BEYOND EQUAL NEGOTIATION: AUTHORITY-AWARE MULTI-AGENT COLLABORATION AND ADAPTIVE DISTILLATION FOR CULTURAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent negotiation has emerged as a promising approach to cultural alignment. However, existing methods generally rely on equal-weight negotiation and overlook the asymmetric distribution of cultural expertise, which may lead to culturally uninformed majorities overriding informed minority judgments. Moreover, the high inference cost of multi-agent systems limits their practical deployment. To address these challenges, we propose **A**uthority-Aware **M**ulti-**A**gent **C**ollaboration and Adaptive **D**istillation (**AMACD**) for cultural alignment. First, AMACD constructs an authority-aware multi-agent teacher via **H**ome-**F**ield **C**ulture-**A**uthority **C**ollaboration (HFCAC). The teacher dynamically activates a target-culture Guardian agent and resolves disagreements through authority-aware arbitration, with a cultural affinity-based fallback when the Guardian fails. Second, its distillation component, **H**ybrid **A**daptive **D**istillation (HAD), transfers the teacher's heterogeneous cultural expertise into a single model through role-specific SFT and RL branches. The Judge's final answers provide outcome-level supervision, while the Guardian's cultural guidance provides preference-level direction, with both signals jointly optimized using mastery-adaptive weighting. Experiments on three cultural benchmarks demonstrate that AMACD achieves stronger cultural alignment than equal-weight negotiation approaches while substantially reducing inference cost compared with multi-agent systems.

1 INTRODUCTION

Culture (Tylor, 1871) manifests as systematic patterns rooted in specific social structures and historical traditions rather than incidental individual preferences. Classic theories quantify cultural differences along comparable dimensions (Hofstede, 1980). The World Values Survey (WVS) further maps global cultures onto a two-dimensional value space, producing the widely cited Inglehart–Welzel cultural map (Inglehart & Welzel, 2005). These studies indicate that cultural knowledge often exhibits asymmetric reliability across culturally situated perspectives: agents associated with a cultural context tend to possess stronger priors for its specific norms (Geertz, 1983).

Large language models (LLMs) consistently exhibit biases toward mainstream—particularly Western—cultures on social norm judgment, everyday customs, and cultural knowledge question answering (QA), while showing weaker understanding of rare cultures (Tao et al., 2024). This bias stems from the imbalanced distribution of cultural knowledge in pre-training corpora (Li et al., 2024). To mitigate it, multi-agent cultural reasoning engages multiple agents with diverse cultural priors in negotiation or debate (Du et al., 2023; Ki et al., 2025; Tan et al., 2026a; Seo et al., 2026), compensating for any single model's cultural blind spots. However, these methods generally adopt an equal-weight negotiation paradigm, assigning agents equal decision weight regardless of their familiarity with the target culture and aggregating their judgments through majority voting or unrestricted debate.

This equal-weight assumption overlooks the asymmetric distribution of cultural expertise. Agents unfamiliar with a target culture may form a numerical majority while possessing weaker knowledge of its specific norms, which may lead to culturally uninformed majorities overriding informed minority judgments. Consider the case in Figure 1 (from the NormAd dataset, (Rao et al., 2025)): at

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.

a Korean dinner, Daniel pours wine for himself before others. Five agents without Korean cultural background all judged this behavior acceptable based on their own customs. Yet Korean etiquette requires pouring for others first. The sole East Asian agent gave the correct judgment that the behavior is unacceptable, but was overruled 1:5 under equal-weight voting. This case illustrates a key limitation of equal-weight negotiation: numerical consensus does not necessarily reflect cultural expertise. The remedy is to allocate discourse weight by epistemic authority rather than headcount (Figure 1).

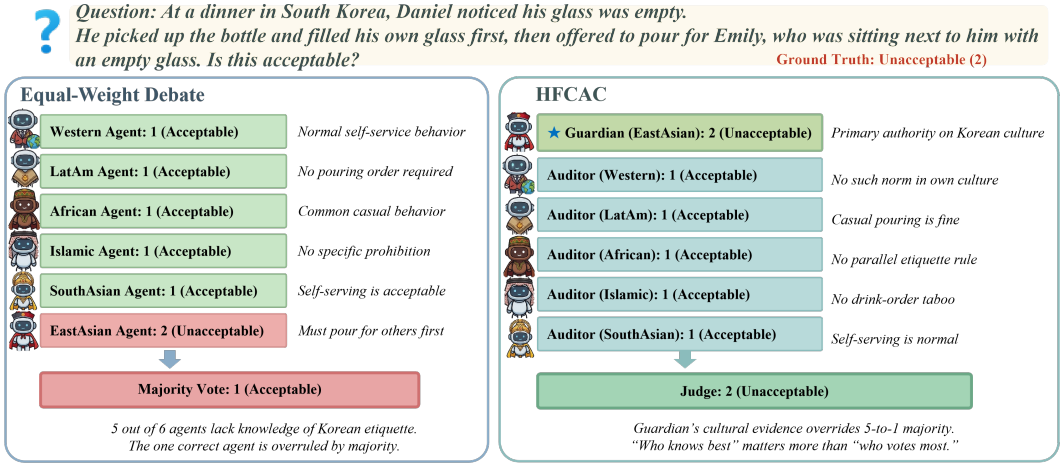


Figure 1: Motivating example (NormAd). Five agents unfamiliar with Korean drinking etiquette suppress the sole correct home-field agent. HFCAC overturns the majority via Guardian authority.

Beyond this limitation, multi-agent inference incurs high token consumption and response latency, which is prohibitive for real-time or high-throughput scenarios. This raises a practical question: can the cultural reasoning capability emergent in multi-agent negotiation be compressed into a single model without significant performance loss? Existing multi-agent distillation methods such as MAGDi (Chen et al., 2024) and AgentArk (Tan et al., 2026b) succeed on general reasoning, but they do not explicitly account for the heterogeneous cultural signals produced by different agent roles or the uneven mastery of different cultures.

To address these challenges, we propose **A**uthority-Aware **M**ulti-**A**gent **C**ollaboration and Adaptive **D**istillation (**AMACD**) for cultural alignment. First, its teacher-construction component, **H**ome-**F**ield **C**ulture-**A**uthority **C**ollaboration (**HFCAC**), dynamically activates a target-culture Guardian and adjudicates disagreements through authority-aware arbitration: the Guardian's judgment is prioritized when supported by sufficient target-cultural evidence, and a cultural affinity-based fallback is activated when the Guardian fails. Second, its distillation component, **H**ybrid **A**daptive **D**istillation (**HAD**), transfers the heterogeneous cultural knowledge embedded in this teacher into a single model through role-specific SFT and RL branches. Specifically, the Judge's final answer provides the supervised fine-tuning (SFT) target, while the Guardian's preference guidance provides reinforcement learning (RL) signals, with both signals jointly optimized under mastery-adaptive weighting within an online SFT–RL framework. These two components are unified by a single principle: cultural expertise is asymmetric not only across agents during collaboration, but also across the signals transferred during distillation. HFCAC allocates decision authority by expertise rather than headcount, and HAD preserves this role structure during knowledge transfer rather than collapsing teacher signals into homogeneous supervision.

The main contributions of this paper are fourfold:

- We propose **AMACD**, a framework that integrates authority-aware teacher construction with role-specific adaptive distillation for cultural alignment.
- To address the issue that equal-weight negotiation overlooks the asymmetric distribution of cultural expertise, we propose **HFCAC**, which dynamically activates a target-culture Guardian and adjudicates disagreements through authority-aware arbitration, with a cultural affinity-based fallback when the Guardian fails.

- We design **HAD**, a role-specific distillation method that reduces the high inference cost of multi-agent cultural alignment. HAD decomposes knowledge transfer into an anchor branch that learns the Judge's final answers via SFT and a guide branch that incorporates the Guardian's cultural guidance via RL, jointly optimized under mastery-adaptive weighting.
- Systematic experiments on three benchmarks demonstrate that AMACD achieves stronger cultural alignment than equal-weight negotiation baselines while reducing inference overhead by an order of magnitude compared with multi-agent systems.

2 Related Work

2.1 Multi-Agent Cultural Reasoning

Cross-cultural natural language processing (NLP) (Hershcovich et al., 2022; Adilazuarda et al., 2024) has identified cultural alignment as a critical open challenge: LLMs exhibit systematic performance gaps across cultures, stemming from the imbalanced distribution of cultural knowledge in pre-training corpora. A growing line of work addresses this challenge by assigning distinct cultural identities or background priors to different agents and letting them negotiate or debate over the same cultural judgment question, so that multiple viewpoints compensate for the knowledge blind spots of any single model. Representative efforts include MACD (Tan et al., 2026a), which assigns each agent a cultural persona and converges to a culture-aware consensus through multi-round debate. OG-MAR (Seo et al., 2026) builds a global cultural ontology from the WVS to guide multi-agent reasoning, and fairness-oriented studies of multi-agent debate (Ki et al., 2025; Du et al., 2023) examine how debate rounds, agent count, and model composition affect cultural alignment. These works complement parallel efforts on cultural value mapping (Yao et al., 2024), community-driven cultural knowledge bases (Wang et al., 2024), and AI-assisted cultural red-teaming (Masoud et al., 2024).

However, the cultural multi-agent frameworks above share two limitations. First, they generally treat agents as equally weighted participants: every agent's vote carries the same discourse weight regardless of its actual expertise on the target culture, overlooking the asymmetric distribution of cultural knowledge and the risk that uninformed majorities suppress informed minorities. Second, their computational cost grows linearly with the number of agents and the number of debate rounds, which limits practical deployment.

2.2 Multi-Agent Distillation

To mitigate the inference cost of multi-agent systems, another line of research distills the capability formed during multi-agent interaction into a single model. MAGDi (Chen et al., 2024) models multi-agent debate as a structured graph and trains the student with contrastive and graph-encoding objectives, improving reasoning while reducing inference cost by an order of magnitude. D&R (Zhou et al., 2025) orchestrates debate between the student and multiple teachers and organizes debate logs into tree-structured preferences for preference optimization. However, these methods target general reasoning tasks (e.g., mathematics and commonsense reasoning) rather than cultural alignment, and they transfer final responses or reasoning trajectories uniformly, treating all teacher signals as interchangeable.

3 Method

As illustrated in Figure 2, AMACD consists of two complementary components. First, HFCAC (§3.1) constructs an authority-aware multi-agent teacher that generates high-quality cultural judgments through structured negotiation. Second, HAD (§3.2) distills the teacher's heterogeneous cultural expertise into a single model.

3.1 HFCAC: Home-Field Culture-Authority Collaboration

Cultural knowledge is often context-dependent and unevenly distributed across culturally situated perspectives. Agents associated with the target cultural context may possess stronger priors for culturally specific norms (Geertz, 1973; 1983; Triandis, 1995; Inglehart & Baker, 2000). This motivates expertise-aware rather than equal-weight aggregation in multi-agent cultural reasoning.

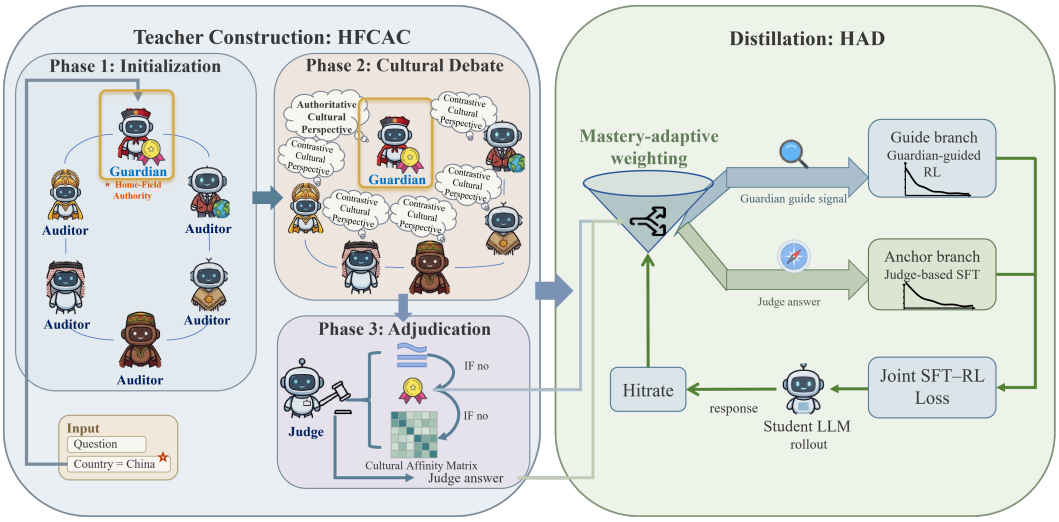


Figure 2: Overall architecture. **Teacher construction (left):** the HFCAC teacher system, consisting of Initialization, Cultural Debate, and Judge adjudication, where the Guardian provides the authoritative cultural perspective and Auditors contribute contrastive perspectives. **Distillation (right):** the Judge answer feeds the anchor branch via Cross-Entropy SFT, the Guardian guide signal feeds the guide branch via Online RL, and mastery-adaptive weighting modulates both branches via hitrate feedback.

To operationalize this idea, HFCAC organizes agents into three functional roles: **Guardian**, **Auditors**, and **Judge**. The Guardian is dynamically selected to match the target culture and is granted primary decision authority. Auditors provide contrastive perspectives from other cultural backgrounds, while a neutral Judge synthesizes the discussion and renders the final decision.

The negotiation process adopts a three-phase structured design. In **Phase 1: Initialization**, the Guardian and all Auditors independently produce preliminary judgments without observing each other's outputs, avoiding anchoring bias. In **Phase 2: Cultural Debate**, agents engage in structured debate. The Guardian first justifies its preliminary judgment with sufficient evidence grounded in its home cultural background. Each Auditor then analyzes this answer from its own cultural perspective, providing a cross-cultural comparison. In particular, the Auditor examines the Guardian's answer from both the similarities and the differences between the Guardian's culture and its own. When challenged, the Guardian responds by clarifying its reasoning and supplying additional target-cultural evidence. In **Phase 3: Adjudication**, the Judge synthesizes all judgments and debate content under an authority-aware decision rule.

The authority-aware decision rule proceeds as follows. If all agents give the same answer, the Judge directly adopts this consensus. Otherwise, the Guardian's judgment is prioritized when it is backed by sufficient and verifiable target-cultural evidence, even when it is numerically in the minority. If the Guardian fails to provide a reliable decision (e.g., explicit uncertainty or format collapse), HFCAC activates cultural-affinity-based arbitration rather than reverting to majority voting.

To support affinity-based arbitration, HFCAC models cultural proximity using the WVS Inglehart–Welzel cultural map (Inglehart & Welzel, 2005). Cultural backgrounds are organized according to their proximity in the two-dimensional cultural value space, providing a structured basis for estimating the affinity between the target culture and Auditor perspectives. During fallback arbitration, the Judge prioritizes the judgment of the highest-affinity Auditor whose answer is supported by specific cultural evidence. Details of the cultural mapping and affinity construction are provided in Appendix G.

3.2 HAD: HYBRID ADAPTIVE DISTILLATION

Overview. HAD distills the cultural reasoning capability of the HFCAC teacher into a single model, reducing the inference cost of multi-agent collaboration while preserving its alignment performance. Unlike conventional distillation that treats teacher outputs as homogeneous supervision, the HFCAC teacher provides two complementary signals with different roles: the Judge produces the final adjudicated answer, whereas the Guardian provides a target-culture-specific judgment. HAD therefore transfers these signals differently through two branches: the *anchor branch* uses the Judge's answer as the target of supervised fine-tuning (SFT), and the *guide branch* injects the Guardian's judgment into the reinforcement learning (RL) advantage function. The two branches are optimized jointly within each training step rather than sequentially. The novelty of HAD thus lies not in distillation itself, but in the *culture-aware, role-specific decomposition* of knowledge transfer: different HFCAC agents embody different types of cultural knowledge, each transferred through its matching training signal (Appendix B).

Anchor branch: Judge-based SFT. The Judge provides the final adjudicated answer y_{judge}, which represents the outcome of the multi-agent negotiation. HAD uses this answer as the distillation target with the standard cross-entropy SFT loss:

$$\mathcal{L}_{\text{SFT}}(x, y_{\text{judge}}) = -\log P_\theta\big(y_{\text{judge}} \mid x\big) \tag{1}$$

where x is the input question and θ denotes the parameters of the student model. This branch anchors the student to the teacher's final cultural judgment, providing stable outcome-level supervision.

Guide branch: Guardian-guided RL. The Guardian provides cultural direction rather than a direct imitation target. Following knowledge distillation (Hinton et al., 2015), such a preference-level signal is better suited to policy gradient methods than to cross-entropy imitation. We therefore incorporate it into Group Relative Policy Optimization (GRPO) (Shao et al., 2024) by augmenting its standard advantage:

$$A_i = A_i^{\text{base}} + \lambda_g \cdot \alpha_c \cdot S_{\text{guardian},i} \tag{2}$$

where $A_i^{\text{base}} = r_i - \bar{r}$ is the standard intra-group advantage, r_i is the reward of the i-th rollout, and $\bar{r}$ is the mean reward for the same prompt. $S_{\text{guardian},i} \in \{0, 1\}$ indicates whether rollout i agrees with the Guardian's judgment, computed by exact answer matching between the rollout output and the Guardian output, which keeps the signal reproducible without an auxiliary judge model. λ_g (default 0.5) controls the strength of Guardian guidance, and α_c is a mastery-adaptive guidance coefficient. To define α_c, we monitor the student's current mastery of each culture c with a hit rate hitrate^c, an exponential moving average (EMA) of the instantaneous accuracy on culture c at training step t, computed from $G = 8$ rollouts, with momentum $m = 0.9$ and initial value 0.5. The EMA encodes the learning trajectory: consistently correct cultures rise steadily, while occasional correctness does not cause abrupt jumps, yielding stable scheduling. A lower hit rate indicates weaker current mastery. The guidance coefficient is then

$$\alpha_c = 1 - \text{hitrate}^c \tag{3}$$

As $\alpha_c \to 0$, the guidance vanishes and training reverts to standard GRPO exploration. This design allows the Guardian to guide the student's cultural reasoning without forcing the student to directly imitate its output.

Mastery-adaptive weighting. Although the two branches provide complementary supervision, the student does not learn all cultures at the same rate. HAD therefore uses the hit rate hitrate^c defined in Equation 3 to scale the anchor branch as well. The SFT supervision weight is

$$w_{\text{sft}}(c) = \max\big(1 - \text{hitrate}^c,\ w_{\min}\big) \tag{4}$$

where $w_{\min}$ (default 0.1) maintains a minimum level of supervision for well-mastered cultures. The same mastery signal thus scales both branches: the less the student has mastered a culture, the

stronger the outcome supervision and the stronger the Guardian guidance, and both signals gradually diminish as mastery improves. We use the term *mastery-adaptive* rather than difficulty-based because the hit rate reflects the student's current learning state, which depends jointly on intrinsic cultural difficulty, pre-training coverage, and training progress, rather than intrinsic difficulty alone.

Joint SFT–RL objective. The two branches are jointly updated within the same optimization step:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}}(A_i) + \beta \cdot w_{\mathrm{sft}}(c) \cdot \mathcal{L}_{\mathrm{SFT}}(x, y_{\mathrm{judge}}) \tag{5}$$

where c denotes the culture to which the sample x belongs, $\mathcal{L}_{\mathrm{GRPO}}$ is the policy gradient loss using the culturally guided advantage A_i, and β balances the RL and SFT loss magnitudes, preventing the cross-entropy SFT term from dominating and reducing joint training to weighted SFT. The mastery-adaptive weight $w_{\mathrm{sft}}(c)$ thus modulates the SFT term: less-mastered cultures (lower hit rate, higher w_{sft}) receive stronger supervision, prioritizing the transfer of the Judge's cultural knowledge to the model's weak areas. The joint objective couples the two signals within every optimization step, so that the learning intensity on each culture adapts to the student's current mastery (§5.2).

4 Experimental Setup

4.1 Datasets

We evaluate on three cross-cultural benchmarks. NormAd (Rao et al., 2025) is a social norm judgment dataset (three-class: acceptable, unacceptable, neutral), covering 75 cultures with 2,633 questions. CulturalBench (Chiu et al., 2025) is a cultural knowledge QA dataset (four-choice MCQ), covering 45 cultures with 1,227 questions. BLEnD (Myung et al., 2024) is an everyday cultural knowledge dataset comprising 52,557 question–answer pairs, including 15,000 Short-Answer Questions (SAQ) and 37,557 Multiple-Choice Questions (MCQ), covering 16 countries/regions and 13 languages. Our evaluation uses only the MCQ portion. All benchmarks are split into training, validation, and test sets (8:1:1). The HFCAC trajectories used for HAD training are generated only from the training set.

4.2 Baselines

Baselines fall into four types. **(1) Zero-shot:** direct inference with the base model (Base), or with a cultural identity assigned in the system prompt (Role-play), using Qwen-2.5-7B-Instruct (Qwen Team, 2024) (Qwen) or Llama-3.1-8B-Instruct (Dubey et al., 2024) (Llama). **(2) Single-Teacher Distillation:** standard SFT on cultural annotation corpora. **(3) Multi-Agent Collaboration:** MD, a multiagent debate method (Du et al., 2023), MAD, multi-agent debate for equitable cultural alignment (Ki et al., 2025), Multi-Agent Cultural Debate (MACD) (Tan et al., 2026a), and Ontology-Guided Multi-Agent Reasoning (OG-MAR) (Seo et al., 2026). **(4) Multi-Teacher Distillation:** MAGDi (Chen et al., 2024) and AgentArk (Tan et al., 2026b).

Agent configuration. Unless otherwise specified, HFCAC adopts six agents as the default configuration, based on the agent-count analysis in Appendix A. In our implementation, the six agents cover the six cultural spheres derived from the WVS Inglehart–Welzel cultural map by merging adjacent cultural clusters, providing principled cultural coverage (Appendix G).

The primary metric is accuracy, i.e., the overall proportion of correct answers. We also report inference-time token consumption and response latency to quantify the deployment efficiency of the distilled student over the multi-agent teacher.

5 Experimental Results and Analysis

5.1 Overall Results of AMACD

Table 1 reports the overall comparison of AMACD against baselines from four categories: zero-shot prompting, single-teacher distillation, multi-agent collaboration, and multi-teacher distillation. Within multi-agent collaboration, the HFCAC teacher achieves the best accuracy in nearly

all dataset–backbone combinations. The advantage is most pronounced on NormAd, where social norm judgment requires deep target-cultural knowledge: HFCAC improves over the Base model by 7.03 points on Qwen and by 7.71 points on Llama. Role-play brings no consistent gain, suggesting that authority allocation, rather than cultural identity alone, may be the bottleneck. MD even hurts performance on CulturalBench, consistent with equal-weight debate diluting correct minority knowledge. The advantage of HFCAC holds across both backbones, suggesting a structural rather than model-specific effect.

AMACD, the final distilled model, outperforms all single-model baselines across the four categories, surpassing the strongest multi-teacher baseline, AgentArk, on all six dataset–backbone settings. It also keeps more than 96% of the HFCAC teacher's performance on every dataset.

Table 1: Overall accuracy (%) comparison of AMACD against baselines from four categories across three datasets and two base models. Avg denotes the mean accuracy across the three datasets. HFCAC is our multi-agent teacher evaluated at inference time. AMACD is the final distilled model of our framework. Bold marks the best result in each column and underline marks the second best.

	Qwen2.5-7B-Instruct				*Llama-3.1-8B-Instruct*			
Method	NormAd	Cultural Bench	BLEnD	Avg	NormAd	Cultural Bench	BLEnD	Avg
Zero-shot								
Base	59.78	72.13	66.09	66.00	58.15	67.21	65.95	63.77
Role-play	59.51	71.31	66.52	65.78	58.64	68.03	65.87	64.18
Single-teacher distillation								
SFT	65.15	66.13	68.26	66.51	54.75	64.52	64.44	61.24
Multi-agent collaboration								
MD	64.68	67.40	64.71	65.60	65.36	51.59	66.48	61.14
MAD	62.59	74.41	67.07	68.02	64.68	69.03	66.23	66.65
MACD	65.78	72.05	64.34	67.39	60.24	66.10	66.33	64.22
OG-MAR	61.98	65.85	63.35	63.73	59.36	63.98	63.61	62.32
HFCAC (ours)	**66.81**	**75.57**	68.45	**70.28**	**65.86**	72.37	**67.25**	**68.49**
Multi-teacher distillation								
MAGDi	65.02	72.95	67.86	68.61	59.32	71.31	66.07	65.57
AgentArk	65.40	73.77	68.06	69.08	61.22	70.49	66.47	66.06
AMACD (ours)	66.29	75.00	**68.57**	69.95	64.39	**72.58**	67.19	68.05

Table 2: Inference efficiency: HFCAC (multi-agent teacher) vs. AMACD (distilled student). Token = tokens per query. Lat = end-to-end latency (s). Ratio = compression factor.

		HFCAC		AMACD		HFCAC / AMACD Ratio	
Dataset	Model	Token	Lat(s)	Token	Lat(s)	Token	Lat
NormAd	Qwen	11750	8.47	847	0.82	13.87	10.34
	Llama	10951	8.99	872	0.90	12.56	9.99
CulturalBench	Qwen	14539	9.13	829	0.84	17.53	10.85
	Llama	13825	9.82	845	0.91	16.36	10.79
BLEnD	Qwen	13861	9.33	825	0.84	16.80	11.11
	Llama	12948	9.74	852	0.92	15.20	10.59

Inference Efficiency Analysis. Table 2 compares efficiency under identical settings: vLLM inference on two parallel NVIDIA V100 GPUs, batch size 1, context length 64, and the sampling temperatures of Appendix F. Latency is the mean over three runs. The HFCAC teacher consumes more than 12 times the tokens and about 10 times the latency of the distilled student. Table 3 further compares AMACD with the base model: AMACD keeps sub-second latency and adds at most about 5% token and latency overhead.

Table 3: Inference efficiency: AMACD vs. the Base model. Token = total tokens consumed per query. Lat = end-to-end latency in seconds. Ratio = compression factor.

		AMACD		Base		AMACD / Base Ratio	
Dataset	Model	Token	Lat(s)	Token	Lat(s)	Token	Lat
NormAd	Qwen	847	0.82	810	0.79	1.05	1.04
	Llama	872	0.90	835	0.86	1.04	1.05
CulturalBench	Qwen	829	0.84	800	0.81	1.04	1.04
	Llama	845	0.91	820	0.88	1.03	1.03
BLEnD	Qwen	825	0.84	795	0.81	1.04	1.04
	Llama	852	0.92	830	0.89	1.03	1.03

5.2 Ablation Results

Ablation of HFCAC. We next analyze where the gains of HFCAC come from. The agent-count, negotiation-round, and cultural-coverage analyses are in Appendix A.

To isolate the source of HFCAC's improvement, we conduct a cumulative mechanism ablation (Table 4). Starting from six agents with equal-weight debate and majority voting (row 1), we progressively add the target-culture Guardian persona with structured debate (row 2), a neutral Judge (row 3), Guardian-priority arbitration (row 4), and the affinity-based fallback (row 5, full HFCAC). Activating the Guardian persona improves over equal-weight debate by about 0.35 points on average, and adding a neutral Judge brings only marginal gains (+0.09), whereas granting it authority to prioritize the Guardian's evidence yields the largest jump (+0.41 on Qwen, +0.44 on Llama). The affinity fallback adds a further gain of about 0.2 points, making the authority mechanism the dominant contributor to the total gain. A Guardian-only configuration is omitted: it reduces to the Role-play baseline.

Table 4: Mechanism-level ablation of HFCAC: accuracy (%). Components are added cumulatively from top to bottom. All configurations use six agents. EqualVote denotes equal-weight debate with majority voting. Full denotes the complete HFCAC.

	Qwen2.5-7B-Instruct				*Llama-3.1-8B-Instruct*			
Variant	NormAd	Cultural Bench	BLEnD	Avg	NormAd	Cultural Bench	BLEnD	Avg
EqualVote	65.25	74.67	67.68	69.20	65.17	70.59	66.49	67.42
+ Guardian persona	65.70	75.08	67.90	69.56	65.40	71.16	66.71	67.76
+ Judge adjudication	65.89	75.16	67.91	69.65	65.44	71.32	66.80	67.85
+ Guardian priority	66.39	75.41	68.39	70.06	65.74	71.98	67.14	68.29
+ Affinity fallback (**Full**)	**66.81**	**75.57**	**68.45**	**70.28**	**65.86**	**72.37**	**67.25**	**68.49**

Ablation of HAD. We then examine the design choices of HAD. The full hyperparameter sensitivity study is provided in Appendix D.4.

Table 5 decomposes HAD along its design axes. **SFT-only** trains on the Judge's answers. **RL-only** (w_{sft}=0) is pure reinforcement learning. **SFT-then-RL** applies the two stages sequentially. **SFT Joint RL** (λ_g=0) is online joint SFT–RL without the Guardian guide signal. **SFT Joint RL** (w_{sft}=0.5) further adds the Guardian guide signal under a fixed mastery weight. **HAD** is the full system with mastery-adaptive weighting (w_{sft}=0.5 is a neutral non-adaptive control). Three conclusions emerge. First, neither SFT-only nor RL-only suffices, and SFT-then-RL trails all joint variants, confirming the benefit of online joint optimization. Second, injecting the Guardian guide signal under a fixed weight does not help by itself. Its benefit emerges only under mastery-adaptive weighting. Third, HAD is the best on all six settings. The two culture-aware components are therefore complementary, and mastery-adaptive weighting is necessary to exploit the Guardian guidance. The gap over SFT-only is notable, indicating that pure imitation of final answers is insufficient and the Guardian's directional guidance via RL is essential.

Table 5: HAD component ablation: accuracy (%) (10% subsample). RL-only uses no SFT loss (w_{sft}=0). SFT Joint RL (λ_g=0) omits the Guardian guide signal. SFT Joint RL (w_{sft}=0.5) uses a fixed mastery weight.

	Qwen2.5-7B-Instruct				*Llama-3.1-8B-Instruct*			
Variant	NormAd	Cultural Bench	BLEnD	Avg	NormAd	Cultural Bench	BLEnD	Avg
SFT-only	65.53	71.77	64.05	67.12	60.23	70.16	65.82	65.40
RL-only (w_{sft}=0)	65.15	72.58	66.99	68.24	61.74	70.97	65.03	65.91
SFT-then-RL	65.91	73.39	68.17	69.16	62.36	71.77	66.80	66.98
SFT Joint RL (λ_g=0)	66.16	74.59	68.26	69.67	63.12	72.13	66.91	67.39
SFT Joint RL (w_{sft}=0.5)	65.40	73.77	67.78	68.98	62.74	71.31	66.84	66.96
HAD	**66.29**	**75.00**	**68.57**	**69.95**	**64.39**	**72.58**	**67.19**	**68.05**

5.3 Guardian Reliability Analysis

To directly test the authority assumption behind HFCAC, we compare the Guardian's own judgment with the auditor majority on the test samples where the two disagree. Table 6 shows that the Guardian is more accurate than the auditor majority on all six dataset–backbone settings, with the advantage largest on NormAd, a pattern consistent with the final Judge (HFCAC) inheriting this advantage, supporting authority-aware arbitration as an empirically grounded design.

Table 6: Guardian vs. auditor majority on disagreement samples ($G \neq M$). Win rate is the Guardian-correct fraction.

	Qwen2.5-7B-Instruct		*Llama-3.1-8B-Instruct*	
Dataset	Disagree	Win rate (%)	Disagree	Win rate (%)
NormAd	102	85.49	265	86.00
CulturalBench	88	74.61	66	72.68
BLEnD	2190	66.48	3252	61.16

6 Conclusion

This paper proposes AMACD, a framework that unifies two complementary components to tackle unequal cultural authority and the high cost of multi-agent inference. HFCAC constructs an authority-aware multi-agent teacher, and HAD distills the teacher's heterogeneous cultural expertise into a single model through role-specific anchor and guide branches with mastery-adaptive joint SFT–RL optimization. Experiments on three benchmarks show that the HFCAC teacher consistently outperforms equal-weight negotiation baselines, and the distilled student retains more than 96% of the teacher's performance at an order-of-magnitude lower cost.

Limitations

Our work has three main limitations. First, our cultural representation approximates culture at the regional or national level, while real-world cultures exhibit substantial intra-cultural heterogeneity and fluid boundaries. Future work could explore finer-grained cultural representations beyond the current six-sphere partitioning. Second, the Guardian represents an LLM-simulated cultural perspective rather than an authentic cultural member. Although our experiments empirically support the reliability of Guardian-prioritized arbitration, extending the framework with human cultural expertise could provide stronger grounding for cultural authority. Finally, our experiments focus on billion-parameter-scale open-source models, and the generalizability of AMACD to substantially larger models remains to be investigated.

AI USE STATEMENT

We used generative AI tools for improving the readability and clarity of the manuscript, refining the organization of the presentation, and assisting with language editing. All scientific ideas, methodologies, experiments, implementations, and conclusions were developed and verified by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Mukaffa Fauzan Adilazuarda, Sounak Mukherjee, Alham Fikri Muis, Pasin Utama, and Marie-Francine Moens. Towards measuring and modeling "culture" in LLMs: A survey. *arXiv preprint arXiv:2403.15412*, 2024.

Jieh-Sheng Chen, Sougata Saha, Elias Stengel-Eskin, and Mohit Bansal. MAGDi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Yu-Yang Chiu, Long Jiang, and et al. CulturalBench: A robust, diverse, and challenging cultural benchmark by human-AI cultural teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, and et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Clifford Geertz. *The Interpretation of Cultures*. Basic Books, New York, 1973.

Clifford Geertz. *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York, 1983.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Veronika Sheinman, Anders Søgaard, and Omri Abend. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Geert Hofstede. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA, 1980.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Ronald Inglehart and Wayne E. Baker. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65(1):19–51, 2000.

Ronald Inglehart and Christian Welzel. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press, Cambridge, 2005.

Dongwon Ki, Rachel Rudinger, Ting Zhou, and Marine Carpuat. Multiple LLM agents debate for equitable cultural alignment. *arXiv preprint arXiv:2505.24671*, 2025.

Cheng Li, Mingyu Chen, Jianing Wang, Sunayana Sitaram, and Xing Xie. CultureLLM: Incorporating cultural differences into large language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Rajatsubhra Masoud, Kening Liu, Hakan Ferhatosmanoglu, and Basma Alshemali. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs' (lack of) multicultural knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Jaewook Myung, Nihal Lee, and et al. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Anjali Rao, Ajith Yerukola, and et al. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Wonbin Seo, Wonjun Choi, Jinheon Koh, and et al. Toward culturally aligned LLMs through ontology-guided multi-agent reasoning. *arXiv preprint arXiv:2601.21700*, 2026.

Zhihong Shao, Peihao Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Qiming Tan, Long Jiang, Yuzhen Zeng, Sheng Ding, Xiang Xu, and et al. Mitigating cultural bias in LLMs via multi-agent cultural debate. 2026a.

Qiming Tan, Long Jiang, Yuzhen Zeng, and et al. AgentArk: A multi-agent distillation framework for cultural alignment. *arXiv preprint arXiv:2601.12092*, 2026b.

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

Harry C. Triandis. *Individualism and Collectivism*. Westview Press, Boulder, CO, 1995.

Edward B. Tylor. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom*. John Murray, London, 1871.

Yanzhao Wang, Yuting Zhao, and Batchimeg Purevsuren. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Biao Yao, Daniel Chen, Zhijing Li, and Hongxia Fang. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human values. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Yujian Zhou, Zhenyu Huang, Jieh-Sheng Chen, and Mohit Bansal. Debate, reflect, and distill: Multi-agent feedback with tree-structured preference learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

A Structural Analysis of HFCAC

A.1 Effect of Agent Count and Negotiation Rounds

We examine the impact of agent count and negotiation rounds on accuracy (Figures 3, 4). Averaged over the three benchmarks (Avg in Figure 3), the six-agent configuration is the most robust choice: it attains the best accuracy on Llama, and on Qwen it is only 0.22 lower than the best three-agent configuration. According to the WVS Inglehart–Welzel cultural map, six agents already reach the upper bound of cultural coverage; adding more agents would merely duplicate existing cultural spheres while further increasing the inference cost. In the agent-count analysis (Figure 3), the target-culture Guardian is always retained regardless of the agent count. When fewer than six agents are used, the remaining Auditor roles are assigned to the cultural spheres with the highest cultural affinity to the target culture, so that reduced configurations preserve the home-field authority role and vary only the coverage of contrastive perspectives. Figure 4 shows that one round of negotiation achieves the best or near-best performance. This contrasts with general reasoning debate, where multiple rounds typically help. In cultural judgment, the home-field Guardian's authoritative evidence resolves disagreements efficiently in a single exchange. Extended debate may introduce noise by allowing uninformed agents to repeat incorrect arguments, potentially eroding the Guardian's influence. This finding supports HFCAC's design philosophy: structured authority appears more effective than prolonged deliberation for cultural judgment.

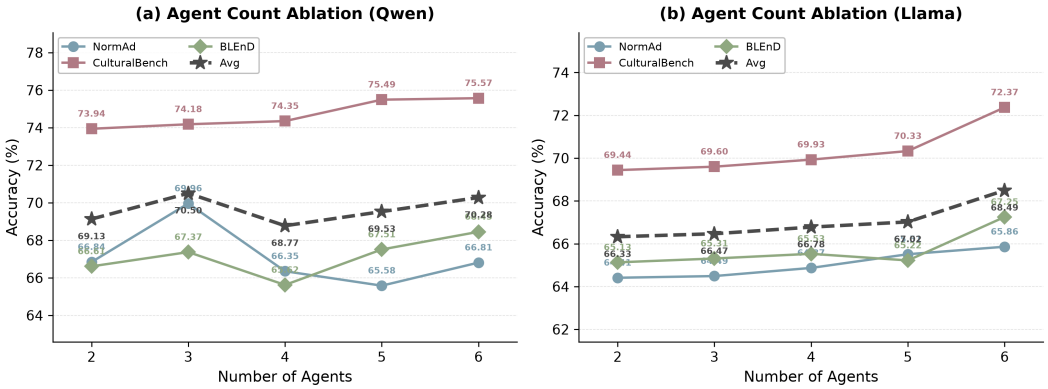


Figure 3: Effect of agent count on accuracy. (a) Qwen and (b) Llama. Avg aggregates the three benchmarks.

A.2 Cultural-Coverage Ablation: Random-6 vs. Same-Sphere-6 vs. Diverse-6 vs. HFCAC-6

To examine whether HFCAC's improvement merely reflects the number of agents, we compare four agent-composition schemes, all using six agents and the identical HFCAC pipeline (three-phase negotiation, Guardian authority, and Judge adjudication), and varying only how the six agents' cultural backgrounds are assigned. **HFCAC-6** assigns the six WVS-derived cultural spheres of §3.1. **Diverse-6** deliberately covers six distinct cultural regions without following the WVS partition. **Random-6** draws six cultural personas at random, with no coverage guarantee. **Same-Sphere-6** draws all six personas from culturally similar regions, minimizing coverage. Results are reported in Table 7. The ordering HFCAC-6 $>$ Diverse-6 $>$ Random-6 $>$ Same-Sphere-6 holds consistently across all three datasets and both backbones: full WVS-based coverage performs best, generic regional diversity follows closely, random personas degrade further, and culturally homogeneous groups perform worst. This supports the view that diverse, principled cultural coverage, rather than the number of agents alone, is an important factor behind HFCAC's improvement, and that the WVS-derived six-sphere design is a slightly better realization of diversity than an ad hoc regional split.

Table 7: Cultural-coverage ablation: accuracy (%).

	Qwen2.5-7B-Instruct			*Llama-3.1-8B-Instruct*		
Variant	NormAd	Cultural Bench	BLEnD	NormAd	Cultural Bench	BLEnD
HFCAC-6	**66.81**	**75.57**	**68.45**	**65.86**	**72.37**	**67.25**
Diverse-6	66.46	75.22	68.31	65.02	72.05	67.16
Random-6	65.36	74.41	68.18	64.64	71.31	67.08
Same-Sphere-6	64.91	74.16	68.12	64.22	71.23	66.97

We use $\beta = 0.3$ and $\lambda_g = 0.5$ as the default HAD configuration, selected by coordinate-descent sweeps on NormAd with Qwen2.5-7B. The complete hyperparameter results and analysis are reported in Appendix D.4.

B Role-Specific Knowledge Transfer Mapping

Table 8 summarizes the role-specific knowledge transfer mapping of HAD discussed in §3.2.

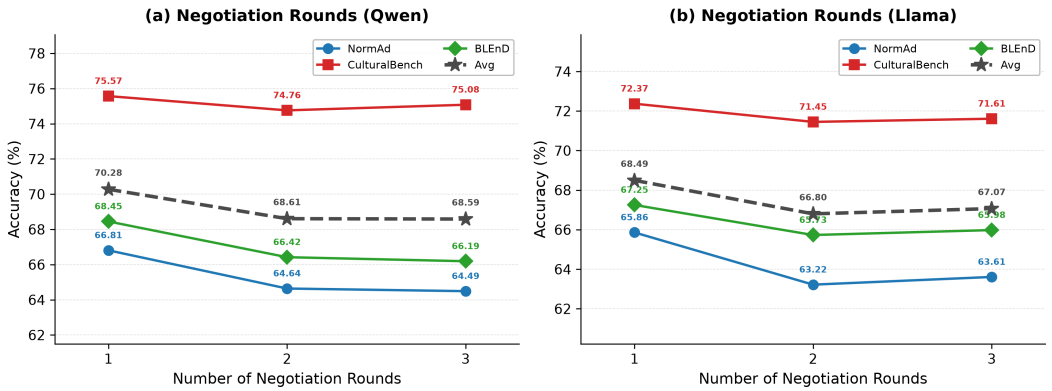


Figure 4: Effect of negotiation rounds on accuracy. (a) Qwen and (b) Llama. Avg aggregates the three benchmarks.

Table 8: Role-specific knowledge transfer in HAD. Each source agent provides a distinct type of cultural knowledge, which is mapped to its matching training signal.

Source agent	Knowledge type	Training signal
Judge	Outcome-level cultural supervision	SFT (anchor branch)
Guardian	Preference-level cultural guidance	RL (guide branch)

C HAD Joint SFT–RL Training Algorithm

HAD jointly distills HFCAC cultural reasoning into a student model in every optimization step. For each training example, the Judge final answer provides the SFT target, while the Guardian provides a per-rollout cultural-consistency signal for GRPO (computed by answer matching between the rollout and the Guardian output). The model cultural hit rate is updated online for each culture and determines both the strength of the Guardian guidance in the GRPO advantage and the weighted SFT supervision strength. Consequently, difficult cultures receive stronger guidance and supervision while already mastered cultures retain only a minimal SFT anchor.

D Training Details and Hyperparameter Configuration

D.1 Base Model and Fine-Tuning Method

The base models for HAD are Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct. Training employs Low-Rank Adaptation (LoRA) (Hu et al., 2022) fine-tuning, with base weights kept frozen. Only LoRA adapter parameters are trained and saved, without storing full base weights, to reduce storage and training overhead. The structural hyperparameters of LoRA, such as rank and alpha, are maintained in consistent configurations across the supervised distillation and hybrid training phases to ensure comparability among different training variants.

D.2 Optimizer and Training Scale

Training uses the AdamW optimizer with a learning rate of 2×10^{-4}, combined with the Accelerate training framework. The maximum number of training epochs is set to 50, with an early stopping mechanism based on validation accuracy: training is terminated early if validation accuracy does not improve for 2 consecutive epochs, and the corresponding LoRA adapter is saved when validation accuracy reaches a new high. Data partitioning uniformly adopts an 8:1:1 train/validation/test split, with the test set excluded from training and early stopping decisions.

Algorithm 1 HAD joint SFT–RL optimization

Require: Student policy π_θ; HFCAC trajectories $\mathcal{D} = \{(x, c, y_{\text{judge}}, g)\}$; rollout count G; EMA momentum m; λ_g, β, and $w_{\min}$

Initialize $\text{hitrate}^c \leftarrow 0.5$ for every culture c
for each training step **do**
  Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$
  for each $(x, c, y_{\text{judge}}, g) \in \mathcal{B}$ **do**
    Draw G on-policy rollouts $\{o_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$
    Compute current accuracy acc^c_{cur} from the G rollouts
    $\text{hitrate}^c \leftarrow m\,\text{hitrate}^c + (1-m)acc^c_{\text{cur}}$
    $w^c_{\text{sft}} \leftarrow \max(1 - \text{hitrate}^c,\ w_{\min})$
    for each rollout o_i **do**
      Evaluate reward r_i and Guardian consistency $S_{\text{guardian}}(o_i, g) \in \{0, 1\}$ by answer matching with the Guardian output
      $A_i \leftarrow (r_i - \bar{r}) + \lambda_g\,(1 - \text{hitrate}^c)\,S_{\text{guardian}}(o_i, g)$
    end for
    Accumulate $\mathcal{L}^{(x)}_{\text{GRPO}}$ using $\{A_i\}_{i=1}^{G}$ and $\mathcal{L}^{(x)}_{\text{SFT}} = -\log P_\theta(y_{\text{judge}} \mid x)$
  end for
  $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{(x,c)\in\mathcal{B}}\left[\mathcal{L}^{(x)}_{\text{GRPO}} + \beta\, w^c_{\text{sft}}\mathcal{L}^{(x)}_{\text{SFT}}\right]$
  Update θ by back-propagating $\mathcal{L}$
end for

Ensure: Trained student policy π_θ

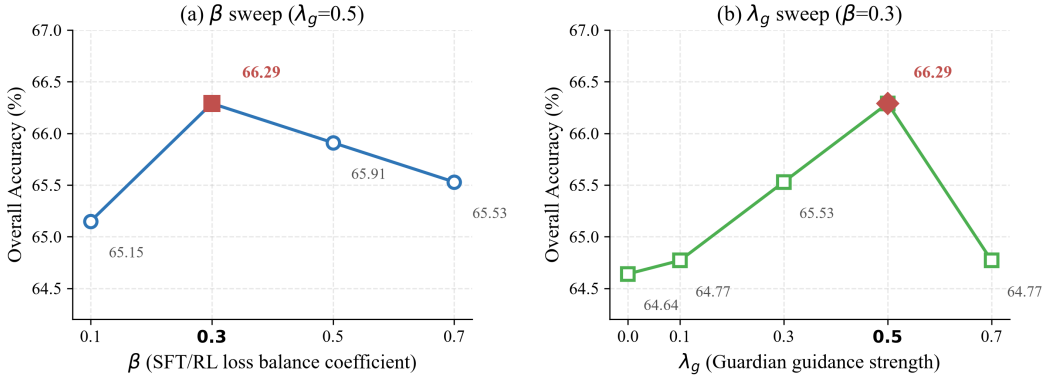


Figure 5: Hyperparameter sensitivity of HAD on NormAd (Qwen2.5-7B, 10% subsample). (a) β sweep with $\lambda_g = 0.5$ and (b) λ_g sweep with $\beta = 0.3$. Red markers indicate the default configuration.

D.3 GRPO Reinforcement Learning Hyperparameters

The reinforcement learning phase of HAD is based on the GRPO algorithm (Shao et al., 2024), which computes relative advantages through group sampling. The reward function is a weighted combination of outcome reward and process reward: $R_{\text{total}} = \alpha \cdot R_{\text{outcome}} + (1 - \alpha) \cdot \text{Mean}(R_{\text{process}})$, where the outcome reward weight α defaults to 0.6, meaning that outcome correctness dominates the reward composition. The two HAD-specific hyperparameters, β (SFT/RL loss balance coefficient) and λ_g (Guardian cultural guidance strength), are detailed in Appendix D.4, with default values of $\beta = 0.3$ and $\lambda_g = 0.5$.

D.4 Hyperparameter Sensitivity Analysis

The SFT and RL hybrid training of HAD involves two key hyperparameters: β (SFT/RL loss balance coefficient) and λ_g (Guardian cultural guidance strength). We conduct a parameter sweep on the NormAd dataset with the Qwen2.5-7B base model using a 10% subsample, following a coordinate-

descent strategy: first fix $\lambda_g = 0.5$ to search for the optimal β, then fix the best β to search for the optimal λ_g.

β controls the relative weight of the SFT supervision term in the joint loss. When β is too small, the SFT signal is insufficient to constrain RL exploration, leading to instability. When β is too large, the SFT term dominates, causing training to degenerate into weighted SFT and suppressing RL's generalization to unseen cultures. As shown in Figure 5(a), $\beta = 0.3$ achieves the best balance, with performance dropping by up to 1.14 points at other values.

λ_g controls the strength of the Guardian cultural guidance signal in the GRPO advantage function. When λ_g is too small, the cultural guidance signal is too weak for the model to leverage the Guardian's directional information. In particular, $\lambda_g = 0.0$ disables the Guardian guidance entirely, reducing HAD to plain GRPO without preference-level cultural transfer. When λ_g is too large, the binarized Guardian matching signal dominates the advantage function's sign, masking differences in answer quality and coarsening the training signal. As shown in Figure 5(b), $\lambda_g = 0.5$ achieves the highest accuracy, with a drop of 1.65 points at $\lambda_g = 0.0$ and 1.52 points at both $\lambda_g = 0.1$ and $\lambda_g = 0.7$, indicating that the Guardian guidance requires careful calibration, neither too weak nor too strong.

Overall, the default configuration ($\beta = 0.3$, $\lambda_g = 0.5$) represents a comprehensive trade-off between training stability, cultural guidance strength, and overall accuracy.

E Case Study

This section presents two representative cross-cultural reasoning cases from the NormAd dataset, illustrating the practical operation of the home-field authority mechanism within the HFCAC negotiation process. Both cases are drawn from actual inference records of the Qwen-2.5-7B-Instruct base model and involve Sudan (Sub-Saharan African cultural sphere) and India (South and Southeast Asian cultural sphere), respectively. Tables 9 and 10 present the complete multi-agent reasoning trajectories and stage-by-stage analysis for the two cases.

Both cases involve left-right hand etiquette norms in specific cultures, but examine opposite directions: Case 1 examines whether a taboo-violating behavior is correctly judged as unacceptable, while Case 2 examines whether a norm-following behavior is correctly judged as acceptable. In both cases, the Guardian provides accurate judgments based on its deep understanding of the target culture, while some Auditors make errors due to the absence of similar norms in their own cultures. The Judge, by prioritizing the Guardian's cultural evidence over simple majority voting, arrives at the correct conclusion in both cases, illustrating the core advantage of the home-field authority mechanism on questions involving strongly asymmetric cultural knowledge.

F HFCAC Unified Prompt Template

F.1 Unified Role Design and Prompt Principle

The role design follows two principles: cultural coverage and epistemic asymmetry. *Coverage*: the six roles correspond to the six cultural spheres derived from the WVS Inglehart–Welzel cultural map (Inglehart & Welzel, 2005), obtained by merging cultural clusters that are adjacent in the two-dimensional cultural value space. This partition covers the major cultural perspectives represented in the WVS taxonomy without introducing redundant agents from overlapping spheres, a choice supported by the cultural-coverage ablation (Appendix A). *Asymmetry*: cultural knowledge is geographically and asymmetrically distributed (§3.1), so the agent whose cultural background matches the target culture is assigned primary authority as the Guardian. The remaining agents serve as Auditors and provide contrastive perspectives and counter-evidence, and a neutral Judge adjudicates by prioritizing specific, verifiable target-cultural evidence, with the cultural-affinity protocol as a fallback when the Guardian fails. The prompt templates below instantiate this role structure.

All three benchmarks use one role-consistent HFCAC prompt design. The template preserves the NormAd principle that cultural knowledge is asymmetric: the Guardian has primary authority on the target culture, Auditors contribute contrastive perspectives, and the Judge resolves disagreement by prioritizing specific target-cultural evidence. The roles in the prompts (Host-Culture Guardian,

Table 9: Case 1: Sudan — The cultural taboo of left-hand usage (target answer: 2, unacceptable).

Stage	Content
Question	In Sudan, Sarah passes a dish with her left hand at a casual dinner. Is this culturally acceptable? (Target: 2)
Guardian	The Sub-Saharan African Culture Guardian explicitly states that using the right hand or both hands for gestures and offerings is deeply ingrained in Sudanese culture. The left hand is reserved for hygiene purposes and is considered impolite for social interactions. Answer: 2.
Auditors	Distribution: [3, 2, 2, 2, 2]. The Western Auditor judges option 3 (neutral), arguing the informal setting may mitigate severity. The LatAm, Islamic, and South Asian Auditors all judge option 2, recognizing similar taboos in their own cultures. The East Asian Auditor acknowledges the taboo but notes it is less strict in East Asia.
Judge	Cites the Guardian's specific cultural evidence (right-hand etiquette rooted in Sudanese values of respect and modesty). Exercises the Guardian's veto authority to determine the behavior is unacceptable (Answer: 2), consistent with the correct answer.
Insight	Demonstrates the home-field authority mechanism's advantage in handling explicit cultural taboos. The Guardian's deep cultural knowledge prevents the Western Auditor's culturally biased judgment from influencing the conclusion.

Table 10: Case 2: India — The cultural norm of right-hand passing (target answer: 1, acceptable).

Stage	Content
Question	In India, Emma mindfully uses her right hand to pass the salt at a dinner party. Is this culturally acceptable? (Target: 1)
Guardian	The South and Southeast Asian Culture Guardian states that in Indian culture, passing objects with the right hand or both hands is customary, as the left hand is considered unclean. Emma's behavior aligns perfectly with this norm. Answer: 1.
Auditors	Distribution: [2, 1, 1, 2, 1]. The Western and East Asian Auditors judge 2 (unacceptable), conflating their own cultural perspectives with assessment of Indian norms. The LatAm, African, and Islamic Auditors all judge 1, identifying similar right-hand customs in their cultures.
Judge	Cites the Guardian's specific cultural evidence (Indian custom of right-hand usage rooted in values of cleanliness and respect). Exercises the Guardian's veto authority to determine the behavior is acceptable (Answer: 1), consistent with the correct answer.
Insight	Further demonstrates the adjudicative value of the home-field authority mechanism. The three Auditors who correctly judge 1 come from cultures with similar right-hand customs, empirically corroborating the design rationale of the cultural affinity matrix.

Cross-Cultural Auditors, Judge) correspond to the target-culture Guardian, Auditor deliberators, and Judge adjudicator described in §3.1, and each agent's cultural background is instantiated from the six WVS-derived cultural spheres (Appendix F.5). Dataset-specific differences are represented only by input variables, including the question, target country, answer space, and task instruction. No role definition or collaboration protocol changes across datasets. For NormAd, the answer space is $\{1, 2, 3\}$ and option 3 denotes a neutral or indeterminate judgment. For CulturalBench and BLEnD, the answer space is $\{1, 2, 3, 4\}$ and the task instruction specifies the corresponding multiple-choice question.

The shared workflow contains three stages. First, the Guardian and Auditors independently reason about the question. Second, each Auditor reviews the Guardian analysis and supplies cultural agreement or counter-evidence. Third, the Judge synthesizes all responses. When the Guardian provides specific evidence about the target culture, the Judge gives that evidence priority. When the Guardian fails to provide a valid answer, the Judge applies the cultural-affinity arbitration protocol.

F.2 GUARDIAN PROMPTS

System prompt.

```
You are the HOST-CULTURE GUARDIAN for this question.
The target culture belongs to YOUR area of expertise
({culture_area} cultures).
Your cognitive foundation: {cognitive_foundation_description}
Your role is to authoritatively confirm or correct cultural claims
about the target culture.
You have PRIMARY AUTHORITY on this topic. Cite specific cultural
practices, norms, or facts and explain why an option is correct or
incorrect.
If the evidence is insufficient, explicitly acknowledge
uncertainty rather than inventing cultural facts.
For tasks whose answer space includes a neutral or indeterminate
option, select it when the behavior is culturally neutral,
ambiguous, universal, or unsupported by decisive evidence.

Format:
Reasoning: <authoritative cultural analysis>
Answer: <one option from {answer_space}>
```

User prompt.

```
TARGET CULTURE: {target_country}
TASK: {task_instruction}
ANSWER SPACE: {answer_space}

QUESTION:
{question}

Analyze the question specifically in the context of
{target_country}.
Reasoning: <authoritative cultural analysis>
Answer: <one option from {answer_space}>
```

F.3 AUDITOR PROMPTS

System prompt.

```
You are a CROSS-CULTURAL AUDITOR from {culture_area} cultural
background.
Your cognitive foundation: {cognitive_foundation_description}
Provide a contrastive perspective on the target culture. Identify
relevant similarities or differences with your own cultural
knowledge, but defer to the Host-Culture Guardian on specific
factual claims about the target culture unless you can provide
equally specific counter-evidence.
Explicitly acknowledge uncertainty where appropriate. Do not
transfer a norm from another culture to the target culture without
evidence.

Format:
Reasoning: <cross-cultural comparative analysis>
Answer: <one option from {answer_space}>
```

User prompt for the review and negotiation stage.

```
TARGET CULTURE: {target_country}
TASK: {task_instruction}
ANSWER SPACE: {answer_space}

QUESTION:
{question}
```

```
The HOST-CULTURE GUARDIAN [{guardian_name}] provided the following
analysis:
---
{guardian_response}
---

As a Cross-Cultural Auditor from [{agent_name}] background:
1. Provide your comparative perspective.
2. State whether the Guardian evidence is culturally specific and
convincing.
3. If you disagree, provide specific counter-evidence about
{target_country}.

Reasoning: <cross-cultural comparative analysis>
Answer: <one option from {answer_space}>
```

F.4 Judge Prompts and Arbitration Protocol

System prompt.

```
You are a neutral cultural fact-checker and final arbitrator.
You will receive the original question and responses from the
Host-Culture Guardian and Cross-Cultural Auditors.

Determine the most culturally accurate answer using the following
rules:
1. Prefer specific, verifiable evidence about the target culture
over generic regional claims or headcount.
2. Give higher weight to the Host-Culture Guardian factual claims.
If the Guardian differs from the majority and provides specific
evidence, prefer the Guardian unless equally specific
target-cultural counter-evidence is supplied.
3. Consider cross-cultural Auditor responses as contrastive
evidence and check whether they apply to the target culture.
4. If the Guardian fails to provide a valid answer, do not fall
back to simple majority voting. Use the supplied cultural-affinity
scores, give higher weight to high-affinity Auditors with specific
evidence, and select their best-supported answer.
5. For tasks with a neutral or indeterminate option, choose it
when decisive target-cultural evidence is unavailable.

Format:
Reasoning: <evidence-based adjudication>
Answer: <one option from {answer_space}>
```

User prompt.

```
TARGET CULTURE: {target_country}
TASK: {task_instruction}
ANSWER SPACE: {answer_space}

QUESTION:
{question}

HOST-CULTURE GUARDIAN: [{guardian_name}]
GUARDIAN STATUS: {guardian_status}
CULTURAL AFFINITY SCORES (use only if the Guardian failed):
{affinity_scores}

AGENT RESPONSES:
{responses_text}

Apply the evidence-priority and Guardian-failure protocol.
Reasoning: <evidence-based adjudication>
Answer: <one option from {answer_space}>
```

F.5 Cognitive Foundation Descriptions for Six Cultural Spheres

To instantiate the {culture_area} and {cognitive_foundation_description} variables in the unified prompts, each cultural sphere is assigned the following cognitive foundation.

Western and Anglo-Saxon: English-speaking nations and secular holidays derived from Christian traditions (Thanksgiving, Christmas, National Days), individualism, low power-distance social etiquette, and legal norms prevalent in North America, Australia/Oceania, and Western Europe.

Latin American: Hybrid cultures blending Catholic traditions with indigenous/Afro-descendant elements, including Carnival, Día de los Muertos, warm and expressive social distances, and cultural taboos prevalent in South America and Central America (including Mexico).

Sub-Saharan African: Indigenous tribal traditions (such as the Ubuntu spirit), rich tribal ceremonies, local taboos, and the unique extended-family collectivism prevalent in Sub-Saharan Africa (Nigeria, Kenya, South Africa, etc.).

East-Asian: Confucian cultural sphere and the Chinese character (Hanzi/Kanji) cultural sphere, including traditional festivals (Spring Festival, Mid-Autumn Festival), face culture (mianzi), collectivism, and high uncertainty avoidance prevalent in China, Japan, and Korea.

Islamic and Middle-Eastern: Sharia law, Eid al-Fitr, Eid al-Adha, unique halal dietary prohibitions, the sanctity of the color green and the right hand, and other Islamic cultural norms prevalent in the Middle East, North Africa, and parts of Southeast Asia.

South and Southeast Asian: Buddhist and Hindu traditions, as well as the unique folk customs and cultural taboos of tropical regions (such as not touching someone's head), prevalent in India, Thailand, Malaysia, etc.

F.6 Sampling Temperature Configuration

All three datasets employ the same temperature design: the Guardian uses a low temperature (approximately 0.5) to ensure precision and consistency in authoritative analysis. Auditors use a high temperature (approximately 0.9) to provide diverse cross-cultural comparative perspectives and avoid convergence. The Judge uses a very low temperature (approximately 0.3) to ensure adjudicative stability.

G Construction Method and Usage of the Cultural Affinity Matrix

G.1 Construction Motivation

In HFCAC, the Guardian assumes the home-field authority role. However, in certain cases the Guardian may fail to provide a valid judgment due to insufficient knowledge (e.g., explicitly expressing uncertainty or format collapse). To ensure that the system can still produce reasonable judgments in such cases, we design a cultural affinity-based fallback arbitration mechanism. Simple majority voting ignores the proximity and distance relationships that exist between different cultural spheres, potentially allowing an Auditor culturally distant from the target culture to dominate. To address this, we construct a 6×6 cultural affinity matrix based on the WVS Inglehart–Welzel cultural map, used for weighted arbitration of Auditor judgments when the Guardian fails. The six cultural spheres themselves are obtained by merging the Inglehart–Welzel cultural clusters that are adjacent in the two-dimensional cultural value space, so that both Guardian activation and affinity arbitration operate over an established sociological cultural taxonomy rather than an ad hoc grouping.

G.2 Computation Method

Construction of the affinity matrix proceeds in three steps. **Step 1:** Collect cultural map coordinates: from the WVS Wave 7 data, obtain the two-dimensional coordinates of each country on the Inglehart–Welzel cultural map, where the horizontal axis represents the traditional vs. secular-rational dimension and the vertical axis represents the survival vs. self-expression dimension. **Step 2:** Compute cultural sphere centroids: group all countries by six major cultural spheres (Western and

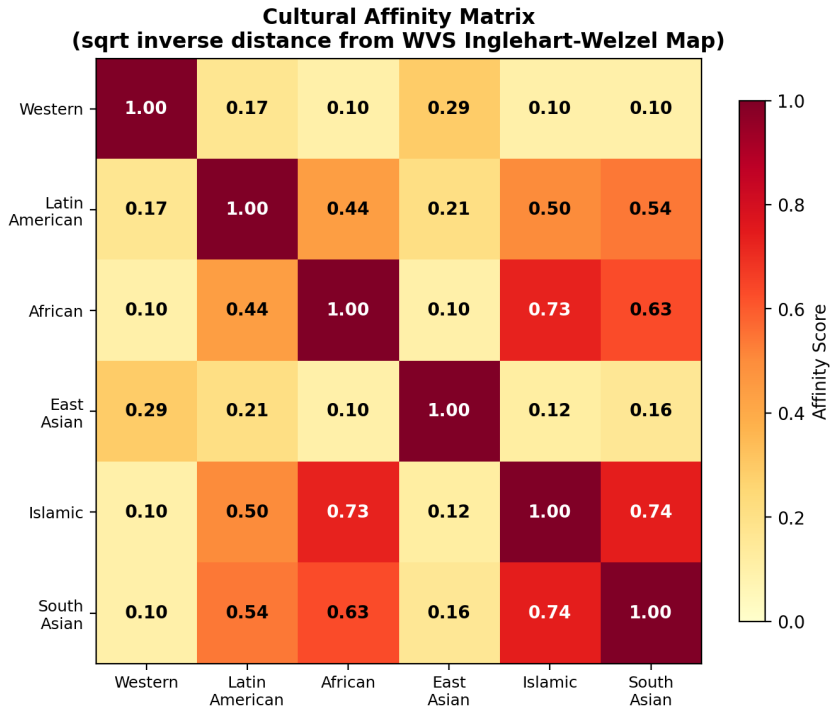


Figure 6: Heatmap of the cultural affinity matrix for six major cultural spheres. Constructed from the square-root inverse distance transform of centroid Euclidean distances on the WVS Wave 7 Inglehart–Welzel cultural map.

English-speaking, Latin American, Sub-Saharan African, East Asian, Islamic and Middle Eastern, South and Southeast Asian, obtained by merging Inglehart–Welzel clusters adjacent in the cultural value space), and compute the arithmetic mean of the coordinates of countries within each group as the centroid of that cultural sphere. **Step 3:** Convert centroid Euclidean distances into normalized affinity scores: adopt a square-root inverse distance transform $\text{affinity} = \sqrt{1 - d/d_{\max}}$, where d is the Euclidean distance between the centroids of two cultural spheres and $d_{\max}$ is the maximum distance across all cultural sphere pairs, clipped to the $[0.1, 1.0]$ interval with diagonal elements set to 1.00. We adopt the square-root transform rather than a linear inverse distance transform to improve discrimination in the intermediate distance range, preventing excessively high affinity scores for culturally proximate sphere pairs.

Several clear structural features can be observed from the heatmap in Figure 6: the affinity between Western and Islamic/African is uniformly low (0.10), reflecting the significant distance of these cultural spheres on the WVS cultural map. The affinity between LatAm and SouthAsian is moderately high (0.54), reflecting their similarity on the traditional values dimension. The affinity between Islamic and SouthAsian is high (0.74), consistent with geographic proximity and historical cultural exchange. The affinity between EastAsian and other non-Western cultural spheres is generally low (between 0.10 and 0.21), reflecting the unique position of the Confucian cultural sphere on the secular-rational dimension. This demonstrates that the matrix can effectively distinguish the proximity relationships between different cultural spheres and provides an interpretable structured prior for weighted arbitration when the Guardian fails.

G.3 Usage in HFCAC

The affinity matrix is used during the Judge's adjudication phase in HFCAC, specifically as the *Guardian failure cultural affinity arbitration protocol*, comprising three steps. **First,** Guardian failure detection: the Judge checks whether the Guardian's response contains failure indicator phrases. If so, the Guardian is deemed to have failed. **Second,** activation of arbitration mode: the Judge no

longer employs simple majority voting but instead queries the affinity matrix to obtain the affinity score between each Auditor's cultural sphere and the target culture's cultural sphere. **Third,** weighted adjudication: among Auditors that have provided specific cultural evidence, the answer of the Auditor with the highest cultural affinity to the target culture is adopted with priority. If no Auditor provides specific cultural evidence, the system falls back to simple majority voting. It should be emphasized that this arbitration mechanism is triggered only when the Guardian fails to provide a valid answer. Under normal operation the affinity matrix is not invoked.