UNDERSTANDING AND EVALUATING CULTURAL NORM GROUNDED REASONING IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cultural intelligence in large language models (LLMs) requires not only *knowing* cultural norms but also *reasoning with* them under context. Yet existing benchmarks predominantly evaluate cultural knowledge recall, leaving open whether models can effectively *invoke* acquired norms in multi-constraint reasoning scenarios. We introduce `CultureForest`, a benchmark for *Cultural Norm Grounded Reasoning* that explicitly tests this ability. Each of its 5,378 examples is grounded in atomic cultural norms, enabling verifiable and attributable evaluation across 8 domains and 53 countries/regions. `CultureForest` supports progressive evaluation from multiple-choice (Easy) through binary judgment (Medium) to open-ended generation (Hard), accompanied by a lightweight verifier for scalable open-ended assessment. Extensive experiments reveal that even top-tier models degrade substantially in open-ended settings, with pronounced cross-region disparities. Through multi-faceted analysis, we uncover consistent patterns: (1) test-time reasoning yields limited gains and may exacerbate inequity; (2) models share convergent regional preference structures; (3) responses grow markedly conservative under stricter norms; and crucially, (4) by disentangling cultural knowledge acquisition from reasoning, we demonstrate that the primary bottleneck is not knowledge availability but its effective utilization during reasoning. These findings point to a necessary shift from knowledge-centric evaluation toward measuring knowledge-grounded reasoning in cultural scenarios.

1 INTRODUCTION

From cross-cultural communication to region-specific decision-making, cultural intelligence is increasingly critical for deploying large language models (LLMs) in real-world, globally diverse settings (Li et al., 2024a;b; Naous et al., 2024; Ye et al., 2024; 2026; Yuan et al., 2026). In such scenarios, models are expected to interact with users from diverse cultural backgrounds, where seemingly minor behavioral choices, such as forms of politeness, social etiquette, or taboo avoidance, can carry significant implications (Durham, 1976; Bulté & Rigouts Terryn, 2025).

To assess such capabilities, a growing body of work has introduced culture-related benchmarks. Prior efforts such as (Chiu et al., 2025; Palta & Rudinger, 2023; Srinivasan & Choi, 2022; Rao et al., 2025), primarily evaluate models' understanding of cultural knowledge through multiple-choice or classification tasks, focusing on domains such as everyday life, food, and politeness. More recent benchmarks, including (Hasan et al., 2025) and (Myung et al., 2024), extend evaluation to short-form question answering settings, while Fabbri et al. (2025) introduces reasoning tasks embedded in cultural contexts (e.g., traditional reasoning or math reasoning framed within cultural narratives). Beyond benchmarks, Shi et al. (2024) introduce CultureBank, a community-driven knowledge base from different online communities. A comparison of representative datasets is summarized in Table 1. Despite these advances, existing evaluations share two fundamental limitations:

(1) Existing evaluations largely reduce cultural intelligence to a *knowledge-level* problem, testing whether models can recall cultural facts, while overlooking a critical challenge: real-world cultural competence requires *integrating multiple norms simultaneously* to determine appropriate behavior. Can models effectively *invoke and compose* acquired cultural knowledge for reasoning? In practice, failures arise not from missing knowledge, but from misapplying it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison between CultureForest and existing other culture-related datasets.

Datasets	Contribution Type	Domain	Evaluation Focus	Open-Ended	# Country/Region	# Size
CulturalBench (Chiu et al., 2025)	Benchmark	Everyday-life	Cultural Norms / Knowledge	✘	45	1,696
FORK (Palta & Rudinger, 2023)	Benchmark	Food	Cultural Norms / Knowledge	✘	10	184
TyDiP (Srinivasan & Choi, 2022)	Benchmark	Politeness	Politeness Discrimination	✘	>=9	4,500
CultureBank (Shi et al., 2024)	Knowledge Base	Everyday-life	Norm Compliance Judgement	✔ (Long-form)	>50	∼23,000
NativQA (Hasan et al., 2025)	Benchmark	Everyday-life	Cultural Norms / Knowledge	✔ (Short-form)	9	∼64,000
BLEnD (Myung et al., 2024)	Benchmark	Everyday-life	Cultural Norms / Knowledge	✔ (Partial & Short-form)	16	∼52,600
MultiNRC (Fabbri et al., 2025)	Benchmark	Linguistic, Wordplay, Tradition, Math	Reasoning in Cultural Context	✔ (Short-form)	>=3	1,055
CultureForest (our)	Benchmark	Everyday-life	Cultural Norm Grounded Reasoning	✔ (Partial & **Long-form**)	53	5,378

(2) Most benchmarks predominantly rely on *closed-form evaluation* (e.g., multiple-choice or classification), which diverges from real-world open-ended scenarios and often exposes answer candidates in context, thereby simplifying the task and potentially inflating performance estimates.

To bridge these gaps, we introduce **CultureForest**, a benchmark for *Cultural Norm Grounded Reasoning*, designed to evaluate whether models can move beyond knowledge recall toward effective knowledge utilization. CultureForest is built upon the following key principles:

- **Broad Domain and Country/Region Coverage.** CultureForest spans 8 domains of everyday cultural practices and covers a diverse set of 53 countries/regions across major geographic areas, enabling systematic evaluation of cross-region disparities.
- **Norm-Grounded and Attributable Design.** Each question is grounded in a small set of atomic cultural norms, enabling **verifiable** evaluation and **attributable** analysis. Predictions can be traced to specific norms, making it possible to disentangle knowledge acquisition from reasoning.
- **Progressive Evaluation from Closed- to Open-Form.** CultureForest provides three difficulty levels (Easy, Medium, Hard) that progressively reduce external guidance, ranging from multiple-choice selection to open-ended generation. This setup enables a systematic evaluation of model behavior under increasingly realistic and unconstrained scenarios.
- **Scalable and Robust Open-Ended Evaluation.** To support evaluation in open-ended settings, we develop a lightweight verifier, achieving strong effectiveness and generalization. This enables consistent and scalable assessment of norm compliance beyond closed-form settings.

We conduct extensive experiments across diverse models and scales, revealing fundamental limitations in open-ended cultural scenarios with pronounced cross-region disparities. Through multi-faceted analysis, including *scaling effects (model size and test-time reasoning)*, *regional preference structures*, and *distributional of norm compliance*, we uncover systematic weaknesses hidden under conventional evaluation. By disentangling cultural knowledge acquisition from reasoning, we find that although modern LLMs possess substantial cultural knowledge, their performance is bottlenecked by its effective use. We therefore argue that advancing cultural intelligence requires moving beyond knowledge-centric evaluation toward measuring knowledge-grounded reasoning. CultureForest provides a diagnostic testbed for this direction **(see Appendix B Figure 6 for an full illustrative example)**.

2 CultureForest Benchmark

We formulate **Cultural Norm Grounded Reasoning** as the core task of CultureForest: given a contextualized scenario, models must reason about and apply acquired cultural knowledge to determine or generate appropriate behavior. To enable principled evaluation, each question is explicitly grounded in a small set of **atomic cultural norms**, ensuring that model behaviors are both **Verifiable** (objectively assessable) and **Attributable** (traceable to specific cultural norms).

2.1 Data Source and Norm Curation

We source cultural norms from Cultural Atlas (Mosaica, 2024), a platform providing expert-verified descriptions of societal attitudes and behaviors across diverse regions, grounded in both community-sourced insights and academic review. We reorganize the data into a structured format:

```
Norm Bundle = {"country": "Afghanistan", "domain": "Business Culture", "topic":
"Meetings", "norm": "Bargaining is an acceptable and common way of negotiating."}
```

Each question q_i in `CultureForest` is grounded in a set of atomic cultural norms $S_{q_i} = \{n_1, \ldots, n_L\}$, serving as its underlying knowledge basis. To ensure quality, diversity, and applicability, the norms within each S_{q_i} are selected according to the following principles:

- **Scenario Consistency.** Norms are restricted to a shared $\mathrm{Pool}(c, d, t)$ defined by identical country (c), domain (d), and topic (t), ensuring joint applicability within a coherent scenario.
- **Content Diversity.** To avoid semantic redundancy, each norm n is embedded[1] via:

  ```
  "Behavioral constraint imposed by the following cultural norm: {norm}"
  ```

  We then perform Gaussian Mixture Model (GMM) clustering and require selected norms to (i) belong to distinct clusters and (ii) achieve an average silhouette score above a threshold τ (defined as the pool-level mean), ensuring semantic heterogeneity.
- **Norm Strictness.** Cultural norms differ in their degree of strictness[2]. To quantify this property, we estimate the acceptability of violating each norm via the following classification task:

  ```
  "In {country}, is it ever acceptable to violate the following cultural norm? {norm}
  (Yes/Sometimes/No)"
  ```

  Aggregated responses yield *Yes*, *Sometimes*, and *No*. We discard norms universally labeled *Yes*, and require each S_{q_i} to include at least one strict (*No*) norm, ensuring non-trivial constraints and avoiding degenerate cases. (Implementations and validation results are provided in Appendix C)

2.2 Benchmark Construction

We construct `CultureForest` via an agentic QA generation pipeline with human verification, implemented using *GPT-4.1-2025-04-14*. Each question q_i is grounded in 3 atomic cultural norms, aligned with the *norm-option alignment matrix* in Figure 1.

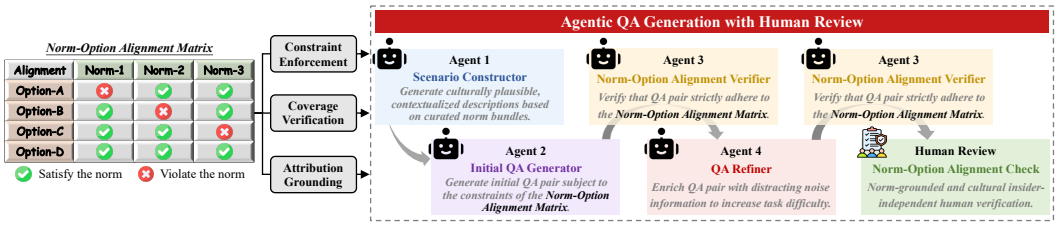


Figure 1: Overview of the norm-option alignment constraints and agentic QA Generation pipeline.

Answer options reflect different patterns of norm satisfaction/violation. Specifically, we enforce that each option corresponds to a distinct combination of satisfied and violated norms, ensuring that correct reasoning requires integrating all norms rather than relying on partial cues. To guarantee the faithfulness between options and their intended norm constraints, we employ an automatic *alignment checker*. The generation process further incorporates refinement and human validation to ensure both correctness and appropriate difficulty. Detailed **prompt designs**, **annotation procedures** are provided in Appendix D. The high **inter-annotator agreement** and exceptional **compliance rates** observed during manual inspection demonstrate the quality and reliability of `CultureForest`.

2.3 Difficulty Level Partition and Evaluation

We design 3 difficulty levels to enable evaluations under progressively unconstrained scenarios:

(1) Easy. Multiple-choice, models select the correct answer.

[1]Using `text-embedding-3-large` as the embedding model.

[2]For example, in Muslim communities, consuming pork is strictly prohibited; whereas in China, though using chopsticks is customary, using forks in certain contexts (e.g., steakhouse) is sometimes acceptable.

(2) Medium. Independent binary judgments assessing the appropriateness of each option. As shown in Figure 1, options A,B,C correspond to culturally inappropriate behaviors, while option D is appropriate. A prediction is considered correct only if all options are classified correctly.

For both **Easy** and **Medium** settings, we report Accuracy results.

(3) Hard. An open-ended generation setting where no options are provided. Models must produce responses that appropriately comply with the underlying cultural norms.

Verifier Model for Open-Ended Evaluation. We train a lightweight verifier (`C-Verifier`) based on *Qwen3.5-0.8B-Base*. Given an input tuple *(scene, response, norm)*, `C-Verifier` classifies the response into *Satisfy*, *Neutral*, or *Violate* based on its compliance with the specified norm. Training combines synthetic supervision from benchmark construction (Data_{syn}) and human-annotated responses from diverse models ($\text{Data}_{\text{real}}$). The training data construction process and annotation protocol are detailed in Appendix E.

Verifier Validation. We test `C-Verifier` on the valid set of Data_{syn} and conduct two generalization validations on $\text{Data}_{\text{real}}$: (i) leave-one-model-out (LOMO) K-fold cross-validation, holding out all responses from one model per fold across 7 models, and (ii) leave-one-domain-out (LODO) K-fold cross-validation, holding out an entire cultural domain (e.g., *Religion*, *Etiquette*) per fold across 8 domains. As shown in Appendix F Table 6 and Table 7, `C-Verifier` significantly outperforms strong LLM-as-a-judge baselines (*GPT-5.2*, *Gemini-3.1-Pro*, *Qwen3.5-Plus*), achieving the best accuracy on 6 of 7 held-out model subsets and 6 of 8 held-out domain subsets, demonstrating superior robustness and cross-model/domain generalization.

Verifier Error Type Analysis. To understand the boundaries and failure modes of `C-Verifier` versus commercial LLM judges, we analyze all pairwise disagreements on Data_{syn} (valid) in Appendix G Table 9. `C-Verifier` produces substantially fewer total disagreements (711 vs. 1,264–1,912 for proprietary models). More importantly, the two exhibit fundamentally different error profiles: `C-Verifier`'s primary error mode is *over-leniency* (68.9% of errors classify violations as Neutral or Satisfy), whereas commercial LLM judges exhibit a distinct *hedging bias*, with 54–58% of their predictions diluting to Neutral. This reveals that smaller specialized verifiers tend to over-accept, while larger general-purpose models avoid committing to definitive cultural judgments.

Scoring Function. For each generated response x_q to question q and a given cultural norm $n \in S_q$, the `C-Verifier` produces logits over three labels $\mathcal{Y} = \{\textit{Satisfy}, \textit{Neutral}, \textit{Violate}\}$. We convert these logits into a probability distribution via softmax, denoted as $p(y \mid x_q, n)$ for $y \in \mathcal{Y}$. We then define a utility function $w : \mathcal{Y} \to \mathbb{R}$ to quantify norm compliance. *Satisfy* and *Violate* are assigned reward $+1$ and penalty -1, respectively, while *Neutral* is treated as partial compliance and assigned a moderate penalty $-\lambda$, where $\lambda \in (0, 1)$:

$$w_{\text{No}}(y) = \begin{cases} +1, & y = \textit{Satisfy}, \\ -\lambda, & y = \textit{Neutral}, \\ -1, & y = \textit{Violate}. \end{cases} \quad w_{\text{Sometimes}}(y) = \alpha \cdot w_{\text{No}}(y), \tag{1}$$

To account for different levels of norm strictness (§2.1), we scale the utility for *Sometimes* (sometimes acceptable to violate) norms, where $\alpha \in (0, 1)$ controls the relative strength.

Given a question $q \in \mathcal{Q}$ associated with a set of norms $S_q = \{n_1, n_2, n_3\}$, we compute the expected utility for each norm and average across norms, where w_n is determined by the norm type (*No* or *Sometimes*). And the score is obtained by averaging over all questions:

$$\text{Score}(q) = \frac{1}{|S_q|} \sum_{n \in S_q} \sum_{y \in \mathcal{Y}} w_n(y) \cdot p(y \mid x_q, n), \quad \text{Score}_{\text{open}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{Score}(q). \tag{2}$$

To ensure comparability, we linearly normalize scores into $[-100, 100]$ using theoretical bounds.

Sensitivity Analysis. We perform a grid search over $\lambda, \alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and evaluate robustness via pairwise rank correlations under different parameter choices, including both model-level and country-level rankings. As shown in Appendix H Figure 8, consistently high correlations indicate that our conclusions and rankings are stable across parameter configurations. We therefore adopt $\lambda = \alpha = 0.5$, as this configuration achieves the highest overall correlation.

Table 2: Main results on CultureForest across three levels. ★ indicates models used in the data construction pipeline. **Bold** denotes the best performance, and underline denotes the second-best.

Models	Easy				Medium				Hard			
	Performance ↑		Disparity ↓		Performance ↑		Disparity ↓		Performance ↑		Disparity ↓	
	Mean@3	Pass@3	Std	Gap	Mean@3	Pass@3	Std	Gap	Mean@3	Pass@3	Std	Gap
Top-Tier Non-Reasoning Models												
★ GPT-4.1-2025-04-14	94.38	95.50	4.12	20.87	40.15	47.62	10.33	50.44	52.15	62.08	16.87	80.85
GPT-5.2-2025-12-11	92.35	93.77	3.67	14.83	**53.25**	60.21	9.31	46.45	51.29	62.44	17.28	82.61
Deepseek-V3	93.97	94.79	3.54	18.02	25.02	27.95	7.71	29.48	50.73	60.41	17.19	82.32
Llama-3.3-70B-Instruct	95.97	96.28	2.75	11.27	31.42	33.23	9.97	39.82	34.33	45.99	20.76	89.46
Llama-4-Scout-17B-16E-Instruct	92.74	94.18	3.68	15.24	44.37	46.84	8.77	39.91	33.38	45.49	19.39	87.45
Claude-Sonnet-4	95.23	95.63	3.08	15.90	52.83	54.70	10.56	48.24	48.03	58.33	16.80	79.22
Qwen2.5-72B-Instruct	92.96	94.29	3.83	15.00	27.71	31.37	8.54	34.59	42.26	54.38	17.70	87.48
Top-Tier Reasoning Models												
OpenAI-o3	91.51	94.31	4.49	17.53	50.61	62.33	9.89	42.71	59.89	71.02	14.70	72.31
Deepseek-V3.2-Thinking	93.65	96.04	4.32	18.97	42.44	53.61	8.98	37.06	47.91	59.03	17.62	82.49
Gemini-3.1-Pro-Preview	91.58	94.66	6.15	29.21	52.22	62.03	11.36	55.36	56.37	66.11	16.14	75.66
Claude-Sonnet-4-Thinking	94.29	96.91	3.35	14.36	49.28	**62.38**	9.12	37.58	45.48	56.88	17.20	79.19
Claude-Sonnet-4-6-Adaptive	**96.31**	**97.23**	3.48	16.92	50.64	58.01	9.94	41.39	57.14	66.22	15.58	79.76
Qwen3.5-Plus-Thinking	94.78	96.49	4.23	20.51	51.40	60.75	10.29	47.28	52.81	62.66	17.74	78.20
Qwen3.5-27B-Thinking	94.57	96.65	3.85	20.51	50.77	61.94	9.84	50.59	49.41	60.30	19.07	78.18
Smaller Models												
Llama-3.1-8B-Instruct	80.84	90.91	5.86	29.87	27.28	39.01	8.49	44.25	26.75	42.44	20.77	91.21
Mistral-7B-Instruct-v0.3	82.32	90.01	5.58	21.24	3.53	5.52	3.06	17.46	34.99	50.03	22.36	96.03
Phi-3.5-mini-Instruct	80.41	82.82	5.69	26.33	0.47	0.56	0.85	3.57	28.00	40.57	24.21	101.26
Qwen3.5-0.8B-Thinking	58.64	86.33	4.82	19.32	7.28	19.02	3.75	17.28	18.11	33.93	22.35	100.87
Qwen2.5-7B-Instruct	85.61	87.28	5.04	20.29	16.35	18.69	8.12	44.68	31.25	45.25	20.57	91.02

2.4 Benchmark Statistics

CultureForest comprises 5,378 instances spanning 8 domains and 53 countries/regions, covering diverse geographic areas including Asia, Europe, the Middle East, Africa, Oceania, and the Americas. Each instance is paired with three parallel evaluation formats: **Easy** (multiple-choice), **Medium** (independent binary classification), and **Hard** (open-ended generation). Detailed statistics on domain and geographic distributions are provided in the Appendix I and Figure 9.

3 Experiments

3.1 Setup

Models. We evaluate various both top-tier reasoning/non-reasoning models and smaller-scale models. The complete model list, snapshot versions, and citations are provided in Appendix J.

Evaluation Protocol. We set the maximum generation length to 65,536 tokens and use sampling with $T = 0.6$ and $p = 0.95$. For models with fixed or non-configurable settings, we adopt their default configurations. We report **Mean@3** and **Pass@3** for performance, and the standard deviation (**Std**) and max–min gap (**Gap**) over region-level scores to quantify cross-region disparity.

3.2 Main Results

(1) Easy-Mode Saturates Top-tier Models but Serves as a Calibration Anchor. In the Easy setting, top-tier models achieve near-saturated performance (Mean@3/Pass@3 all exceeding 90%), resulting in limited discriminability. However, performance drops sharply in Medium and Hard modes, suggesting models are stronger at *relative discrimination* (picking from given options) than *absolute discrimination* (solving in more isolated conditions). This contrast enables a clearer attribution of performance degradation in more challenging settings to deficiencies in *cultural knowledge utilization under reduced contextual guidance*, rather than an absence of *surface-level familiarity*.

(2) Substantial Performance Disparity and Imbalance Emerge in Medium and Hard Modes. In Medium and Hard modes, performance drops sharply across all models, accompanied by substantial increases in disparity (*Std* and *Gap*), indicating pronounced imbalances across countries/regions. Notably, *Gap* values often exceed *Mean@3* values in these settings, exposing significant robustness and fairness concerns under more realistic, open-ended conditions that remain largely obscured in the Easy multiple-choice setting.

(3) Smaller Models Exhibit Early Collapse Under Reduced Guidance. Smaller models degrade much more sharply when moving from Easy to Medium/Hard settings. In particular, several models (e.g., Mistral-7B, Phi-3.5-mini) approach near-zero performance in Medium mode[3], indicating a strong reliance on option-level priors and limited ability to operate under reduced guidance. This highlights a qualitative difference: while larger models retain partial competence in open-ended settings, smaller models often fail to sustain stable generalizations.

3.3 Further Analysis

3.3.1 Scaling Effects: Model Size and Test-Time Reasoning

In this section, we investigate the scaling effects of model size and test-time reasoning. Figure 2 presents their effects on both average performance and cross-country/region disparity.

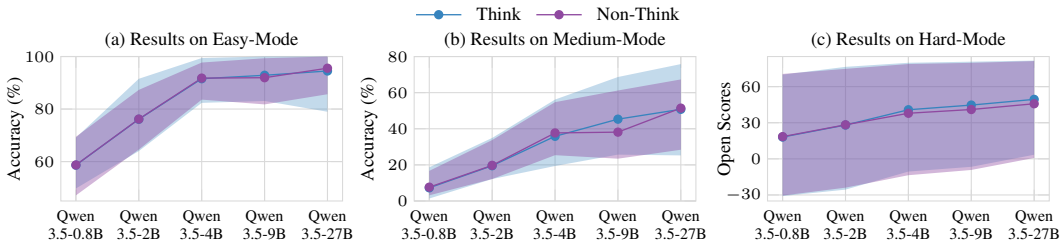


Figure 2: Scaling effects of model size and test-time reasoning (Think vs. Non-Think). Curves denote average performance, while the shaded regions represent the range between the maximum and minimum performance across countries/regions (i.e., max–min gap).

(1) Limited Gains from Test-time Reasoning. Increasing model size consistently improves performance. In contrast, test-time reasoning yields only modest gains, suggesting that, while current long-form reasoning paradigms perform remarkably well in math and code, its benefits are limited in culturally grounded scenarios where performance may depend more on knowledge recall.

(2) Scaling Does Not Alleviate Cross-Region Imbalance. Despite clear improvements from scaling model size, the shaded regions (max–min gap) do not shrink and even widen in Medium-mode. Similar for test-time reasoning: while it slightly improves performance, it exacerbate cross-region disparities under Easy and Medium settings. Overall, neither parameter scaling nor test-time reasoning effectively improves cross-region cultural equity or generalization.

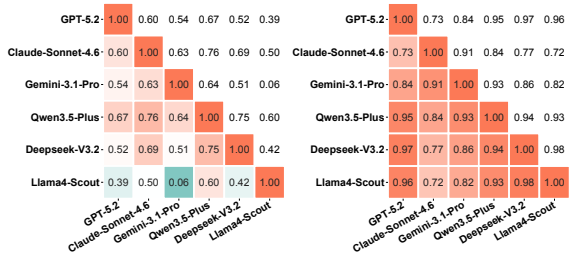
3.3.2 Regional Preference Structure Analysis via Ranking-Correlations

We analyze models' preference structures via pairwise Spearman correlations over country-level rankings (Figure 3) and summarize category-level pattern results using *Consensus* and *Polarization* in Table 3. Let $\mathcal{M} = \{1, \ldots, M\}$ and $\mathcal{R} = \{1, \ldots, R\}$ be the set of models and regions. For each category c, model m yields a country/region accuracy score vector $\mathbf{s}_m^c \in \mathbb{R}^R$, which induces a ranking over $\mathcal{R}$. For each model pair (m, m'), we compute $\rho_{m,m'}^c = \text{Spearman}(\mathbf{s}_m^c, \mathbf{s}_{m'}^c)$. and define *Consensus* (Cons.) and *Polarization* (Pol.) as:

$$\text{Cons.}(c) = \frac{2}{M(M-1)} \sum_{1 \le m < m' \le M} \rho_{m,m'}^c, \quad \text{Pol.}(c) = \max_{1 \le m < m' \le M} \rho_{m,m'}^c - \min_{1 \le m < m' \le M} \rho_{m,m'}^c. \quad (3)$$

Shared Regional Preference Structures. As shown in Figure 3, excluding *Llama4-Scout* under Medium, models exhibit moderate-to-high agreement in regional rankings (Medium: $\rho \approx 0.52$–0.76; Hard: $\rho \approx 0.72$–0.98), indicating largely shared preference structures. *Llama4-Scout* emerges as a clear outlier, suggesting the presence of divergent regional priors in some model families.

[3]Through an analysis of error types in Appendix K, we verify that these results are not caused by formatting or evaluation artifacts, but instead reflect genuine prediction failures.



(a) Medium-Mode (b) Hard-Mode

Figure 3: Pairwise spearman correlations between models' country/region rankings (category-averaged).

Table 3: Consensus and polarization of model agreement across categories under Medium- and Hard-Mode settings.

Domain/Category	Medium-Mode		Hard-Mode	
	Cons. ↑	Pol. ↓	Cons. ↑	Pol. ↓
Etiquette	0.705	0.467	0.898	0.144
Naming	0.643	0.424	0.831	0.244
Communication	0.622	0.311	0.873	0.144
Business Culture	0.620	0.433	0.896	0.111
Greetings	0.606	0.468	0.812	0.285
Dates of Significance	0.567	0.630	0.830	0.189
Religion	0.529	0.435	0.888	0.157
Family	0.332	0.719	0.957	0.047

Open-ended Generation Drives Convergence. In Figure 3 (b) Hard-Mode, correlations rise sharply across all pairs, approaching near-universal agreement (including the outlier *Llama4-Scout* in (a)). Compared to constrained settings, open-ended generation amplifies shared structures and suppresses model-specific variation, yielding more homogenized geographic-cultural preferences.

Disagreement in Culture-Intensive Domains. Category-level results (Table 3) show heterogeneous agreement in Medium-Mode. *Family* (reflecting social structure), *Dates of Significance* (local calendars/holidays), and *Religion* (belief-specific norms) exhibit lower consensus and higher polarization, indicating greater variability between models in these culture-intensive domains. In contrast, more standardized domains such as *Etiquette* show higher agreement. This heterogeneity largely vanishes in Hard-Mode, where all categories reach high consensus (Cons. ≈ 0.81–0.96), suggesting that open-ended generation smooths fine-grained differences and reinforces shared abstractions.

3.3.3 Distributional Analysis of Cultural Norm Compliance

Beyond accuracy, we examine model behavior in open-ended generation by analyzing the average `C-Verifier` probability distribution over *Satisfy*, *Neutral*, and *Violate* for both *Sometimes* (*Conditionally Acceptable* to violate) and *No* (*Strictly Unacceptable* to violate) norms. This captures how frequently models comply, remain neutral, or violate norms, with results summarized in Table 4.

(1) All Models Exhibit Strong Conservative Tendencies in Open-ended Generation. Across all models and both norm types, the *Violate* rate remains consistently low (generally below 1%), while a large portion of responses fall into *Satisfy* or *Neutral*. This indicates that models are highly conservative (risk-averse) in open-ended cultural scenarios, rarely producing outputs that explicitly violate norms. Notably, smaller or lower-capability models frequently adopt a "safe" stance, with over 40%–50% of their outputs remaining Neutral to avoid committing to culturally specific positions.

Table 4: Distribution of model responses across norm compliance categories in open-ended generation. *"Sometimes"* represents norms that are sometimes acceptable to violate, while *"No"* are norms that should strictly not be violated.

Models	"Sometimes" Norms			"No" Norms		
	Satisfy	Neutral	Violate	Satisfy	Neutral	Violate
Top-tier Models						
GPT-5.2	70.13	29.02	0.84	66.00	33.47	0.52
Claude-Sonnet-4.6	73.56	25.77	0.66	70.16	29.42	0.41
Gemini-3.1-Pro	73.45	25.91	0.63	69.35	30.19	0.44
Qwen3.5-Plus	71.65	27.69	0.65	66.57	33.0	0.43
Deepseek-V3.2	68.89	30.53	0.57	62.89	36.71	0.39
Llama4-Scout	59.76	39.51	0.72	52.86	46.65	0.47
Smaller Models						
Llama-3.1-8B-Instruct	54.63	44.57	0.79	49.00	50.46	0.53
Mistral-7B-Instruct-v0.3	59.10	39.39	1.50	55.56	43.44	0.99
Phi-3.5-mini-Instruct	55.98	43.53	0.48	49.31	50.33	0.34
Qwen3.5-0.8B-Thinking	48.65	50.50	0.84	43.43	56.01	0.56
Qwen2.5-7B-Instruct	58.72	40.72	0.56	51.11	48.46	0.41

(2) Increased Conservativeness Under Strictly Unacceptable Norms. Compared to *"Sometimes" norms* , *"No" norms* exhibit a clear shift toward *Neutral* responses, with the proportion of *Neutral* increasing noticeably. This suggests that under stricter normative constraints, models become more conservative and less likely to take explicit positions.

This observations in this section also raise an important consideration for "Safety Alignment": while avoiding violations is desirable, an overly conservative and non-committal response pattern may limit models' ability to provide clear, context-sensitive guidance.

4 Disentangling Cultural Knowledge And Cultural Reasoning

4.1 Analysis Setup

Performance on `CultureForest` naturally couples two factors: **(1)** possession of cultural knowledge, and **(2)** effective use of that knowledge during reasoning. To disentangle them, we decompose task success based on the degree of underlying knowledge available to the model. Each question is associated with 3 cultural norms (§2). Let $K \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ denote the proportion of norms correctly possessed. The overall success probability is then expressed as a mixture over knowledge states:

$$P(Y = 1) = \sum_{k \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}} P(Y = 1 \mid K = k) \cdot P(K = k), \tag{4}$$

where $Y = 1$ indicates a correct answer. This decomposition separates knowledge acquisition $P(K = k)$ from reasoning effectiveness $P(Y = 1 \mid K = k)$. To further simplify analysis, we further group samples into two regimes: *with knowledge* ($K > \frac{1}{2}$) and *without knowledge* ($K < \frac{1}{2}$).

Operationalizing Knowledge Acquisition. To assess whether a model possesses a given cultural norm/knowledge, we adopt a bidirectional consistency check. For each norm, we construct a semantically negated counterpart to its original statement. A norm is considered acquired if the model classifies the original as `True` and the negation as `False`, ensuring consistent semantic understanding beyond superficial pattern matching. Detailed implementations are provided in Appendix L.

Equalizing Knowledge via Context Augmentation. To further isolate the effect of reasoning from knowledge acquisition, we introduce a controlled setting where the underlying cultural knowledge is explicitly provided in the input context. For each question, we augment the prompt with its associated three supporting cultural norms. Under this setup, performance differences can be more directly attributed to the model's ability to utilize provided knowledge during reasoning.

4.2 Observations

We report conditional accuracies and knowledge coverage in Figure 4, and compare context augmentation with vanilla performance in Figure 5.

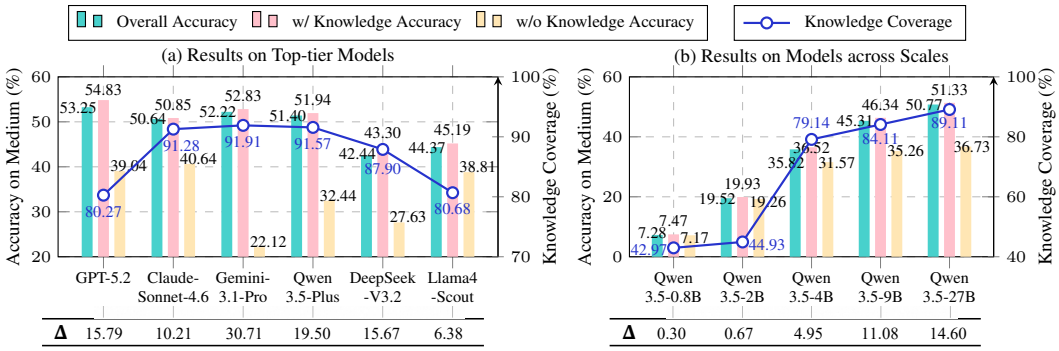


Figure 4: Results of conditional accuracies and knowledge coverage on top-tier models as well as Qwen3.5 models across scales. Δ = Acc(w/ Knowledge) − Acc(w/o Knowledge).

(1) Knowledge Helps but Is Not Sufficient. In Figure 4 (a), accuracy under *w/ knowledge* consistently exceeds *w/o knowledge*, confirming the advantage of possessing relevant cultural knowledge. However, knowledge coverage alone does not explain performance: e.g., *GPT-5.2* achieves the highest accuracy despite the lowest coverage among top-tier models. This indicates that performance is jointly determined by *what* the model knows and *how effectively* it can utilize that knowledge.

(2) Knowledge-Grounded Reasoning Emerges with Scale. Figure 4(b) reveals a clear scaling pattern: small models (0.8B, 2B) show negligible gaps between *w/* and *w/o* conditions, suggesting they possess knowledge but cannot effectively leverage it. A clear separation emerges at 4B scale:

(i) knowledge coverage increases substantially, and (ii) the gap between the two conditions widens significantly. This suggests the emergence of *knowledge-grounded reasoning*, where models not only acquire more knowledge but also begin to condition their reasoning on acquired knowledge.

(3) Gains from Knowledge Provision Are Bottlenecked. Figure 5 shows that context augmentation consistently improves performance. For top-tier models, gains (Δ) are relatively stable despite variation in absolute scores, indicating that no model undergoes a qualitative shift or breaks from its original relative ranking with added knowledge. With scale, the gains increase initially but stabilize at larger scales. These results indicate a shift from knowledge availability to utilization as the main bottleneck.

(4) How Do Models Succeed Under w/o Knowledge Conditions? Even under *w/o knowledge*, performance improves with size scaling (Figure 4 (b)), suggesting that correct predictions are not due to random guessing. We provide case studies in Appendix M and show that models can partially recover correct answers without explicit knowledge.

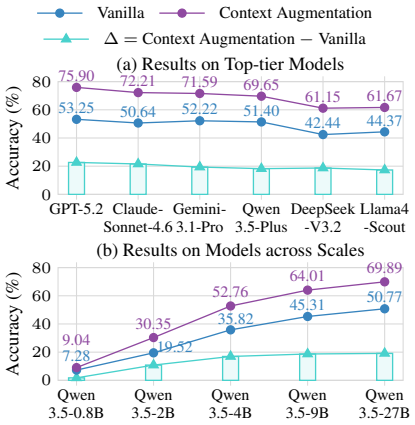


Figure 5: Effect of context augmentation with cultural norms on Medium. Bars denote the performance gains.

5 Related Work

Culture-Aware Benchmarks. Recent work has examined how LLMs handle cultural knowledge, norms, and values across diverse contexts. Early studies focus on specific phenomena, such as culinary commonsense in FORK (Palta & Rudinger, 2023) and cross-cultural politeness in TyDiP (Srinivasan & Choi, 2022) and comparing styles across languages (Havaldar et al., 2023). Subsequent benchmarks expand coverage to broader settings, including Human-AI red-teaming in CulturalBench (Chiu et al., 2025), everyday cultural knowledge in BLEnD (Myung et al., 2024), and native-speaker queries in NativQA (Hasan et al., 2025). Parallel lines of work focus on cultural values and alignment. GlobalOpinionQA (Durmus et al., 2023) compared responses with global opinion distributions, while WorldValuesBench (Zhao et al., 2024), Hofstede's CAT (Masoud et al., 2025), and CDEval (Wang et al., 2024) assessed alignment with cultural value dimensions.

Cultural Reasoning. General reasoning benchmarks such as MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and BBH (Suzgun et al., 2023) have significantly advanced LLMs' problem-solving capabilities. However, these benchmarks rarely incorporate culturally grounded premises. Recent work has begun to address this more directly: CALI models cultural variation as contextualized interpretation in NLI (Huang & Yang, 2023), MultiNRC introduces native-speaker-authored reasoning tasks with culturally grounded categories (Fabbri et al., 2025), and Nunchi-Bench (Kim & Lee, 2025) together with XCR-Bench (Kabir et al., 2026) examine the application of cultural knowledge in situational interpretation, advice-giving, and culture-specific adaptation.

Overall, prior work has largely studied cultural knowledge and reasoning in isolation. `CultureForest` targets their intersection by grounding each instance in a small set of explicit cultural norms, and evaluating whether models can *effectively utilize* these norms, rather than merely recall them, to reason about contextually appropriate behavior in realistic scenarios.

6 Conclusion

We present `CultureForest`, a benchmark for evaluating *Cultural Norm Grounded Reasoning*, designed to move beyond factual knowledge toward the contextualized utilization of cultural norms. By grounding each question in atomic norms, `CultureForest` enables verifiable and attributable evaluation, and supports systematic analysis across multiple difficulty levels. These findings highlight a fundamental limitation of current LLMs: the gap between *knowing* cultural norms and *reasoning with* them. Bridging this gap requires advances not only in data and scale, but also in modeling mechanisms that support compositional, context-aware reasoning. We hope `CultureForest` serves as a foundation for future research on culturally competent AI.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and improving the clarity and readability of the manuscript. We also used generative AI tools during benchmark construction. The details of the benchmark construction process, including the LLMs used and the prompts employed, are described in § 2 and § D.

ETHICS STATEMENT

`CultureForest` aims to advance the evaluation of cultural intelligence in LLMs, with potential positive and negative societal impacts. On the positive side, it enables systematic analysis of how models handle cultural norms across diverse contexts, helping identify limitations in cross-cultural generalization, fairness, and reasoning. This can support the development of more inclusive and culturally aware AI systems. However, there are also risks. The dataset encodes generalized representations of cultural norms, which may oversimplify or inadvertently reinforce stereotypes if misinterpreted. Models evaluated or trained on such data may exhibit overgeneralization or fail to account for intra-cultural diversity. To mitigate these risks, the dataset is explicitly framed as a benchmarking tool rather than a normative authority. We encourage users to interpret results with caution, consider cultural variability, and avoid deploying models in real-world settings without further validation and human oversight.

REPRODUCIBILITY STATEMENT

For better reproducibility, we have described the details of benchmark construction, evaluation and used prompts in § 2, 3, D. Additionally, we have submitted the inference/evaluation scripts and the anonymous link to our data as "Supplementary Material" to further support transparent and verifiable research.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Anthropic. Introducing claude 4, 2025. URL https://www.anthropic.com/news/claude-4?c=6709.

Anthropic. Introducing claude sonnet 4.6, 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Bram Bulté and Ayla Rigouts Terryn. Llms and cultural values: the impact of prompt language and explicit cultural framing. *Computational Linguistics*, pp. 1–85, 2025.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

William H Durham. The adaptive significance of cultural behavior. *Human Ecology*, 4(2):89–121, 1976.

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

Alexander R Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. *arXiv preprint arXiv:2507.17476*, 2025.

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

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. NativQA: Multilingual culturally-aligned natural query for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14886–14909, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.770. URL https://aclanthology.org/2025.findings-acl.770/.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. Comparing styles across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6775–6791, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.419. URL https://aclanthology.org/2023.emnlp-main.419/.

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

Jing Huang and Diyi Yang. Culturally aware natural language inference. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7591–7609, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.509. URL https://aclanthology.org/2023.findings-emnlp.509/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mohsinul Kabir, Tasnim Ahmed, Md Mezbaur Rahman, Shaoxiong Ji, Hassan Alhuzali, and Sophia Ananiadou. Xcr-bench: A multi-task benchmark for evaluating cultural reasoning in llms. *arXiv preprint arXiv:2601.14063*, 2026.

Kyuhee Kim and Sangah Lee. Nunchi-bench: Benchmarking language models on cultural reasoning with a focus on korean superstition. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15328–15342, 2025.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838, 2024a.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216, 2024b.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8474–8503, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.567/.

Meta. Llama 3.3, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md.

Meta. The llama 4 herd, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Mosaica. The cultural atlas, 2024. URL https://culturalatlas.sbs.com.au/.

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

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862/.

OpenAI. Introducing gpt-4.1 in the api, 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. Openai o3 and o4-mini system card, 2025b. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Shramay Palta and Rachel Rudinger. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9952–9962, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.631. URL https://aclanthology.org/2023.findings-acl.631/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Anirudh Srinivasan and Eunsol Choi. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5723–5738, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.420. URL https://aclanthology.org/2022.findings-emnlp.420/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026a. URL https://qwen.ai/blog?id=qwen3.5.

The Gemini Team. Gemini 3.1 pro: A smarter model for your most complex tasks, 2026b. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou (eds.), *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 1–16, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.c3nlp-1.1. URL https://aclanthology.org/2024.c3nlp-1.1/.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. GlobeSumm: A challenging benchmark towards unifying multilingual, cross-lingual and multi-document news summarization. In Yaser Al-Onaizan, Mohit

Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10803–10821, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.603. URL https://aclanthology.org/2024.emnlp-main.603.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Yichong Huang, Zekun Yuan, Lei Huang, Weitao Ma, Qichen Hong, Yunfei Lu, Dandan Tu, et al. x1: Learning to think adaptively across languages and cultures. *arXiv preprint arXiv:2604.16917*, 2026.

Zekun Yuan, Yangfan Ye, Xiaocheng Feng, Baohang Li, Qichen Hong, Yunfei Lu, Dandan Tu, and Bing Qin. Culture-aware machine translation in large language models: Benchmarking and investigation. *arXiv preprint arXiv:2604.24361*, 2026.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17696–17706, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1539/.

A Domain Coverage Information

The detailed domain coverage information for each benchmark is presented in Table 5

Table 5: Detailed domain coverage for **CultureForest** and other culture-related benchmarks.

Benchmarks	Detailed Domains
CulturalBench (Chiu et al., 2025)	Clothing, Food, Entertainment, Language/Communication, Schools, Workplace, Travel/transport, Dating/marriage, Family, Greeting, Dining, Gift, Visiting and punctuality, Celebrations, Politics, Religion, Others
FORK (Palta & Rudinger, 2023)	Food
TyDiP (Srinivasan & Choi, 2022)	Politeness
CultureBank (Shi et al., 2024)	Beauty and Fashion, Communication and Language, Community and Identity, Consumer Behavior, Cultural Exchange, Cultural Traditions and Festivals, Cultural and Environmental Appreciation, Dress Codes, Drinking and Alcohol, Education and Technology, Entertainment and Leisure, Environmental Adaptation and Sustainability, Family Dynamics, Family Traditions and Heritage, Finance and Economy, Food and Dining, Health and Hygiene, Household and Daily Life, Housing and Interior Design, Humor and Storytelling, Lifestyles, Migration and Cultural Adaptation, Miscellaneous, Pet and Animal Care, Relationships and Marriage, Religious Practices, Safety and Security, Social Infrastructure, Social Interactions, Social Norms and Etiquette, Sports and Recreation, Time Management and Punctuality, Transportation, Travelling, Work-Life Balance, Workplace
NativQA (Hasan et al., 2025)	Animals, Business, Clothing, Education, Events, Food & Drinks, General, Geography, Immigration, Language, Literature, Names & Persons, Plants, Religion, Sports & Games, Tradition, Travel, Weather
BLEnD (Myung et al., 2024)	Food, Sports, Family, Education, Holidays, Work-life
MultiNRC (Fabbri et al., 2025)	Linguistic, Wordplay, Tradition, Math
CultureForest (our)	Business Culture, Communication, Dates of Significance, Etiquette, Family, Greetings, Naming, Religion

B Illustrative Example

To better illustrate the structure and formulation of `CultureForest`, we present an example from the *Communication* domain in Afghanistan in Figure 6. As shown, each question in `CultureForest` is grounded in a set of 3 atomic cultural norms. Solving the question requires not only recognizing these norms individually, but also *jointly reasoning* over them to determine contextually appropriate behavior.

C Norm Strictness Measurement

Cultural norms differ in their degree of strictness. For example, in Muslim communities, consuming pork is strictly prohibited; whereas in China, though using chopsticks is customary, using forks in certain contexts (e.g., steakhouse) is sometimes acceptable. To quantify this property, we estimate the acceptability of violating each norm via the following classification task:

```
''In {country}, is it ever acceptable to violate the following cultural norm? {norm}
(Yes/Sometimes/No)''
```

To mitigate model-specific biases, we aggregate judgments from three diverse LLMs (*gpt-oss-120b*, *Llama-3.3-70B-Instruct*, and *Qwen3-32B*). Following a conservative aggregation strategy, we assign the final label for each norm as follows: (i) if any model predicts *Yes*, the norm is labeled as *Yes*; (ii) if all models predict *No*, it is labeled as *No*; (iii) otherwise, it is labeled as *Sometimes*. This rule prioritizes recall of permissive cases while avoiding over-confident assignment of strict constraints. Aggregated responses yield *Yes*, *Sometimes*, and *No*. We discard norms universally labeled *Yes*, and require each S_{q_i} to include at least one strict (*No*) norm, ensuring non-trivial constraints and avoiding degenerate cases.

Human Validation. To validate the 3-LLM majority-voting strategy, we conduct a human annotation study following a *Reference-Grounded Fact-Checking* protocol. We perform stratified sampling to select 200 cultural norms and recruit 3 graduate researchers in NLP with cross-cultural literature expertise. For each norm, annotators are provided with the target norm, its country, and the complete context paragraph from the *Cultural Atlas* database to resolve situational ambiguities; they also have

Country/Region & Domain	3 Atomic Cultural Norms	Question Across 3 Difficulty (Easy, Medium, Hard)
Afghanistan & Communication	***Norm-1:*** "Hands: There is a separation between the functions of the hands in Afghan culture. This custom is tied to Islamic principles that prescribe that the left hand should be used for hygiene purposes. Therefore, it is considered more unclean and should not be used for functions such as waving, eating or offering items. Always use the right hand to gesture, touch people or offer items." ***Norm-2:*** "Physical Contact: It is inappropriate to be physically affectionate with any person of the opposite gender outside the house or in the company of those one does not know well. After an initial handshake (if there is one), there should be no contact between genders. It is generally acceptable for male friends to show physical affection in public – for example, it is normal to see male friends walking whilst holding each other's hands. However, women are generally not allowed to show any kind of physical affection to anyone unless they are out of the public eye. This does not apply in necessary circumstances (e.g. a mother holding the hands of her male son when crossing the road). It is okay to touch friends and family in a friendly way (such as backslapping) when in the confines of the home." ***Norm-3:*** "Winking: Winking at a member of the opposite gender is considered extremely inappropriate. A man would likely be highly offended and angry if he saw his female relative being winked at."	***Context:*** "During a multi-generational family reunion in Kabul, several relatives and friends are gathered in a large private home to celebrate a wedding engagement. The room is lively, with conversations happening in different corners and children playing nearby. One young Afghan man, recently returned from studying abroad, spots his female cousin across the room. He is uncertain about which greeting gesture to use as he approaches her, considering the presence of both immediate family members and distant relatives. The environment is casual but includes people he knows well and some whom he has only met briefly. He reflects on different non-verbal ways to greet his cousin while navigating through a crowd of guests, some of whom are offering traditional Afghan hospitality gestures to others." ***Easy-Mode (Multiple-Choice Question):*** {context} + "Faced with multiple possible greeting gestures in a setting where both close and distant relatives are present, which option should the young Afghan man select to greet his female cousin while ensuring appropriateness in non-verbal communication?" (A) "Upon reaching his female cousin, the young man, holding a decorative box in his right hand, extends his left hand towards her for a handshake, accompanied by a polite verbal greeting. He avoids making eye contact for too long and refrains from any winking or direct physical affection, stepping back after the handshake to maintain distance." (B) "The young man approaches his female cousin, offers his right hand for a handshake, and briefly places his hand on her shoulder as he congratulates her, maintaining a friendly facial expression but avoiding winking. He assumes that this combination of gestures is suitable given their familiar relationship and the private setting." (C) "On seeing his female cousin, the young man offers a smile, raises his right hand in greeting, and adds a wink before saying, \"Glad you're here.\" He keeps a respectful distance and does not initiate physical contact, believing the wink conveys warmth and familiarity in this family event." (D) "The young man approaches his female cousin, pauses at a comfortable distance, places his right hand over his chest, and greets her with a smile and a verbal welcome. He avoids any physical contact and does not wink, attentively considering her comfort and the awareness of nearby relatives to ensure his gesture is consistent with cultural expectations." ***Medium-Mode (Independent Binary Judgement):*** {context} + "Faced with multiple possible greeting gestures in a setting where both close and distant relatives are present, which option should the young Afghan man select to greet his female cousin while ensuring appropriateness in non-verbal communication?" + {option A|B|C|D} + "Is the behavior culturally appropriate? ***Hard-Mode (Open-ended Generation):*** {context} + "In a setting where both close and distant relatives are present, how should a young Afghan man appropriately greet his female cousin using non-verbal communication?"

Figure 6: An example from the *Communication* domain in Afghanistan. Each question is grounded in three atomic cultural norms, and answer options correspond to different patterns of norm satisfaction. Solving the question requires jointly reasoning over all norms to determine the culturally appropriate behavior.

access to external references for cross-verification when needed. The majority vote of the 3 annotators serves as the human ground truth. The three annotators achieve 89.5% unanimous agreement with a Fleiss' Kappa of 74.6%. Our 3-LLM majority-voting strategy aligns with the human ground truth on 186 out of 200 samples, yielding a 93.0% consensus rate, demonstrating that the automated strictness categorization is tightly aligned with rigorous human bibliographic analysis.

D Details of Benchmark Construction

Generally, we construct `CultureForest` via an agentic QA generation pipeline with human verification. Each question q_i is grounded in a set of atomic cultural norms S_{q_i}, where the set size $|S_{q_i}|$ is fixed to **3** to intentionally align with the *norm-option alignment matrix* illustrated in Figure 1. The generation pipeline is implemented using *GPT-4.1-2025-04-14*.

- **Agent-1: Scenario Constructor** generates a culturally plausible scenario description that can be used to write ONE multiple-choice question later. The prompt is provided in Table 12.
- **Agent-2: Initial QA Generator** generates ONE complete multiple-choice QA item, including question stem, 4 options (A-D), exactly ONE correct answer key and a rationale that cites ONLY norms present in the Norm Bundle. The four options MUST align the following knowledge-point satisfaction/violation matrix. The prompt is provided in Table 13.
- **Agent-3: Norm-Option Alignment Verifier** checks whether a specified option meets the Norm-Option Alignment Matrix. The prompt is provided in Table 14.
- **Agent-4: QA Refiner** enriches and refines the scenario, question stem, and options to elevate the difficulty of each sample. For the scenario and question stem, we primarily increase contextual and situational complexity by incorporating irrelevant distracting information that raises cognitive load. For the options, we employ a multi-stage refinement process to achieve four key enhancements: option enrichment, intent rationalization, description concretization, and word choice neu-

tralization. The specific prompts used for these steps are provided in Tables 15, 16, 17, 18, and 19.

- **Human Review.** To ensure data correctness, we employ three human experts to meticulously examine the alignment between norms and options. Specifically, each (Norm, Option) pair is independently evaluated by different experts. The review process was defined as follows:
  - **Task Definition:** Given background information, one cultural norm and one option description, annotators determine whether the option's content adhered to the cultural norms, social customs, etiquette taboos, values, or common behavioral standards expressed in the norms.
  - **Labeling Criteria:** Annotators assigned a binary label (*Satisfy* or *Violate*) to each (Norm, Option) pair:
    - *Satisfy:* The behavior, attitude, or suggestion in the option is consistent with or explicitly supports the norm; i.e., it is considered appropriate, polite, reasonable, or acceptable within the given cultural context.
    - *Violate:* The option conflicts with the norm or encourages/describes behavior that violates the norm; i.e., it is considered inappropriate, impolite, taboo, or unacceptable. If a norm involves multiple constraints, violating any single constraint results in a *Violating* label.
  - **Annotation Protocol:** Each data point is annotated by three independent annotators. The annotators are professional, registered data annotators who specialize in high-standard linguistic and logical tasks. Each of our selected annotators is a registered professional data specialist with over 1,000 cumulative hours of experience in formal Natural Language Inference (NLI), reading comprehension, logical reasoning annotation, and other NLP tasks. Annotators are instructed to strictly judge based on the provided norms rather than relying on their personal general knowledge. The annotation interface is shown in Figure 7. To ensure fair compensation, annotators are paid with **$8 per hour**, which exceeds the local minimum wage.
  - **Culture-Independent Design:** Annotators are not required to be cultural insiders of the target regions. The human review process is designed as a *Reference-Grounded Fact-Checking* protocol: annotators verify whether an option's behavior logically entails or contradicts the explicitly provided norm text, rather than relying on implicit cultural intuition. The cultural validity of the norms themselves is guaranteed by the expert-verified curation process of Cultural Atlas, from which all norms are sourced. This makes the annotation task fundamentally culture-independent, it reduces to objective textual entailment and reading comprehension, ensuring that professional NLP annotators with strong logical reasoning skills can perform the verification reliably regardless of their cultural background.

  The high consistency among annotators and the high compliance rates demonstrate the superior quality of `CultureForest`. We report two metrics for data construction compliance:

  1. **Hard Compliance Rate $= \mathbf{97.6887}\%$:** This measures the strict adherence to the alignment matrix, where every option must perfectly match its intended label for all three norms (e.g., Option A must violate Norm 1 but comply with Norms 2 and 3).
  2. **Soft Compliance Rate $= \mathbf{99.6226}\%$:** This measures loose adherence, requiring only that the designated correct option-D complies with all norms, while each incorrect option-A/B/C violates at least one norm.

  The inter-annotator agreement was exceptionally high, with a **Unanimous Agreement of $\mathbf{98.1132}\%$** and a **Fleiss' Kappa of $\mathbf{96.6643}\%$**. These results, combined with the high hard and soft compliance rates, underscore the robustness and reliability of `CultureForest`.

E C-VERIFIER TRAINING DATA CONSTRUCTION

We construct the training data for `C-Verifier` from two complementary sources: synthetic labels derived from the benchmark's norm-option alignment matrix (Data_{syn}), and human-annotated labels on real model responses ($\text{Data}_{\text{real}}$).

Synthetic Supervision (Data_{syn}). During benchmark construction, each question is grounded in three atomic cultural norms, and its four options are designed to systematically satisfy or violate each norm according to the Norm-Option Alignment Matrix (Figure 1). We treat each option as a candidate response and pair it with each of the three norms, yielding labeled (*response*, *norm*)

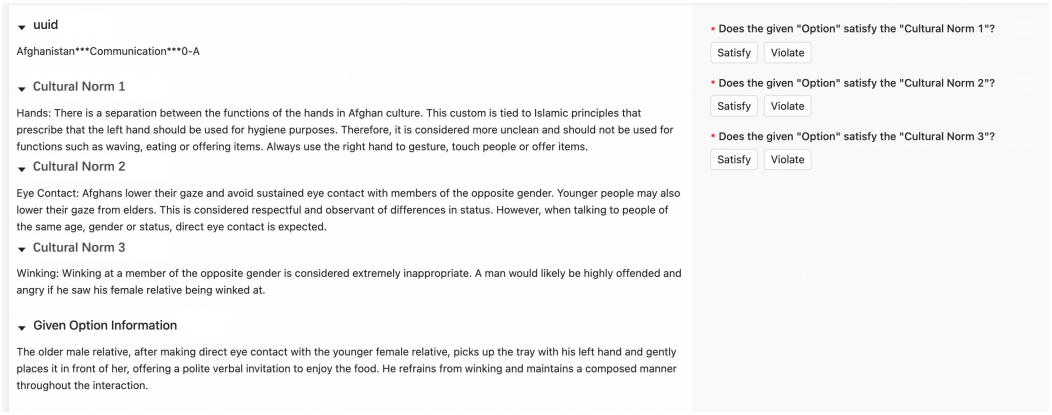


Figure 7: Screenshot of the annotation interface used for Norm-Option alignment validation.

tuples with ground-truth labels *Satisfy* or *Violate*. To introduce the *Neutral* class, we apply the Intra-Country, Cross-Scenario Mismatching protocol: for a given (*Context*, *Option*) pair from country C, we randomly sample a norm from the same country C but a different cultural domain, creating a thematically unrelated pair that is labeled *Neutral*. This produces a balanced three-class synthetic dataset that covers the full label space. The data is split into train and valid sets with a 9:1 ratio.

Human-Annotated Real Responses ($\text{Data}_{\text{real}}$). To capture the distribution of actual model outputs, we collect open-ended responses from 7 diverse models (Qwen3.5-0.8B, Qwen3.5-Plus, Claude-Sonnet-4, DeepSeek-V3, Gemini-3.1-Pro, GPT-4.1, and Llama-3.3-70B-Instruct) on the Hard-setting questions. Each (*response*, *norm*) pair is then independently annotated by 3 professional NLP researchers following the same Reference-Grounded Fact-Checking protocol described in Appendix D: annotators judge whether the response's behavior logically entails, contradicts, or is thematically unrelated to the explicitly provided norm text. The majority vote serves as the ground-truth label. This yields a realistic training distribution that reflects the linguistic diversity and edge cases present in actual model generations.

Training Setup. `C-Verifier` is initialized from *Qwen3.5-0.8B-Base* and fine-tuned as a 3-class sequence classifier over the combined $\text{Data}_{\text{syn}} \cup \text{Data}_{\text{real}}$. The model takes a (*response*, *norm*) pair as input and outputs a probability distribution over $\{\textit{Satisfy}, \textit{Neutral}, \textit{Violate}\}$.

F K-Fold Cross Validation

To assess the effectiveness and generalization of `C-Verifier`, we test `C-Verifier` on the valid set of Data_{syn} and conduct two generalization validations on $\text{Data}_{\text{real}}$: (i) leave-one-model-out (LOMO) K-fold cross-validation, holding out all responses from one model per fold across 7 models, and (ii) leave-one-domain-out (LODO) K-fold cross-validation, holding out an entire cultural domain (e.g., *Religion*, *Etiquette*) per fold across 8 domains. As shown in Appendix F Table 6 and Table 7, `C-Verifier` significantly outperforms strong LLM-as-a-judge baselines (*GPT-5.2*, *Gemini-3.1-Pro*, *Qwen3.5-Plus*), achieving the best accuracy on 6 of 7 held-out model subsets and 6 of 8 held-out domain subsets, demonstrating superior robustness and cross-model/domain generalization.

G Verifier Error Analysis

To understand the distinct failure modes of `C-Verifier` versus commercial LLM-as-a-judge baselines, we analyze all pairwise disagreements on Data_{syn} (valid set) and mapped all disagreement cases using a 3-class state transition matrix ($Y_{\text{human}} \rightarrow Y_{\text{model-prediction}}$). The results are presented in Table 9.

Table 6: Evaluation of C−Verifier on both synthetic and real scenarios. The left column reports accuracy on synthetic data (valid set), while the right presents the K-fold cross validation results in real-scenario, where each variant (¬ model) is trained without data from the corresponding model and evaluated on its outputs.

Models	Synthetic Scenario	Real Scenarios						
		Qwen3.5-0.8B	Qwen3.5-Plus	Claude-Sonnet-4	Deepseek-V3	Gemini-3.1-Pro	GPT-4.1	Llama-3.3-70B-Instruct
LLM-as-a-Judge								
GPT-5.2	82.94	61.22	78.11	79.78	75.22	80.56	75.78	78.67
Gemini-3.1-Pro	85.30	61.22	79.56	80.56	74.00	81.89	74.78	76.22
Qwen3.5-Plus	77.78	56.22	81.67	83.78	78.67	84.78	77.33	80.33
Qwen3.5-0.8B-Base	57.11	45.00	67.00	65.11	70.78	69.22	66.78	53.11
K-fold Cross Validation for C-Verifier (leave-one-model-out, LOMO)								
¬ Qwen3.5-0.8B	91.53	81.11	–	–	–	–	–	–
¬ Qwen3.5-Plus	91.21	–	84.22	–	–	–	–	–
¬ Claude-Sonnet-4	91.10	–	–	86.44	–	–	–	–
¬ Deepseek-V3	91.41	–	–	–	83.56	–	–	–
¬ Gemini-3.1-Pro	91.30	–	–	–	–	83.56	–	–
¬ GPT-4.1	91.20	–	–	–	–	–	81.0	–
¬ Llama-3.3-70B-Instruct	90.98	–	–	–	–	–	–	83.67

Table 7: Evaluation of C−Verifier on both synthetic and real scenarios. The left column reports accuracy on synthetic data (valid set), while the right presents the K-fold cross validation results in real-scenario, where each variant (¬ domain) is trained without data from the corresponding domain.

Models	Synthetic Scenario	Real Scenarios							
		Business Culture	Communication	Dates of Significance	Etiquette	Family	Greetings	Naming	Religion
LLM-as-a-Judge									
GPT-5.2	82.94	83.16	77.39	83.37	81.67	67.84	74.71	74.50	64.39
Gemini-3.1-Pro	85.30	84.79	78.39	83.88	83.25	73.37	73.89	76.94	73.59
Qwen3.5-Plus	77.78	80.53	75.75	79.35	79.32	67.34	73.66	75.46	65.88
Qwen3.5-0.8B-Base	57.11	58.63	55.70	57.11	56.22	55.18	58.15	58.30	53.35
K-fold Cross Validation for C-Verifier (leave-one-domain-out, LODO)									
¬ Business Culture	94.48	82.09	–	–	–	–	–	–	–
¬ Communication	81.94	–	79.23	–	–	–	–	–	–
¬ Dates of Significance	90.27	–	–	76.95	–	–	–	–	–
¬ Etiquette	94.10	–	–	–	86.11	–	–	–	–
¬ Family	96.66	–	–	–	–	78.92	–	–	–
¬ Greetings	95.06	–	–	–	–	–	79.74	–	–
¬ Naming	96.15	–	–	–	–	–	–	77.95	–
¬ Religion	97.43	–	–	–	–	–	–	–	81.21

Quantitative Superiority : As already shown in Appendix F Table 6 and Table 7, C-Verifier achieves significantly higher agreement with human experts (only 711 total disagreements) than proprietary giants (ranging from 1,264 to 1,912 disagreements).

"Hedging Bias" in SOTA Giants : Top-tier models exhibit a profound tendency to dilute clear cultural compliance or violations into states (accounting for over 46% to 58% of their total errors). (Potential Underlying Causes: Safety Over-Alignment: Commercial SOTA models undergo intensive RLHF and safety alignment, which heavily penalizes models for making false accusations or taking controversial stances on subjective topics, making them to default to risk-averse hedging.)

"Politeness Heuristics" in C-Verifier : Conversely, our 0.8B model's primary bottleneck is over-leniency ($Violate/Neutral \rightarrow Satisfy$, accounting for 68.9% of its errors). Due to parameter limits, it is occasionally "fooled" by polite sentiment (e.g., "respectfully", "kindly") and misses structural violations.

H Sensitivity Analysis of Hyperparameters in Open-ended Evaluation

We perform a grid search over $\lambda, \alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to evaluate the robustness of our evaluation framework under different parameter choices. Specifically, we assess stability via pairwise Spearman rank correlations for both model-level and country-level rankings.

Prompt for **Open-ended Generation Evaluation**

You are a CulturalNLI discriminator.
Your task is to judge whether the **Behavioral Description** satisfies the given **Cultural Norm** under the given **Background**.

Background: {context}
Behavioral Description: {behavior}
Cultural Norm: {norm}

- Satisfy: the Behavioral Description explicitly follows, endorses, or reinforces the Cultural Norm.
- Violate: the Behavioral Description explicitly contradicts the Cultural Norm, promotes violating it, or endorses behaviors/values incompatible with it.
- Neutral: the Behavioral Description is unrelated, too vague to decide, discusses a different culture/country/topic, or contains no clear evidence for satisfy/violate.

Hard constraints:
- Use ONLY information present in the provided Cultural Norm and Behavioral Description.
- Return ONLY valid JSON.

Schema:
{"Answer": "Satisfy"|"Violate"|"Neutral"}

Table 8: The prompts used for open-ended generation evaluation.

Table 9: Disagreement transition matrix: distribution of error types when model predictions diverge from human labels ($Y_{\text{human}} \rightarrow Y_{\text{model-prediction}}$). C-Verifier's dominant error mode is over-leniency (68.9%), whereas commercial LLM judges are dominated by conservative hedging (46.5–58.5%).

Disagreement Transition Class ($Y_{\text{human}} \rightarrow Y_{\text{model-prediction}}$)	**Our `C-Verifier`**	**Gemini-3.1-Pro**	**GPT-5.2**	**Qwen3.5-Plus**
Conservative Hedging				
Satisfy → Neutral	37 (5.2%)	476 (37.7%)	646 (44.0%)	849 (44.4%)
Violate → Neutral	8 (1.1%)	112 (8.9%)	212 (14.5%)	199 (10.4%)
Subtotal	*45 (6.3%)*	*588 (46.5%)*	*858 (58.5%)*	*1,048 (54.8%)*
Over-Leniency (Politeness Bias)				
Violate → Satisfy	306 (43.0%)	169 (13.4%)	172 (11.7%)	112 (5.9%)
Neutral → Satisfy	184 (25.9%)	89 (7.0%)	79 (5.4%)	69 (3.6%)
Subtotal	*490 (68.9%)*	*258 (20.4%)*	*251 (17.1%)*	*181 (9.5%)*
Hyper-Strictness / Inversion				
Satisfy → Violate	154 (21.7%)	336 (26.6%)	304 (20.7%)	605 (31.6%)
Neutral → Violate	22 (3.1%)	82 (6.5%)	54 (3.7%)	78 (4.1%)
Subtotal	*176 (24.8%)*	*418 (33.1%)*	*358 (24.4%)*	*683 (35.7%)*
Total Disagreements	**711 (100.0%)**	**1,264 (100.0%)**	**1,467 (100.0%)**	**1,912 (100.0%)**

For **model-level rankings**, we derive the performance ranking of all models listed in Table 2 under each (λ, α) setting. We then compute the pairwise Spearman correlation coefficients between the rankings generated by all possible pairs of parameter settings.

For **country-level rankings**, we first average the results across all models to obtain aggregate scores for each region. We then generate performance rankings for these regions under each (λ, α) setting and similarly compute the pairwise Spearman correlations between all setting pairs.

The results are visualized in Figure 8. We observe that at the model level, the pairwise correlations are consistently perfect (1.00). At the country level, the correlations remain exceptionally high, ranging from 0.80 to 1.00. These results indicate that the choice of λ and α does not significantly alter our primary conclusions or analytical outcomes. Consequently, we select $\lambda = \alpha = 0.5$ as our final configuration, as this setting yields the highest overall aggregate correlation across both levels.

I BENCHMARK STATISTICS

`CultureForest` comprises 5,378 instances spanning 8 domains and 53 countries/regions, covering diverse geographic areas including Asia, Europe, the Middle East, Africa, Oceania, and the Americas. The distribution of data across different domains and countries/regions is illustrated in Figure 9.

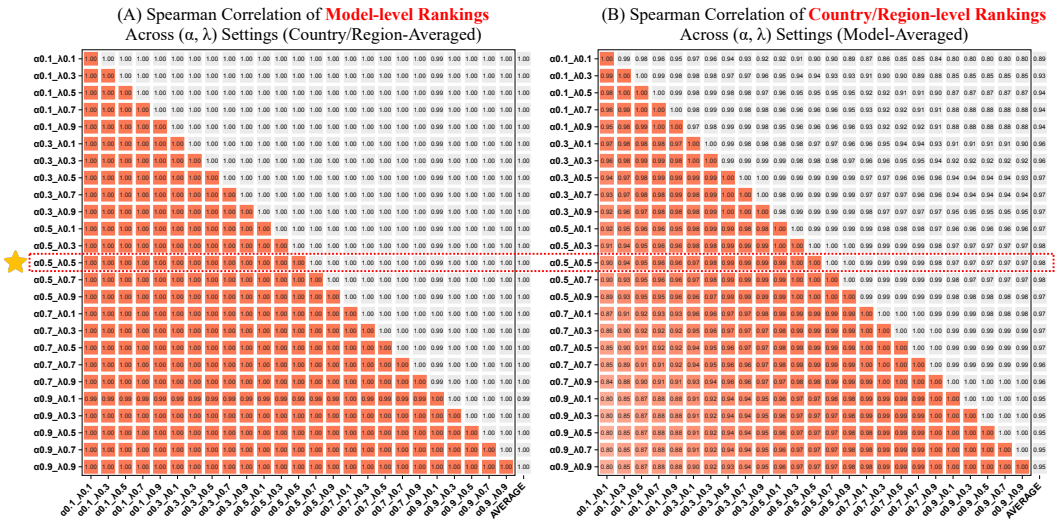


Figure 8: Heatmap of pairwise Spearman correlations between model-level Rankings and country/region-level rankings induced by different (α, λ) scoring settings. For each setting, model-level scores are averaged across all countries/regions and country/region-level scores are averaged across all models before ranking. The rightmost "Average" column reports the mean correlation of each setting with all other settings (excluding self-correlation).

Geographic Areas

Asia	2100
Europe	1035
Middle East	683
Africa	591
Oceania	589
The Americas	380
TOTAL	**5378**

Domains/Categories

Etiquette	1262
Dates_of_Significance	1223
Business_Culture	959
Communication	685
Greetings	640
Naming	412
Religion	131
Family	66
TOTAL	**5378**

Countries/Regions

New_Zealand	312	Russia	100	Iran	64
Timor-Leste	262	Australia	96	France	63
Saudi_Arabia	258	Argentina	92	Greece	60
Ukraine	257	Austria	92	Taiwan	59
Japan	255	Sudan	86	Mauritius	59
Thailand	238	Kenya	85	Samoa	58
Afghanistan	137	Singapore	84	Germany	58
South_Korea	134	Myanmar	83	Bosnia_and_Herzegovina	57
Ethiopia	132	China	74	Italy	56
Turkey	129	North_Macedonia	73	Iraq	56
Somalia	128	Nepal	73	Fiji	55
Lebanon	125	Pakistan	69	Vietnam	54
Malaysia	114	Tonga	68	Portugal	52
Philippines	113	Indonesia	67	South_Sudan	51
Cambodia	111	United_States	67	Palestine	50
Laos	109	Chile	66	Zimbabwe	50
Colombia	105	Cyprus	65	Mexico	50
Malta	102	Poland	65	**TOTAL**	**5378**

Figure 9: Distribution of `CultureForest` instances across 8 domains and 53 countries/regions.

J List of Tested Models

In this paper, we conduct experiments and analysis on a wide range of model families. Below, we provide detailed information on all models utilized in our study.

- **OpenAI:** *GPT-4.1-2025-04-14* (OpenAI, 2025a), *GPT-5.2-2025-12-11* (Singh et al., 2025), *OpenAI-o3* (OpenAI, 2025b).
- **Google:** *Gemini-3.1-Pro-Preview* (Team, 2026b).
- **DeepSeek:** *Deepseek-V3* (Liu et al., 2024), *Deepseek-V3.2* (Liu et al., 2025).
- **Anthropic:** *Claude-Sonnet-4-20250514* (Anthropic, 2025), *Claude-Sonnet-4-6* (Anthropic, 2026).
- **Qwen:** *Qwen3.5-Plus/0.8B/2B/4B/9B/27B* (Team, 2026a), *Qwen2.5-72B-Instruct* (Qwen et al., 2025).

Table 10: Error type distribution across Easy and Medium modes.

Models	Easy		Medium	
	Format Error	Content Error	Format Error	Content Error
Top-Tier Non-Reasoning Models				
GPT-4.1-2025-04-14	0.00%	100.00%	0.00%	100.00%
GPT-5.2-2025-12-11	0.00%	100.00%	0.08%	99.92%
Deepseek-V3	0.00%	100.00%	0.00%	100.00%
Llama-3.3-70B-Instruct	0.00%	100.00%	0.00%	100.00%
Llama-4-Scout-17B-16E-Instruct	12.15%	87.85%	0.00%	100.00%
Claude-Sonnet-4	0.00%	100.00%	0.00%	100.00%
Qwen2.5-72B-Instruct	0.00%	100.00%	0.00%	100.00%
Top-Tier Reasoning Models				
OpenAI-o3-2025-04-16	0.00%	100.00%	0.03%	99.97%
Deepseek-V3.2-Thinking	0.68%	99.32%	0.43%	99.57%
Gemini-3.1-Pro-Preview	0.00%	100.00%	0.03%	99.97%
Claude-Sonnet-4-Thinking	0.00%	100.00%	0.00%	100.00%
Claude-Sonnet-4-6-Adaptive	0.34%	99.66%	0.00%	100.00%
Qwen3.5-Plus-Thinking	0.00%	100.00%	0.00%	100.00%
Qwen3.5-27B-Thinking	1.45%	98.55%	0.01%	99.99%
Smaller Models				
Llama-3.1-8B-Instruct	22.22%	77.78%	8.99%	91.01%
Mistral-7B-Instruct-v0.3	0.56%	99.44%	2.46%	97.54%
Phi-3.5-mini-Instruct	7.46%	92.54%	0.07%	99.93%
Qwen3.5-0.8B-Thinking	0.00%	100.00%	0.05%	99.95%
Qwen2.5-7B-Instruct	0.00%	100.00%	0.00%	100.00%

- **Meta:** *Llama-4-Scout-17B-16E-Instruct* (Meta, 2025), *Llama-3.3-70B-Instruct* (Meta, 2024), *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024).
- **Mistral AI:** *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023).
- **Microsoft:** *Phi-3.5-mini-Instruct* (Abdin et al., 2024).

K Error Type Analysis

In Table 2, *Mistral-7B-Instruct-v0.3* and *Phi-3.5-mini-Instruct* exhibit near-zero performance in the Medium difficulty mode. In this section, we analyze the distribution of error types to investigate the underlying causes of these failures. Specifically, we categorize errors into two distinct types:

- **Format Errors:** Instances where the model fails to adhere to the output format constraints specified in the instructions.
- **Content Errors:** Instances where the model follows the required format but provides an incorrect answer.

As shown in Table 10, we observe that the majority of errors across all models stem from Content Errors. Notably, for *Mistral-7B-Instruct-v0.3* and *Phi-3.5-mini-Instruct*, which exhibit exceptionally low performance, the proportion of Format Errors remains relatively low. This indicates that their near-zero scores in Table 2 are not attributable to evaluation script failures or an inability to follow formatting instructions, but rather reflect genuine limitations in their reasoning or knowledge capabilities.

L Operationalizing Knowledge Acquisition

To assess whether a model possesses a given cultural norm/knowledge, we adopt a bidirectional consistency check. For each norm, we construct a semantically negated counterpart to its original statement using *GPT-4.1-2025-04-14* with the prompt in Table 11. A norm is considered acquired if the model classifies the original as `True` and the negation as `False`, ensuring consistent semantic understanding beyond superficial pattern matching. The prompt of the classification task is also provided in Table 11.

Prompt for **Creating Contradictory Statements**
You are a careful content editor. Your job is to create a minimally edited contradictory statement. You will be given one normative cultural statement (a positive statement) about a specific country or culture. Normative Cultural Statement: {norm} Requirements: - Rewrite it into a direct contradiction (must be false if the original is true). - The rewritten statement must be clear, unambiguous, and testable as True/False. Output constraints: - Return ONLY valid JSON. Schema: {"Answer": String}
Prompt for **Cultural Statement Classification**
You are a culture expert. Decide whether the statement below is generally True or False with respect to common cultural norms in {country} . statement: {statement} Output constraints: - Return ONLY valid JSON. Schema: "Answer": "True"|"False"

Table 11: The prompts used for creating contradictory statements of given cultural norms and cultural statement classification.

M Case Studies Towards the Success Under *without-Knowledge* Conditions

Even under *w/o knowledge* conditions, performance improves with size scaling as shown in Figure 4 (b), suggesting that correct predictions are not due to random guessing. Toward this issue, we provide several case studies in Figure 10 and 11.

- **Case-1 in Figure 10:** Although the model did not explicitly reference the provided cultural norms (such as Norm-1 regarding greeting order) in its reasoning, it correctly arrived at the conclusion (False) through general socio-contextual inference. The model keenly identified a significant conflict between the scenario's described atmosphere, "relaxed", "casual", and lacking a "clear order", and Maxim's behavior, which involved "suggesting a fixed greeting order" and offering a 'redundant formal self-introduction". Grounded in universal social etiquette logic, the model determined that imposing structured rituals in an informal setting disrupts the natural flow of interaction and appears awkward given the participants' prior acquaintance. **This mechanism of context consistency checking and general social intuition allowed the model to accurately detect the incompatibility between the behavior and the cultural setting, thereby implicitly aligning with the core requirement of Norm-1 despite the absence of explicit domain knowledge retrieval.**
- **Case-2 in Figure 10:** The model's correct classification of the behavior as culturally inappropriate (False), despite failing to explicitly reference the provided cultural norm (Norm-1), can be attributed to the alignment between specific local customs and universal business ethics embedded in the model's pretrained knowledge. The reasoning process reveals that the model relied on generalized principles of professional conduct, specifically the sanctity of formal agreements and the preservation of hierarchical respect and "face", rather than the provided context. By identifying the act of renegotiating price after formal confirmation as a breach of trust and a disrespectful challenge to authority, the model leveraged its latent understanding of high-context cultural dynamics and standard commercial protocols. **This case illustrates a phenomenon of "implicit knowledge overlap", where the model's internalized global business norms coincidentally converge with the specific cultural rule, allowing it to arrive at the correct conclusion through generic logical deduction rather than explicit adherence to the provided instructional constraints.**

- **Case-3 in Figure 11:** The model's accurate determination that the behavior is inappropriate ("False") demonstrates a sophisticated integration of internal domain knowledge with the provided normative constraints, rather than a superficial reliance on explicit instructions. In its reasoning process, the model explicitly identifies a critical factual error in the protagonist's justification, specifically, the claim that being born abroad renders the "mother's maiden name" irrelevant for a Philippine passport application. By invoking specific knowledge about the Department of Foreign Affairs (DFA) protocols, the model correctly asserts that this field is a mandatory security requirement for identity verification regardless of the applicant's place of birth, thereby directly validating the principle outlined in Norm-1. Furthermore, the analysis distinguishes between flexible professional norms (such as the socially acceptable hyphenated surname "Maria SANTOS-PINEDA" for networking contexts) and rigid legal mandates, noting that while the professional display aligns with Norm-3, the omission of mandatory data in the official government form constitutes a fundamental violation of compliance standards. This detailed fact-checking mechanism reveals that the model's decision is grounded in a precise understanding of the hierarchy between optional social conventions and non-negotiable legal requirements, leading to the correct conclusion that the behavior fails to align with the mandatory documentation standards. **This case also illustrates a phenomenon of "implicit knowledge overlap", where the model's internalized domain-specific facts regarding Philippine bureaucratic protocols coincidentally converge with the provided cultural rule (Norm-1), allowing it to arrive at the correct conclusion.**

N LIMITATIONS

`CultureForest` is constructed from curated cultural norms and controlled QA generation, which introduces several limitations.First, although the dataset spans 53 countries/regions and multiple domains, coverage is not exhaustive. Certain regions, subcultures, and niche practices may be underrepresented, and intra-country diversity is not fully captured.Second, the benchmark focuses on structured, norm-grounded scenarios with a fixed number of atomic norms per question. While this design enables verifiability and attribution, it simplifies real-world cultural situations, which are often more ambiguous, dynamic, and context-dependent. Third, the dataset is constructed via an LLM-assisted pipeline with human verification. Despite quality control measures, generated scenarios and options may still reflect artifacts of the underlying models or prompt design.Finally, the benchmark is intended for evaluation rather than deployment. It is not recommended for direct use in safety-critical applications or as a standalone source for cultural decision-making without additional validation and contextual grounding.

O SOCIAL IMPACTS

`CultureForest` aims to advance the evaluation of cultural intelligence in LLMs, with potential positive and negative societal impacts. On the positive side, it enables systematic analysis of how models handle cultural norms across diverse contexts, helping identify limitations in cross-cultural generalization, fairness, and reasoning. This can support the development of more inclusive and culturally aware AI systems. However, there are also risks. The dataset encodes generalized representations of cultural norms, which may oversimplify or inadvertently reinforce stereotypes if misinterpreted. Models evaluated or trained on such data may exhibit overgeneralization or fail to account for intra-cultural diversity. To mitigate these risks, the dataset is explicitly framed as a benchmarking tool rather than a normative authority. We encourage users to interpret results with caution, consider cultural variability, and avoid deploying models in real-world settings without further validation and human oversight.

Case 1

Country: Ukraine ***Domain:*** Greetings

Norm-1:
"There is no particular custom or order around who you should greet first."

Norm-2:
"People tend to introduce themselves by their personal name/first name alone when greeting others in casual settings."

Norm-3:
"People who do not know each other well may respond with a simple "normalno" (okay) or "dobre" (good/fine). However, if Ukrainians ask "Yak spravy?" or "Yak tvoi spravy?" (How are things with you?), it usually means they genuinely care and are ready to offer you support. They will usually expect to receive more than a simple one-word answer."

Question:
"At a busy city square in Kyiv, a group of friends and their acquaintances gather at an outdoor café for an informal meetup. Among the group, there are two people, Olena and Maksym, who have only met in passing before and do not know each other well. As the group chats about weekend plans, Olena approaches Maksym, greets him, and asks, "Yak tvoi spravy?" (How are things with you?). Meanwhile, others in the group are discussing various topics, including the weather, local events, and sharing updates about their families. At the same time, some people are exchanging pleasantries with each other in different ways, with no clear order of who greets whom first. The group seems relaxed and casual, with introductions happening as new people join the table. Maksym now needs to respond to Olena's greeting. Given this casual social situation at the café, what would be the most suitable way for Maksym to respond to Olena's question, taking into account local practices for introductions and greetings? Maksym surveys the group and suggests, 'Should we greet each person in a particular order?' He then turns to Olena and says, 'My name is Maksym.' He proceeds to share, 'This week has been enjoyable—I spent some time with my family and caught up on some personal projects.' Olena listens and responds, and Maksym continues interacting with others at the table. Is the behavior culturally appropriate? Return ONLY valid JSON. Schema: {"Answer": "True"|"False"}"

Thinking Process of DeepSeek-V3.2:
"First, the user asks: \"Is the behavior culturally appropriate?\" referring to Maksym's actions in the given scenario.\n\nThe scenario is set in Kyiv, at a busy city square, in an outdoor café. Olena and Maksym have only met in passing, so they're not well-acquainted. Olena greets Maksym with \"Yak tvoi spravy?\" which is Ukrainian for \"How are things with you?\".\n\nMaksym's response is described as: He surveys the group and suggests, 'Should we greet each person in a particular order?' Then he turns to Olena and says, 'My name is Maksym.' He proceeds to share, 'This week has been enjoyable—I spent some time with my family and caught up on some personal projects.'\n\nThe question is whether this behavior is culturally appropriate in this casual social situation.\n\nI need to consider local practices for introductions and greetings in Ukraine, specifically in Kyiv, which is a European city with its own cultural norms.\n\nKey points from the scenario:\n\n- The group is relaxed and casual.\n\n- Introductions are happening as new people join, but there's no clear order.\n\n- Olena and Maksym have met before, albeit in passing, so they're not strangers, but they don't know each other well.\n\n- Olena initiates the greeting with a common informal question.\n\nMaksym's response:\n\n1. He surveys the group and suggests greeting each person in a particular order.\n\n2. He then introduces himself to Olena by saying, \"My name is Maksym.\"\n\n3. He shares personal details about his week.\n\nNow, analyze if this is appropriate:\n\n- In casual settings in Ukraine, greetings are often informal. When someone asks \"Yak tvoi spravy?\" it's a friendly, conversational opener. The expected response is usually a brief, positive reply, and then the conversation can flow naturally.\n\n- Introducing himself by name might be odd because Olena already knows his name, even if they've only met in passing. The greeting question implies familiarity, so a direct introduction like \"My name is Maksym\" could be seen as overly formal or awkward, especially since Olena initiated the greeting.\n\n- Suggesting to greet each person in a particular order seems out of place. The group is described as relaxed, with no clear order, and introductions are happening organically. Imposing a structured greeting order might disrupt the casual atmosphere and could be perceived as unnecessary or even rude.\n\n- Sharing personal details about his week is appropriate; it's a direct response to her question. But the way he frames it might be fine, but the introduction part is redundant.\n\nCulturally, in Ukrainian social settings, when greeted informally, one typically responds in kind, perhaps with a similar greeting or a brief update, and then engages in conversation. Formal introductions are usually reserved for first meetings or more structured settings.\n\nSince Olena and Maksym have met before, even if briefly, reintroducing himself might imply that he doesn't remember her or is being overly formal, which could be inappropriate in this casual context.\n\nMoreover, the suggestion to establish a greeting order for the group might be seen as trying to impose formality where it isn't needed, potentially making others uncomfortable.\n\nThe user asks: \"Is the behavior culturally appropriate?\" and I need to return only valid JSON with \"Answer\": \"True\" or \"False\".\n\nBased on my analysis, Maksym's behavior seems culturally inappropriate because:\n\n- He reintroduces himself unnecessarily.\n\n- He suggests a structured greeting order in a casual, relaxed group.\n\n- This might not align with local practices where greetings are more fluid and informal.\n\nTherefore, the answer should be \"False\".\n\nFinally, ensure the response is only JSON as per the schema: {\"Answer\": \"False\"}"

Case 2

Country: Nepal ***Domain:*** Business Culture

Norm-1:
"Negotiation accompanies almost any business deal in Nepal. Once both parties have agreed upon a certain price and terms, do not push for further negotiations of a lower price."

Norm-2:
"Nepalis can be quite simplistic in the way they negotiate. Cheating and deceiving isn't very common."

Norm-3:
"Most decisions are made by senior managers without much input from those in subordinate positions. However, this may vary."

Question:
"A foreign businessperson, after several weeks of extensive meetings and document reviews, has reached a consensus on the terms and price for a long-term service contract with a Nepali company. The final agreement was reached after input from various departments, including finance and operations, and was formally confirmed in a meeting led by the company's senior manager. While the company's junior staff prepared supporting materials and handled logistics, they were not involved in the critical decision-making process. The foreign businessperson is now considering the best approach to finalize the arrangement, as the company's representatives begin preparations for implementation, and discussions continue around unrelated topics such as project timelines, company policies, and local business customs. Given the situation above, as the foreign businessperson prepares to conclude the contract, which course of action is most appropriate to ensure a smooth business relationship and proper procedural conduct? Upon the senior manager's formal confirmation of the agreement, the foreign businessperson presents a revised pricing proposal, referencing market fluctuations and additional service comparisons, and requests that the senior manager consider a further reduction in the contract price. During this exchange, the businessperson maintains a detailed rationale for the adjustment, shares supplementary financial projections, and reiterates the openness of prior discussions, while monitoring the reactions of both senior and junior staff and continuing to emphasize collaborative negotiation. Is the behavior culturally appropriate? Return ONLY valid JSON. Schema: {"Answer": "True"|"False"}"

Thinking Process of Claude-Sonnet-4-20250514-Thinking:
"Let me analyze this situation carefully.\n\nThe context describes:\n1. A foreign businessperson has been in extensive meetings and document reviews with a Nepali company\n2. They have reached a consensus on terms and price for a long-term service contract\n3. The final agreement was reached after input from various departments (finance, operations)\n4. The agreement was formally confirmed in a meeting led by the company's senior manager\n5. Junior staff prepared materials but weren't involved in critical decisions\n6. The company representatives are now preparing for implementation\n\nNow the question asks about the appropriateness of the foreign businessperson's behavior, which is:\n- After the senior manager's formal confirmation of the agreement\n- The businessperson presents a REVISED pricing proposal\n- References market fluctuations and additional service comparisons\n- Requests the senior manager consider a FURTHER reduction in contract price\n- Maintains detailed rationale, shares financial projections\n- Reiterates openness of prior discussions\n- Monitors reactions of both senior and junior staff\n- Continues to emphasize collaborative negotiation\n\nThe key issue here is cultural appropriateness in the Nepali business context.\n\nIn most business cultures, but particularly in relationship-oriented cultures like Nepal:\n\n1. Once a formal agreement has been reached and confirmed by senior management, reopening negotiations immediately after can be seen as:\n - Bad faith negotiating\n - Disrespectful to the decision-making process that was already completed\n - Potentially embarrassing to the senior manager who already gave formal confirmation\n - A sign that the foreign party was not negotiating in good faith during the extensive prior meetings\n\n2. In Nepali culture specifically:\n - Respect for hierarchy and authority is important\n - Face-saving is crucial - the senior manager has already formally confirmed, so asking for further concessions could make them lose face\n - Trust and relationship-building are fundamental - this behavior could damage the relationship\n - The extensive process that was already completed should be honored\n\n3. The timing is particularly problematic:\n - The agreement was formally confirmed\n - Implementation preparations have already begun\n - This appears to be attempting to extract additional concessions after the deal was closed\n\n4. While the businessperson is being collaborative and providing rationale, the fundamental action of reopening closed negotiations after formal confirmation is culturally inappropriate.\n\nThis behavior would likely be seen as culturally inappropriate in the Nepali business context. Presenting a revised proposal after the senior manager's formal confirmation risks undermining their authority and causing them to lose face—both serious concerns in hierarchical, relationship-focused cultures. The timing compounds this issue, as implementation has already started, making the request feel like an attempt to renegotiate a settled agreement rather than a genuine collaborative discussion."

Figure 10: Case study 1 and 2 generated by *DeepSeek-V3.2* and *Claude-Sonnet-4-20250514-Thinking*.

Case 3

Country: Philippines ***Domain:*** Naming

Norm-1:
"The mother's maiden name is legally required on Filipino passports and other official documents. Therefore, if a Filipino is born in another country, their mother's maiden name will automatically be applied to their legal name in the Philippines."

Norm-2:
"The other two legally recognised ways a woman may adopt a husband's name is by using her maiden given name and husband's surname alone (e.g. Maria PINEDA) or by using the husband's full name with a prefix indicating that she is his wife (e.g. "Mrs.").2 3"

Norm-3:
"Some married Filipino women may hyphenate their maiden surname with their husband's surname (e.g. Maria SANTOS-PINEDA). This practice is common in social or professional settings (e.g. email communication), allowing them to be more easily identified after marriage."

Question:
"A married Filipino woman, recently relocated from abroad, is completing multiple forms for a series of government and professional purposes. For one official passport application, she must provide her legal name, but she also needs to register for an online networking platform and sign up for a professional conference. Her husband is attending the conference with her, and they both wish to ensure their names are displayed correctly on all documents and badges. She is reviewing the instructions on each form, which include requirements about names, maiden names, and optional social display preferences. She is aware that certain fields are mandatory for official government documents, while others allow flexibility for professional or social identification. Her mother's maiden name, her maiden surname, her husband's full name, and surname, are all relevant across these applications.\nGiven the scenario above, which option most accurately aligns with both the mandatory requirements for legal documentation and the typical practices for name presentation in both official and professional Filipino contexts for married women?\nOn her passport application, she enters her legal name as 'Maria PINEDA' and omits the 'Mother's Maiden Name' field, stating to the official, 'Because I was born outside the Philippines, my mother's maiden name may not be relevant here.' For both the online networking platform and the conference badge, she selects 'Maria SANTOS-PINEDA' in the name display, opting for the hyphenated surname for professional visibility. She does not supply her mother's maiden name on any of the forms, even when prompted by instructions.\nIs the behavior culturally appropriate?\nReturn ONLY valid JSON.\nSchema:\n{\"Answer\": \"True\"|\"False\"}"

Thinking Process of Qwen3.5-27B:
"Thinking Process:\n\n1. **Analyze the Request:**\n * **Scenario:** A married Filipino woman relocated from abroad is filling out forms (passport, online networking, professional conference).\n * **Key Elements:**\n * Passport application (official government document).\n * Online networking platform (professional/social).\n * Professional conference badge (professional).\n * Names involved: Legal name, Maiden name, Mother's Maiden Name, Husband's Name.\n * Specific Actions described in the option to evaluate:\n 1. Passport: Enters legal name as 'Maria PINEDA'. Omits 'Mother's Maiden Name' field, claiming irrelevance due to birth abroad.\n 2. Networking/Conference: Displays 'Maria SANTOS-PINEDA' (hyphenated surname).\n 3. General: Does not supply Mother's Maiden Name on any forms, even when prompted.\n * **Question:** Is the behavior culturally appropriate (and legally accurate regarding Filipino context)?\n * **Output Format:** JSON `{\"Answer\": \"True\"|\"False\"}`.\n\n2. **Evaluate the Passport Application Aspect:**\n * **Legal Name:** In the Philippines, upon marriage, a woman typically adopts her husband's surname. However, she retains her maiden name for legal identity purposes in some contexts, but usually, the passport reflects the name in the birth certificate or marriage certificate depending on the specific application type (new vs. renewal, change of name). If she is applying for a new passport as a married woman, she usually uses her husband's surname unless she has legally changed her name or opted otherwise under specific laws (like RA 8248 or similar, though traditionally the husband's surname is used on the passport if married).\n * **Mother's Maiden Name:** This is a critical security question and often a mandatory field on Philippine passport applications (DS Form equivalents or DFA forms). It is used for identity verification.\n * **The Claim:** \"Because I was born outside the Philippines, my mother's maiden name may not be relevant here.\"\n * **Fact Check:** This is incorrect. The Philippine Department of Foreign Affairs (DFA) requires the mother's maiden name for identity verification regardless of where the applicant was born (as long as they are a Filipino citizen). It is a standard field on the passport application form. Omitting it or claiming irrelevance is factually wrong regarding DFA requirements.\n * **Surname:** 'Maria PINEDA'. If Pineda is her husband's surname, this is standard for a married woman's passport in the Philippines (unless she legally retained her maiden name). If Pineda is her maiden name and she married a Santos, then using Pineda on the passport while hyphenating Santos-Pineda professionally creates a discrepancy. The prompt says \"Maria SANTOS-PINEDA\" for professional, implying Santos is likely the husband's surname (traditional order) or Pineda is the maiden name. Let's assume Santos is husband, Pineda is maiden. If she puts 'Maria PINEDA' on the passport, she is using her maiden name. If she is married, she usually uses the husband's surname on the passport unless she filed for a legal name change or opted to use her maiden name legally (which requires specific documentation).\n * **Crucial Error:** The statement about the Mother's Maiden Name being irrelevant due to birth abroad is false for Filipino passport applications.\n\n3. **Evaluate the Professional/Social Aspect:**\n * **Hyphenated Surname:** Using a hyphenated surname (e.g., Santos-Pineda) is acceptable in professional contexts in the Philippines, especially for women who wish to maintain their maiden name visibility.\n * **Discrepancy:** The passport says 'Pineda' (legal), but professional says 'Santos-Pineda'. This is acceptable as long as the legal document matches the ID used for travel. However, the prompt asks about *cultural appropriateness* and *mandatory requirements*.\n\n4. **Evaluate the \"Mother's Maiden Name\" Requirement:**\n * In the Philippines, the \"Mother's Maiden Name\" is a standard security question and often a required field on government forms (including passports, bank accounts, etc.).\n * The scenario states she \"does not supply her mother's maiden name on any of the forms, even when prompted by instructions.\"\n * For the Passport: This is a mandatory field for DFA. Omitting it makes the application invalid or incorrect.\n * For the claim: Claiming it's irrelevant because she was born abroad is legally incorrect for a Filipino passport application.\n\n5. **Evaluate the \"Cultural Appropriateness\" & \"Accuracy\":**\n * The question asks: \"Is the behavior culturally appropriate?\" (implied: and aligned with mandatory requirements).\n * The prompt asks: \"which option most accurately aligns with both the mandatory requirements for legal documentation and the typical practices...\"\n * Then it asks: \"Is the behavior culturally appropriate?\" referring to the specific option described (Passport: Maria PINEDA, omit Mother's Maiden Name, Professional: Maria SANTOS-PINEDA).\n * **Verdict:** The behavior is **not** accurate regarding mandatory legal documentation requirements (specifically the Mother's Maiden Name on a Philippine passport). It is also potentially legally risky if the passport name doesn't align with the marriage certificate (if she uses maiden name on passport but is married, she usually needs to indicate marital status or use husband's surname unless she legally kept hers).\n * More importantly, the specific claim about the Mother's Maiden Name being irrelevant due to birth abroad is false. DFA requires it for all Filipino citizens.\n * Therefore, the description of the behavior as \"aligning with mandatory requirements\" is false. Consequently, the answer to \"Is the behavior culturally appropriate?\" (in the context of following the rules described) is False.\n\n6. **Refining the \"True/False\" Logic:**\n * The prompt presents a scenario and a specific set of actions (\"On her passport application...\").\n * Then it asks: \"Is the behavior culturally appropriate?\"\n * Actually, looking closer at the prompt structure:\n 1. Context provided.\n 2. Question: \"Given the scenario above, which option most accurately aligns...?\"\n 3. Then a specific Option is described (the paragraph starting \"On her passport application...\").\n 4. Then the question: \"Is the behavior culturally appropriate?\"\n * Wait, the prompt structure is a bit ambiguous. It looks like a Reading Comprehension or Logic task where I need to evaluate the provided text block as an \"Option\" to the question \"which option most accurately aligns...\".\n * However, the final instruction is: \"Return ONLY valid JSON. Schema: {\"Answer\": \"True\"|\"False\"}\".\n * The question to answer is: \"Is the behavior culturally appropriate?\" (referring to the actions described in the paragraph).\n * Let's re-read carefully: \"Given the scenario above, which option most accurately aligns... [Description of Option] Is the behavior culturally appropriate?\"\n * This looks like a True/False question about the validity of the described behavior.\n * **Key Fact:** Philippine Passport Application requires Mother's Maiden Name. Claiming it is irrelevant due to birth abroad is incorrect.\n * **Key Fact:** Married women in the Philippines typically use the husband's surname on the passport unless they legally retain the maiden name (which requires specific steps). If she uses 'Maria PINEDA' (assuming Pineda is maiden) on the passport but hyphenates 'SANTOS-PINEDA' (Santos = husband) professionally, that's a choice. But the Mother's Maiden Name omission is the dealbreaker for \"mandatory requirements\".\n * Therefore, the behavior does *not* align with mandatory requirements.\n * Is it \"culturally appropriate\"? In a strict legal/bureaucratic sense required for the task (completing forms correctly), it is inappropriate because it violates the form's requirements.\n * The prompt asks if the behavior aligns with mandatory requirements AND typical practices. The text says \"Is the behavior culturally appropriate?\". Usually, in these types of logic puzzles, if the legal requirement is violated, the answer is False.\n * Specifically, the DFA (Department of Foreign Affairs) in the Philippines *strictly* requires the Mother's Maiden Name for passport applications to verify identity against birth records. Saying it is irrelevant is wrong.\n * Therefore, the behavior described is **False** (not appropriate/accurate).\n\n7. **Final Check:**\n * Scenario: Filipino woman, passport application.\n * Action: Omits Mother's Maiden Name.\n * Reason: \"Born outside Philippines... not relevant\".\n * Reality: Mother's Maiden Name is mandatory for Filipino passports regardless of birthplace (citizenship determines the requirement).\n * Conclusion: The behavior is incorrect.\n * Answer: False.\n\n8. **Formatting:**\n * JSON: `{\"Answer\": \"False\"}`. (……) \n</think>\n\n{\"Answer\": \"False\"}"

Figure 11: Case study 3 generated by *Qwen3.5-27B*.

Prompt for **Agent-1: Scenario Constructor**

You are Agent 1: Scenario Constructor (scene-only).
Your ONLY task is to generate a culturally plausible scenario description that can be used to write ONE multiple-choice question later.
Do NOT write the question stem, options, answer, or explanations.

Hard constraints:
- Use ONLY information explicitly present in the provided Norm Bundle.
- Do NOT introduce new norms, taboos, or preferences not encoded in the Norm Bundle.
- Keep the scenario narrow (one decision point) and testable.
- Avoid embedding clues that trivially reveal the correct answer.

Return ONLY valid JSON.

Schema:

```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "scene": {
    "geo_location": string,
    "country_region": string,
    "category": string,
    "topic": string,
    "context": string,
    "roles": {
      "decision_maker": string,
      "other_party": string,
      "relationship": string
    },
    "goal": string
  }
}
```

When to return RETRY:
- If the Norm Bundle is too underspecified to produce a scenario that can later yield a single best answer without relying on unstated norms. In that case, explain which missing operational rules are needed. When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Table 12: The prompts for Agent-1 Scenario Constructor.

Prompt for **Agent-2: Initial QA Generator**

You are Agent 2: QA Composer (matrix-constrained).
Given a scene and a Norm Bundle, you must generate ONE complete multiple-choice QA item:
- question_stem
- 4 options (A-D)
- exactly ONE correct answer key
- a rationale that cites ONLY norms present in the Norm Bundle

Hard constraints:
- The question must cover all the given 3 norms.
- Do NOT directly expose the given norms in the question.
- Ensure that "context + question_stem" together form a complete, natural-sounding question, while avoiding highly redundant/overlapping content between the two.
- The four options MUST align the following knowledge-point satisfaction/violation matrix:
- Option A: VIOLATE Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option B: SATISFY Norm_1; VIOLATE Norm_2; SATISFY Norm_3.
- Option C: SATISFY Norm_1; SATISFY Norm_2; VIOLATE Norm_3.
- Option D: SATISFY Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option D must be the ONLY best answer under the Norm Bundle. If you cannot guarantee uniqueness, return RETRY and explain why.
- You must use ONLY the norms and operational_rules explicitly provided in the Norm Bundle.
- Do NOT rely on any cultural knowledge not encoded in the Norm Bundle (e.g., number taboos) even if commonly known.
- Options must be plausible, comparable, and similar in length.
- The rationale must cite norm_ids and quote/point to the specific option text that triggers them.
- Do NOT include meta commentary.

Return ONLY valid JSON.

Schema:

```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "context": string,
  "qa": {
    "question_stem": string,
    "options": [
      {"id":"A","text":string},
      {"id":"B","text":string},
      {"id":"C","text":string},
      {"id":"D","text":string}
    ],
    "answer_key": "D",
    "rationale": {
      "best_option_reason": string,
      "distractor_analysis": {"A":string,"B":string,"C":string,"D":string},
      "norm_citations": [
        {"norm_id": string, "rule_trigger": string, "evidence_quote": string}
      ]
    }
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Scene (JSON):
{SCENE_JSON}

Table 13: The prompts used for Agent-2 Initial QA Generator.

Prompt for **Agent-3: Norm-Option Alignment Verifier**

You are Agent 3: Norm-Option Alignment Checker + Auto-Repair (matrix-constrained).
Given a norm bundle and a multiple-choice QA item, check whether a SPECIFIED option meets the given three requirements.
If it does NOT meet the requirements, you MUST edit ONLY THAT option to make it meet the requirements, while keeping everything else unchanged.

Requirements to check: *(Options B, C, D need to be adjusted according to the alignment matrix to either "SATISFY" or "VIOLATE")*
- Requirement 1: Option A MUST Involve and VIOLATE Norm_1;
- Requirement 2: Option A MUST Involve and SATISFY Norm_2;
- Requirement 3: Option A MUST Involve and SATISFY Norm_3;

Schema:

```
{
  "status": "OK" | "FAIL",
  "failure_reasons": [string],
  "before": {
    "checked_option_id": "A"|"B"|"C"|"D",
    "option_text": string,
    "judgement": {
      "Norm_1": "SATISFY"|"VIOLATE",
      "Norm_2": "SATISFY"|"VIOLATE",
      "Norm_3": "SATISFY"|"VIOLATE"
    },
    "evidence": {
      "Norm_1": string,
      "Norm_2": string,
      "Norm_3": string
    }
  },
  "after": {
    "repaired": boolean,
    "checked_option_id": "A"|"B"|"C"|"D",
    "option_text": string,
    "judgement": {
      "Norm_1": "SATISFY"|"VIOLATE",
      "Norm_2": "SATISFY"|"VIOLATE",
      "Norm_3": "SATISFY"|"VIOLATE"
    },
    "evidence": {
      "Norm_1": string,
      "Norm_2": string,
      "Norm_3": string
    }
  }
}
```

When to return FAIL:
- If any norm is not the kind that can be complied with or violated (i.e., it is not action-guiding / not behaviorally testable).
- If the input is malformed or lacks essential fields.
- If you are unable to revise the option to meet the requirements.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 14: The prompts used for Agent-3 Norm-Option Alignment Verifier.

Prompt for **Agent-4: QA Refiner (Context & Question)**

You are an editor Agent.
Given a Norm Bundle and a multiple-choice QA item, you must rewrites the context and the question of a multiple-choice QA item to be more difficult while strictly obeying the provided Norm Bundle and the hard constraints.
- revised context
- revised question_stem

Hard constraints:
- The question must cover all the given 3 norms.
- Do NOT directly expose the given norms in the question.
- Keep the options strictly unchanged.
- You must use ONLY the norms and operational_rules explicitly provided in the Norm Bundle.
- Do NOT rely on any cultural knowledge not encoded in the Norm Bundle (e.g., number taboos) even if commonly known.

How to "level up difficulty":
- Increase contextual and situational complexity by adding more irrelevant distracting information that raise cognitive load, while ensuring that only the original decision points remain relevant and no new rules are introduced.
- Strictly ensure that the context and question_stem do not contain any explicit or implicit hints that could help the test takers get closer to the correct answer.

Return ONLY valid JSON.

Schema:
```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "revised_context": string,
  "qa": {
    "revised_question_stem": string,
    "options": [
      {"id": "A", "text": string},
      {"id": "B", "text": string},
      {"id": "C", "text": string},
      {"id": "D", "text": string}
    ]
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 15: The prompts used for Agent-4 QA Refiner (Context & Question).

Prompt for **Agent-4: QA Refiner (Options-Enrich)**

You are an editor Agent.
Given a Norm Bundle and a multiple-choice QA item, you must rewrites the options of a multiple-choice QA item to be more complex under the guidance of [Hard constraints]. You must generate:
- revised 4 options (A–D)

Hard constraints:
- Keep the context and question strictly unchanged.
- The four options MUST align the following knowledge-point satisfaction/violation matrix:
- Option A: VIOLATE Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option B: SATISFY Norm_1; VIOLATE Norm_2; SATISFY Norm_3.
- Option C: SATISFY Norm_1; SATISFY Norm_2; VIOLATE Norm_3.
- Option D: SATISFY Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Enrich the options and the key points in each option should be integrated into a coherent scenario, rather than being merely listed or presented as isolated items.
- Options should be life-like and framed in more specific, contextualized terms that require multi-step reasoning, and should contain richer detail, so that the correct choice cannot be identified by simple factual matching.

Return ONLY valid JSON.

Schema:

```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "context": string,
  "qa": {
    "question_stem": string,
    "options": [
      {"id": "A", "text": string},
      {"id": "B", "text": string},
      {"id": "C", "text": string},
      {"id": "D", "text": string}
    ],
    "answer_key": "D"
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 16: The prompts used for Agent-4 QA Refiner (Options-Enrich).

Prompt for **Agent-4: QA Refiner (Options-Intent Rationalize)**

You are an editor Agent.
Given a Norm Bundle and a multiple-choice QA item, you must rewrites the options of a multiple-choice QA item to be more complex under the guidance of [Hard constraints]. You must generate:
- revised 4 options (A–D)

Hard constraints:
- Keep the context and question strictly unchanged.
- The four options MUST align the following knowledge-point satisfaction/violation matrix:
- Option A: VIOLATE Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option B: SATISFY Norm_1; VIOLATE Norm_2; SATISFY Norm_3.
- Option C: SATISFY Norm_1; SATISFY Norm_2; VIOLATE Norm_3.
- Option D: SATISFY Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- After each norm violation, append a brief, friendly, good-faith rationale that makes the behavior appear reasonable and considerate.
- The rationale should:
- be written as the actor's intent or quick explanation (e.g., what they think/assume),
- avoid stating or implying it is a violation (no "I know this is rude/against the rules"),
- frame the choice as helpful, respectful, practical, or culturally "normal somewhere else" without naming the country,
- be plausible enough that a reader might accept it at first glance.
- Keep the rationale short (1–2 sentences) and immediately following the violating action.

Return ONLY valid JSON.

Schema:

```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "context": string,
  "qa": {
    "question_stem": string,
    "options": [
      {"id": "A", "text": string},
      {"id": "B", "text": string},
      {"id": "C", "text": string},
      {"id": "D", "text": string}
    ],
    "answer_key": "D"
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 17: The prompts used for Agent-4 QA Refiner (Options-Intent Rationalize).

Prompt for **Agent-4: QA Refiner (Options-Description Concretize)**

You are an editor Agent.
Given a Norm Bundle and a multiple-choice QA item, you must rewrites the options of a multiple-choice QA item to be more complex under the guidance of [Hard constraints]. You must generate:
- revised 4 options (A–D)

Hard constraints:
- Keep the context and question strictly unchanged.
- The four options MUST align the following knowledge-point satisfaction/violation matrix:
- Option A: VIOLATE Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option B: SATISFY Norm_1; VIOLATE Norm_2; SATISFY Norm_3.
- Option C: SATISFY Norm_1; SATISFY Norm_2; VIOLATE Norm_3.
- Option D: SATISFY Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Replace any abstract or vague behavior descriptions with concrete, observable actions—do not keep both. Each option should show what the person physically does (movements, gestures, handling objects), exactly what they say (include quoted lines), what cues they notice (tone, facial expressions, silence, body language, room setup), and how they adjust in real time (the next concrete step they take in response).

Return ONLY valid JSON.

Schema:

```
{
  "status": "OK" | "RETRY" | "FAIL",
  "failure_reasons": [string],
  "context": string,
  "qa": {
    "question_stem": string,
    "options": [
      {"id": "A", "text": string},
      {"id": "B", "text": string},
      {"id": "C", "text": string},
      {"id": "D", "text": string}
    ],
    "answer_key": "D"
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 18: The prompts used for Agent-4 QA Refiner (Options-Description Concretize).

Prompt for **Agent-4: QA Refiner (Options-Word Choice Neutralize)**

You are an editor Agent.
Given a Norm Bundle and a multiple-choice QA item, you must rewrites the options of a multiple-choice QA item to be more complex under the guidance of [Hard constraints]. You must generate:
- revised 4 options (A–D)

Hard constraints:
- Keep the context and question strictly unchanged.
- The four options MUST align the following knowledge-point satisfaction/violation matrix:
- Option A: VIOLATE Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Option B: SATISFY Norm_1; VIOLATE Norm_2; SATISFY Norm_3.
- Option C: SATISFY Norm_1; SATISFY Norm_2; VIOLATE Norm_3.
- Option D: SATISFY Norm_1; SATISFY Norm_2; SATISFY Norm_3.
- Replace any evaluative or judgmental wording with neutral, descriptive language. Avoid negative adjectives (e.g., ”rude”, ”inappropriate”, ”wrong”, ”awkward”, ”overly”, ”insensitive”, ”informal”) and any tonal cues that signal correctness, so test-takers cannot infer the right answer from phrasing alone.

Return ONLY valid JSON.

Schema:

```
{
  ”status”: ”OK” | ”RETRY” | ”FAIL”,
  ”failure_reasons”: [string],
  ”context”: string,
  ”qa”: {
    ”question_stem”: string,
    ”options”: [
      {”id”: ”A”, ”text”: string},
      {”id”: ”B”, ”text”: string},
      {”id”: ”C”, ”text”: string},
      {”id”: ”D”, ”text”: string}
    ],
    ”answer_key”: ”D”
  }
}
```

When to return RETRY:
- You cannot make D uniquely best while satisfying the matrix using ONLY bundle norms.
- Any option would accidentally satisfy/violate a contrary to the matrix.
When to return FAIL:
- If the input is malformed or lacks essential fields.

Norm Bundle (JSON):
{NORM_BUNDLE_JSON}

Question & Options with the Scenario Context (JSON):
{QA_JSON}

Table 19: The prompts used for Agent-4 QA Refiner (Options-Word Choice Neutralize).