 |
| VE | | - | - | 0.629 | - | - | - | - | - | 0.602 | 0.598 | 0.610 | 0.696 | 0.732 | 0.704 | 0.736 |

Table 1: Results on Attitude Detection (AD) and Value Selection (VS). For deepseek-distill models, the models always output their thinking process i.e. reasoning. Therefore, we only report reasoning results for those models "W/ R". For larger models like Deepseek-R1, o1, and o3, we report results on the same subset used for human evaluation, due to the high cost of running them on the full dataset. In the "Merged" setting, we merge similar options and in "Multiple" setting, story characters may hold different opinions. The "Merged" and the "Multiple" settings are both under reasoning settings.

### 3.2 CQ across Different LLMs

We first show the results of attitude detection and value selection in Table 1. Overall, larger models outperform smaller models by a lot and **Summarize-then-analyse** prompting can improve performance generally. Human participants achieve an average score of 0.689 and 0.765 on AD and VS respectively.

**The model struggles to detect nuanced attitudes beyond simple binary labels.** Although the attitude detection task should be easier than value selection, its scores are even lower. One reason is that models struggle to distinguish neutral stances, such as *neither agree nor disagree*. While they can easily differentiate between *agree* and *disagree*, the presence of a neutral option can cause confusion. Even when a model correctly identifies *agree*, it may be distracted by the neutral choice and incorrectly select *neither agree nor disagree*. Additionally, models find it challenging to differentiate between varying levels of severity, such as *not often* and *not at all often*. Although it is also challenging for humans, humans can do better in identifying nuanced attitudes than models. We present results after merging options of varying levels of severity in Table 1. Options are shown in Appendix A.2.

**Smaller models fail on long CoT reasoning in cultural intelligence task.** Although CoT reasoning significantly enhances performance in large models, smaller models often struggle with long CoT reasoning. LLaMA models, in particular, perform poorly in adhering to CoT reasoning across both AD and VS tasks. While Qwen 7B follows CoT reasoning well in AD, its performance declines significantly in VS as the reasoning steps become longer. A manual inspection reveals that its final outputs often consist of random values or irrelevant phrases that fail to focus on the given options. The DS-distill models exhibit slightly better instruction-following capabilities, outperforming Qwen 7B and LLaMA 8B in VS. While they do not strictly adhere to the prescribed format, their reasoning process generally aligns with the ideas provided in the prompt. However, their CQ reasoning ability remains weaker than their mathematical reasoning skills, resulting in final scores that are still lower than in the no-reasoning setting.

**Stronger models do not necessarily outperform weaker models in VE.** Value extraction requires strong reasoning and summarization capabilities, and we find that smaller models often struggle with this task, frequently producing nonsensical outputs. As a result, we focus our evaluation on 5 models: Qwen 2.5 32B, GPT-4o-mini, o3-mini, o4-mini, and DeepSeek-V3. We also evaluate a smaller subset of 25 stories using o1, o3 and DeepSeek-R1. Interestingly, unlike attitude detection and value selection—where larger models consistently outperform smaller ones—we observe that a weaker model (Qwen 2.5 32B) can outperform stronger ones like o3-mini and o1. One possible explanation is that current CoT reasoning methods are not well-suited for open-ended generation; models tend to perform better when given predefined options. Nonetheless, DeepSeek-V3 and R1 outperform other models in VE.

### 3.3 Cultural Intelligence across Different Categories of Culture Values

We present results on category-specific datasets in Figure 3, showing the performance of six models in the no-reasoning setting. Complete results across all models and settings are provided in the

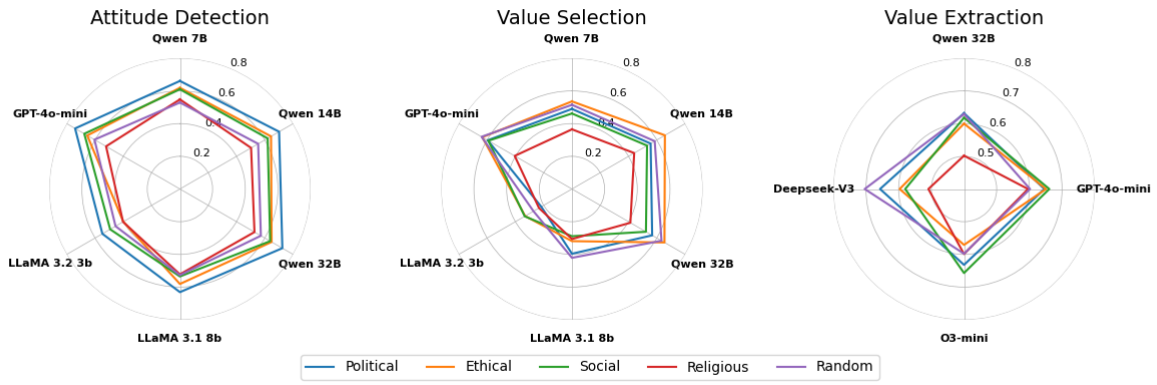


Figure 3: Category specific results. Overall, models perform worst in the Religious setting, and category-specific datasets yield higher scores than randomly sampled ones.

Appendix B.1. Overall, the results show that models perform better in understanding political, social, and ethical values, achieving performance that is better than or comparable to the random setting. However, they perform worse in the religious values domain across both tasks. Specifically, while models show stronger performance in attitude detection on the political dataset, they perform better in value selection on the ethical dataset. This suggests that it is easier for models to infer people's political stances but more challenging for them to identify the specific topics being discussed in political contexts. In contrast, in the ethics domain, models find it easier to identify the main topics being discussed. For value extraction, overall, models perform better on category-specific results except for Deepseek-V3, as they only need to cover a single topic. Similarly, models performed worse in extracting religious values.

### 3.4 Improving Cultural Intelligence on Smaller Models

In Section 3.2, we mention that smaller models struggle with long CoT reasoning. Since these models do not follow instructions well, we initially experiment with few-shot prompting to assess whether it can enhance their reasoning ability. However, because each demonstration is lengthy, we use one-shot prompting instead. The results, shown in Figure 4, indicate that while one-shot prompting improves instruction following and enhances performance on some datasets, it does not necessarily improve reasoning ability. In religious dataset AD task, the performance of Qwen 7B one-shot reasoning even decreases.

**CQ ability distillation for small models.** To further address this issue, we apply supervised fine-tuning on smaller models. Specifically, we distill the reasoning process from o3-mini into smaller models using the random setting dataset consisting of only 500 training samples. After training with LoRA (Hu et al., 2021) for five epochs, we evaluate the models on a category-specific dataset. Since the category-specific dataset contains values not present in the random dataset, it can be considered an out-of-domain dataset approximately. This allows us to assess whether the model has truly learned the reasoning process rather than merely memorizing the values themselves. We fine-tune four models: Qwen-7B, Qwen-14B, LLaMA 3.1-8B, and LLaMA 3.2-3B, and present the full results alongside o3-mini in Figure 4. Except for the ethical dataset, where SFT results show a slight drop, supervised fine-tuning significantly improves performance across all other datasets and models, particularly for LLaMA 3.2-3B. Notably, LLaMA 3.2-3B even slightly outperforms o3-mini in political and religious datasets for attitude detection. Similarly, Qwen-14B surpasses o3-mini in political and religious datasets for value selection. **These results demonstrate that SFT can effectively enhance the cultural intelligence of smaller models by distilling knowledge from larger models.**

**Qualitative analysis.** We also conducted a qualitative analysis to examine the specific aspects in which SFT improves reasoning outputs. After manually reviewing 50 samples, we find out SFT mainly improve these three problems:

- **Inconsistency**: Discrepancies between the reasoning process and the final answer.

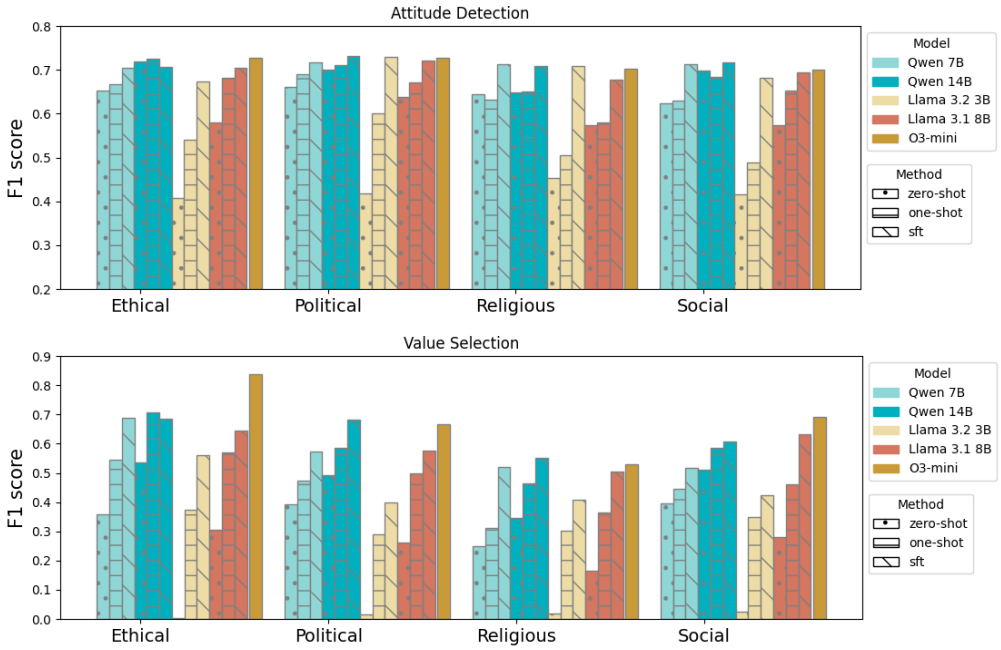


Figure 4: Results of zero-shot/one-shot prompting and SFT across different categories. Overall SFT can significantly improve the performance, especially on smaller models.

- **Logical errors**: Instances where the reasoning exhibits clear logical flaws, such as linking unrelated concepts.
- **Overlooking details**: Cases where conclusions are drawn from general observations but crucial details in the conversation are missed.

Detailed examples are shown in Appendix B.2.

## 4 Related Work

**Culture-Aware LLMs** Culture-aware LLMs account for cross-cultural differences. Prior work has examined their cultural personas and consistency (Kharchenko et al., 2024; Rozen et al., 2024; Yao et al., 2024; Johnson et al., 2022; Saha et al., 2025), showing that prompting language influences expressed values (Zhong et al., 2024). To improve cultural awareness, researchers have proposed both single- and multi-culture models through dataset augmentation and alignment (Nguyen et al., 2023; Lin and Chen, 2023; Abbasi et al., 2023; Li et al., 2024a). New benchmarks and datasets further support cultural knowledge acquisition (Myung et al., 2024; Shi et al., 2024), improving performance on tasks like hate speech detection (Li et al., 2024a)

**Value understanding** Value understanding is key to effective human-LLM interaction. Prior work has focused on detecting general social norms in short texts, often using classification or entailment-based methods (Ren et al., 2024; Kiesel et al., 2023; Zhang et al., 2024). Fung et al. (2022) introduces a method for extracting social norms from conversations, emphasizing norm mining rather than cultural value identification. In contrast, our work focuses on understanding cultural values within long, real-life conversations, contributing to the development of culture-aware LLMs in human-AI interaction.

## 5 Conclusion

We introduce **CQ-Bench**, a benchmark for evaluating LLMs' ability to infer implicit cultural values in conversations. Unlike prior work that focuses on whether LLMs possess cultural knowledge or interpret value from a short text, **CQ-Bench** assesses their cultural reasoning through attitude detection, value selection, and value extraction. Our findings show that LLMs, including state-of-the-art models, struggle with nuanced cultural understanding. Fine-tuning on just 500 examples notably boosts smaller models, suggesting cultural reasoning can be efficiently distilled. **CQ-Bench** exposes gaps in LLMs' cultural adaptability and serves as a foundation for advancing culturally intelligent AI. Future work can build on this to enhance LLM alignment with diverse human values.

## Limitations

This study has several limitations. First, although we construct a multiple-attitude dataset, we only evaluate models on the attitude detection task. Future work should explore how well models can extract individual characters' values within conversations—an essential capability for multi-agent interactions. Second, while we rewrite stories to remove explicit value expressions, the quality of these rewrites is inconsistent; detecting explicit speech is significantly easier for models than generating high-quality implicit alternatives. Third, in the story generation, models sometimes struggle to distinguish between nuanced options, making it difficult for humans to detect the intended attitudes as well. We also observe variability in cultural intelligence (CQ) among humans—while some can achieve up to 80% accuracy in attitude detection, others perform closer to 50%. Finally, in our qualitative analysis of model reasoning, we attempted to use GPT-4o to automatically detect reasoning flaws but found it inadequate for this task. As a result, we relied on manual inspection for a small subset of examples. Future research should investigate automated methods for identifying reasoning errors.

## Ethics Statements

All data used in this study were synthetically generated by large language models and do not contain any real user conversations or personal information. Cultural value statements were sourced from publicly available, anonymized survey instruments, including the World Values Survey and GlobalOpinion datasets.

## 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*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Emmanuel G. Blanchard and Phaedra Mohammed. 2024. On Cultural Intelligence in LLM-Based Chatbots: Implications for Artificial Intelligence in Education. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education*, volume 14829, pages 439–453. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

P Christopher Earley and Soon Ang. 2003. *Cultural intelligence: Individual interactions across cultures*. Stanford University Press.

Ronald Fischer and Ype H Poortinga. 2012. Are cultural values the same as the values of individuals? an examination of similarities in personal, social and cultural value structures. *International Journal of Cross Cultural Management*, 12(2):157–170.

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

Lucrezia Grassi, Carmine Tommaso Recchiuto, and Antonio Sgorbissa. 2024. Enhancing llm-based human-robot interaction with nuances for diversity awareness. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 2287–2294. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*.

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. *Preprint*, arXiv:2106.09685.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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 [cs].

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions. *arXiv preprint*. ArXiv:2406.14805 [cs].

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303.

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 [cs].

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 [cs].

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 [cs].

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. *Preprint*, arXiv:1510.03055.

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*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, and 1 others. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Rajkumar Pujari and Dan Goldwasser. 2024. LLM-Human Pipeline for Cultural Context Grounding of Conversations. *arXiv preprint*. ArXiv:2410.13727 [cs].

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 [cs].

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. ValueBench: Towards Comprehensively Evaluating Value Orientations and Understanding of Large Language Models. *arXiv preprint*. ArXiv:2406.04214 [cs].

Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. 2024. Do LLMs have Consistent Values? *arXiv preprint*. ArXiv:2407.12878 [cs].

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can llms identify cross-cultural communication gaps? *Preprint*, arXiv:2502.09636.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. *arXiv preprint*. ArXiv:2404.15238 [cs].

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

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.

Jing Yao, Xiaoyuan Yi, and Xing Xie. 2024. CLAVE: An Adaptive Framework for Evaluating Values of LLM Generated Responses. *arXiv preprint*. ArXiv:2407.10725 [cs].

Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. 2024. ValueDCG: Measuring Comprehensive Human Value Understanding Ability of Language Models. *arXiv preprint*. ArXiv:2310.00378 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Qishuai Zhong, Yike Yun, and Aixin Sun. 2024. Cultural Value Differences of LLMs: Prompt, Language, and Model Size. *arXiv preprint*. ArXiv:2407.16891 [cs].

# A Dataset

## A.1 Value Set

We adopt seed values from the World value survey(WVS) and Global Opinion. WVS is a global research project that explores people's values and beliefs and how they change over time. The values cover different topics from personal beliefs to political stance. The results include over 200 values and 100 countries. Each value has different option candidates and we provide possible option sets in Table 2.

## A.2 Story Generation

For the random dataset, we start with 50 statements covering seven topics: social, migration, security, science and technology, religious, ethical, and political, following the categorization defined by Li et al. (2024a). To expand the dataset, we focus on four categories—social, religious, ethical, and political—because the WVS and GlobalOpinion datasets contain more values in these areas. To generate coherent stories, we require at least 20 statements per category. We manually select values from WVS and GlobalOpinion, excluding those that do not fit our setting (e.g., "How many times do you go to church every week—everyday"). As a result, we collect 27 seed statements for social values, 23 for religious values, 24 for political values, and 28 for ethical values.

For the multiple attitude dataset, we use the same 50 statements as in the random setting. Each story involves four characters, and we assign one value to each character. Compared to the random dataset, which contains 5 values per story, the multiple attitude dataset includes $5 \times 4$ values. Due to the increased value space, we only conduct attitude detection on the multiple attitude dataset, as value selection becomes challenging when the ground-truth set is already large.

For each story, we will randomly predefined a scenario from those locations: company, school, neighborhood, national park, restaurant, amusement park, and airplane. We remove very short stories (less than 400 words). The length of stories ranges from 500 to 900 words.

## A.3 Story Validation

We conduct three validations: incorporation check, consistency check and implicitness check. For incorporation check and consistency check, which are directly related to the faithfulness of the dataset,

| Example Value | Options |
|---|---|
| Do you think that your country's government should or should not have the right to do the following: Keep people under video surveillance in public areas? | Definitely should not have the right<br>Probably should not have the right<br>Probably should have the right<br>Definitely should have the right |
| In your view, how often do the following things occur in this country's elections: Journalists provide fair coverage of elections? | Very often<br>Not often<br>Not at all often |
| Work is a duty towards society. | Agree<br>Neither agree nor disagree<br>Disagree |
| Apart from weddings and funerals, about how often do you pray? | Frequently<br>Occasionally<br>Never |
| Having a strong leader who does not have to bother with parliament and elections. | Very good<br>Very bad |
| How important is it for people to help others? | Important<br>Not important |

Table 2: Example values and their options. "Definitely should not have the right" and "Probably should not have the right" are similar options with different levels of severity.

we conduct human annotation on a small subset of 50 stories. We use Cohen's Kappa (McHugh, 2012) to calculate inter-annotator agreement. The agreement score for the incorporation check is 0.904. Since we remove missing values from the ground truth, the remaining values are those the model considered as incorporated. Among these, human annotators agree with 98.2% of them.

The agreement score for the consistency check is lower, at 0.3, because the model applies a stricter criterion for consistency. For values marked as inconsistent, we exclude them from attitude detection, as they may confuse the model during attitude prediction. Similarly, during value selection, we do not include statements expressing an opposite attitude in the candidate options, as they could interfere with the model's judgment. For values judged as consistent by the model, human annotators agree with 87.4% of them. We do not conduct consistency checks for the multiple attitude setting.

For the implicitness check, we compare the word count before and after rewriting to assess changes in length. We use Distinct-N (Li et al., 2016) to measure sentence diversity, which captures the number of distinct n-grams within a sentence. Finally, we compute semantic similarity using a Sentence Transformer (Thakur et al., 2021) between the value and both the original and rewritten speech. Ideally, the similarity score between the value and the rewritten speech should be lower, as the model is instructed not to mention the value explicitly. All statistics are reported in Table 3. The results show that refined speech is longer, more diverse, and semantically further from the value.

We generate 500 stories for random setting and 100 stories for each of the rest settings. We show how many datapoints for attitude detection and how many ground truth values in total for value selection in Table 4.

# B Results

## B.1 Category-Specific Results

The no reasoning results are shown in Figure 3. The zero-shot and one-shot reasoning of smaller models are shown in Figure 4. We show the rest of the results in Table 6. The social category dataset includes a single set of options: agree, disagree, and neither agree nor disagree. To study how well models understand the middle stance, we first remove questions where the ground truth is neither agree nor disagree, which eliminates about one-third of the data. We also remove the neither agree nor disagree option from the remaining examples, resulting in a fully binary dataset. On this binary version, o3-mini achieves an accuracy of 0.911. When we reintroduce the neither agree nor disagree option, performance drops to 0.811. However, when

| | | Random | Political | Social | Religious | Ethical | Multiple |
|---|---|---|---|---|---|---|---|
| Word count | Original | 19.94 | 20.21 | 20.04 | 18.95 | 19.60 | 16.60 |
| | Refined | 35.66 | 34.41 | 30.91 | 33.84 | 35.46 | 26.49 |
| Distinct-3 | Original | 0.877 | 0.878 | 0.876 | 0.860 | 0.875 | 0.850 |
| | Refined | 0.941 | 0.939 | 0.939 | 0.937 | 0.940 | 0.919 |
| Distinct-4 | Original | 0.815 | 0.817 | 0.814 | 0.797 | 0.812 | 0.775 |
| | Refined | 0.911 | 0.908 | 0.909 | 0.905 | 0.910 | 0.878 |
| Distinct-5 | Original | 0.755 | 0.755 | 0.752 | 0.742 | 0.750 | 0.702 |
| | Refined | 0.882 | 0.878 | 0.878 | 0.874 | 0.880 | 0.837 |
| Similarity | Original | 0.510 | 0.488 | 0.574 | 0.742 | 0.508 | 0.491 |
| | Refined | 0.433 | 0.408 | 0.473 | 0.355 | 0.415 | 0.425 |

Table 3: Statistics for explicit speech refinement show that the refined outputs exhibit greater linguistic diversity compared to the original speech.

| | Random | Political | Social | Religious | Ethical | Multiple |
|---|---|---|---|---|---|---|
| AD | 1665 | 301 | 213 | 425 | 270 | 1540 |
| VS | 2099 | 402 | 285 | 335 | 351 | - |

Table 4: Total datapoints for attitude detection and total values for value selection tasks.

evaluated on the full dataset—including questions with neither as the correct answer—o3-mini only achieves 0.700 accuracy. This suggests that the inclusion of a middle stance can significantly challenge the model's judgment.

### B.2 Distillation on Smaller Models

We use reasoning results generated by o3-mini to fine-tune smaller models. We use the LLamafactory framework (Zheng et al., 2024) and LoRA to accelerate the fine-tuning (Hu et al., 2021). We train 5 epochs with a GPU of A6000 for 2-3 hours. The rank of LoRA is 8, and the learning rate is 0.0001. We try multiple combinations of hyperparameters before we choose the final hyperparameter set.

We find out the reasoning without fine-tuning has several issues. The first is inconsistency, for example, the reasoning mentioned the usage of alcohol multiple times, but the final answer fail to include values related to alcohol. Or the reasoning does not mention divorce at all, but the final answer choose some values related to divorce. The second is logical errors, for example *"Oliver's preference for a traditional family setup for raising kids matches When jobs are scarce, employers should give priority to people of this country over immigrants–Neither agree nor disagree "*. There is no direct logic between two statements. The third is overlooking details. For example, the one-shot reasoning is "The discussion also involves ethical considerations regarding the use of animals and the environment, which is reflected in the option :Hunting animals as a sport should be banned.–disagree." The reasoning itself sounds reasonable, however, the gold label is "Using animals for entertainment like in zoos/circuses is ethical.–disagree." The SFT reasoning is *"Finally, Harold's statement on **using animals for entertainment***, which he finds shocking, aligns with the value that such practices are unethical." The difference between the two reasoning is that the latter one can catch the details of **using animals for entertainment**, while the former one only catch the use of animals and the environment.

## C Human Annotations

To compare the human performance with LLMs, we ask several human participants from CS undergraduates and master's students. We cover students from different background including 2 Chinese students and 2 Indian students. The instruction is shown below, and we show the exact prompts we use for models to human participants.

| Value | Original Speech | Rewritten Speech |
|---|---|---|
| In your view, how often do the following things occur in this country's elections: Voters are bribed?–Very often | Speech: Voter bribing happens so often , it feels like a norm at this point. Explanation: The speech basically rephrase the value without providing supporting evidence | But honestly, I wasn't surprised. You hear about the rumors and tales around voting—sometimes it seems like it's just part of the landscape now. |

Table 5: An example of implicitness check and rewriting.

| | | Qwen | | Deepseek-distill | | | GPT | Deepseek |
|---|---|---|---|---|---|---|---|---|
| | | **14B** | **32B** | **Q 1.5B** | **Q 7B** | **L 8B** | **4o-mini** | **V3** |
| AD | Political | 0.734 | 0.741 | 0.455 | 0.558 | 0.631 | 0.754 | 0.781 |
| | Social | 0.699 | 0.699 | 0.277 | 0.239 | 0.456 | 0.718 | 0.709 |
| | Ethical | 0.719 | 0.719 | 0.437 | 0.474 | 0.637 | 0.711 | 0.533 |
| | Religious | 0.649 | 0.631 | 0.431 | 0.506 | 0.637 | 0.592 | 0.574 |
| VS | Political | 0.574 | 0.587 | 0.372 | 0.206 | 0.402 | 0.56 | 0.694 |
| | Social | 0.567 | 0.618 | 0.305 | 0.314 | 0.323 | 0.568 | 0.715 |
| | Ethical | 0.674 | 0.741 | 0.366 | 0.341 | 0.389 | 0.64 | 0.815 |
| | Religious | 0.390 | 0.466 | 0.362 | 0.355 | 0.376 | 0.398 | 0.516 |

Table 6: Results on category-specific dataset under zero-shot reasoning setting.

**Human participant instruction**

You will participate in a task designed to evaluate cultural intelligence by assessing your ability to understand individuals' cultural values in conversations. Your responses will be used to compare with model performance. For each task, please read the story and questions carefully.
[Prompt for attitude detection] [Prompt for value selection]

## D Prompts

In this section, we show all the prompts we use for story generation, experiments, and validation.

**Story generation prompt**

You will be provided with 5 cultural values and a location where the conversation happens. Each of them follows the format [culture]–[value]. The first [culture] describes a statement or a situation, and [value] is how you agree with the culture or is the culture common or not.
Your task is to generate a scene including conversations and actions among multiple people and the scene needs to reflect the culture values provided.
Here are some requirements of the scene:
1. It cannot be too short. It should have multiple rounds of interaction among people.
2. It should not be too obvious. It cannot directly spit out or rephrase the values.
3. It cannot be easy to human to understand the culture values behind.
4. You do not need to follow the order of the values. You could mention the values multiple times through the conversation. Make sure the conversation flows well.
5. All the characters should follow the given values. There should not be contradictions between the character's value and the given value.
Here are the cultural values you should follow when generating: values
Here is the pre-defined location of the scene: location
Now using the cultural values to generate a story.

**Incorporation check prompt**

You will be provided a story and values reflected in the story. Your task is to check if the story reflects values? The story does not mention the values directly. You will need some reasoning to analyze the story.
Here is the original story: story
Here are the values to reflect in the story: values
For each value, output if the value is reflected and provide reasoning. In the end, output the values not reflected without the reasoning. Only output the exact and comprehensive value including "–" within it, do not rephrase! If all the values are reflected, just leave it blank. Follow the format: [Value]:[Reasoning, Yes/No] ..... Values not reflected: [Value]

**Missing values incorporation prompt**

You will be provided a story and values which need to be reflected in the story.
Your task is to refine the story to reflect the value provided. You cannot remove anything or replace existing speeches from the story, you can only add conversations to reflect the value.
The refinement should flow with original story well. You cannot add new conversation randomly. Here is the original story: story
Here are the values to reflect in the story: values
Now refine the story.

**Consistency check prompt**

You will be provided a story and values reflected in the story. Your task: for each value, check if all the characters agree with the value. If there is characters who does not agree with the value, you should output the character's name and his speech, and why the speech does not align with the value.
Here is the original story: story
Here are the values to reflect in the story: values
Now check the story and output if there is any contradiction. You can output reasoning to help you analyze. However, in the end, only output where is the contradiction one by one. If there isn't a contradiction, just reply NO. Otherwise, reply where is the contradiction.
Follow the format strictly, do not change the format, output exact values from the values provided, and do not rephrase:
[Reasoning]:
[Value–attitude]:[If all the speeches are aligned with the value]
[Contradictions]:
[Value–attitude]: [*character name*:speech] .....

**Consistency resolve prompt**

You will be provided a story and values which need to be reflected in the story. However, the story includes some contradiction where characters do not agree on certain values. You will be provided where is the contradiction.
The contradiction includes 3 parts:
1. Correct value to follow
2. Character name
3. Character speech
Your task is to replace the speech mentioned in the contradictions with a new speech to make sure the speech is aligned with the values The refinement should flow with the original story well. You cannot add new conversations randomly. Here is the original story: story
Here are the contradictions: contradiction
Ignore the original character's speech. Directly write a new speech that reflects the value.
Here is the rewritten speech:

**Implicitness check prompt**

You will be provided a story and values reflected in the story. Your task is to check if there is obvious speech that directly mentions or rephrases the values. If the story mentions phrases or sentences from values, that would be also counted as directly mentioned If it just reflects the value but does not rephrase the values, it is not considered as obvious speech. Here are some examples of obvious speech:
[Example 1]
Value: In your view, how often do the following things occur in this country's elections: Journalists provide fair coverage of elections—Very often Speech: "I think journalists do their best to cover things fairly" Explanation: The speech basically rephrases the value without providing supporting evidence
[Example 2]: .....
You should output places where the values are directly mentioned. If the values are not directly spit out, do not output them!!!
Here is the original story: story
Here are the values to reflect in the story: values
Now check if any parts of the stories directly mentioned the values. You can output reasoning or evidence and help you analyze. However, in the final answer, output speeches where the values are directly mentioned and the corresponding values. If there are no directly mentioned values, just output NO in the final answer.
Follow the format:
[Reasoning]: reasoning
[Final answer]: [value]:[speech]
[value]:[speech] .....

**Implicitness rewrite prompt**

You will be provided a story and values reflected in the story. You will be also provided with some comments where the values are directly mentioned in the story. Each comment includes a value and a speech where the speech directly reflects the value. Your task is to refine the speech by rewriting the places where they directly mention the values. Instead of speaking out values directly, characters should reflect values implicitly by using anecdotes. Imagine the place is not a safe place to speak out, characters need to respect each other and not offend others. You need to make sure after rewriting, people are able to detect values by the speech. It cannot be too abstract.
Here are some examples:
[Example 1]: Value: How frequently do the following things occur in your neighborhood: Robberies?–Not frequently Original speech: "Well, at least we're not dealing with robberies or anything in our area." Refined speech: "You know in other counties, cars got breaking in is common. At least we don't have to worry about that."
[Example 2]: .....
Here is the original story: story
Here are the values to reflect in the story: values
Here are the comments that you need to refer to when refining the story: comments
Please rewrite the speech one by one in the story. Do not miss any comments.
For each rewritten speech, you need to follow this format: Value:[value] Original speech: [Speech in the comment] Refined speech: [You rewritten speech]
Here is your output:

**Attitude detection prompt**

You will be given a scenario including background, conversation and actions in a location. You task is to detect the culture values reflected from the conversation. You will be provided a statement and you need to detect what attitude they hold to the value. For example, the statement is "The people who don't work turn lazy" and the options are "disagree, agree, neither disagree nor agree". And you need to read the story and choose one attitude from the options. The options could also describe something happening around them. For example, if the statement is "How often does this happen in your neighborhood:Racist behaviors", the options are "Very often, not often, not at all often", you need to choose one based on the story.
Here is the scenario: story
Here is the statement: statement
Here are the options: options
Now choose one option based on the story. Let's think step by step! First, extract the speech that is related to the statement.Then analyse the attitude of those people from the speech. Finally, output the answer.
Follow the format:
[Related speech]: speeches
[Analysis]: analysis of attitude
[Answer]:answer

**Value selection prompt**

You will be given a scenario including background, conversation and actions in a location. You task is to detect the culture values reflected from the conversation.
You will be provided 15 options and you need to select number correct answers from the options. (Only number options are correct).
You need to choose the whole option, for example:
Being a housewife is just as fulfilling as working for pay–agree
Make sure you include '–' in the answer.
Here is the scenario: story
Here are the options: options
To think step by step: 1. You need to first detect what topics the story mentioned according to the options. You need to list the speech related to the topic. 2. Then for each topic, detect what values might be related to the topic. Sometimes, you might find multiple options that might be correct, for uncertain options, you should compare them. 3. Based on the previous detection, analyze what values are aligned with the story, you should always prioritize those values that are strongly related. 4. In the end, output the final answers only. Only choose the required number of values.
Follow the format: [Topic]: topic1: speeches related to the topic .....
[Value detection]: topic1: what values are related to the topic
[Reasoning]: Reasoning
[Final answer]:
[text] – [text]
.....
"""

**Value extraction prompt**

You will be provided with a story. Your goal is to summarize the main themes in the story. Please provide 4-5 sentences to summarize the ideas.
You should be specifically looking for: topic themes. Please pay attention to the character's responses and attitudes towards topic.
Here is the story to summarize: story.
You will be provided with a story. Your goal is to identify the most prevalent topic values in the story. For example, one religious value could be: I strongly believe in God and the afterlife. The value should be a complete sentence, it should not be a phrase like work-life balance. The value should hold a attitude or it reflects social phenomena, it should not be a overview of topic. For example, one social value could be: I think work is a duty towards the society. And one political value could be: I think voters are bribed in our election system. Please identify the values in this story: story by paying attention to how the characters in the story discuss topic. You need to generate 10 values in total, make sure values are specific and detailed. Only focus on one topic in one value. Do not mention several themes in one value.
Here is the story: story
Here is a story summary with some key ideas: summary.
You could output reasoning before you output final answer. But in the end, your output should follow the format: [Final answer]: value1 value2 ....

**LLM evaluation prompt**

You will be given two sets of texts: a set of predicted values and a ground truth set of values. Your task is to determine how many of the ground truth values are fully represented in the predicted values. A ground truth value is considered correct if all of its components are meaningfully discussed in the predicted value, even if there is no exact 1-to-1 match. It could be many-to-1 match i.e. many values to 1 ground truth value. If the ground truth value is fully presented, score 1, if it is partially presented, score 0.5, if it is not mentioned at all, score 0. Additionally, provide a brief justification for your score, explaining which values were correctly or incorrectly represented, in the justification, you should explicitly mention which predicted values are related to the ground truth value. Here are the predictions: pred and ground truth: gt In the reasoning, if the ground truth value is fully represented, you need to point out which predicted value is related to it.
Return the results in this format: [Reasoning]:
[Ground truth 1]: [reasoning]
.....
[Final answer]:
[Ground truth 1]:1
[Ground truth 2]:0.5
.....