

A Game-Theoretic Negotiation Framework for Cross-Cultural Consensus

Guoxi Zhang^{1*} Jiawei Chen^{1*} Tianzhuo Yang^{1*}
Jiaming Ji¹ Yaodong Yang^{1,2†} Juntao Dai^{1,2†}

¹ Institute for Artificial Intelligence, Peking University

² Beijing Academy of Artificial Intelligence, China

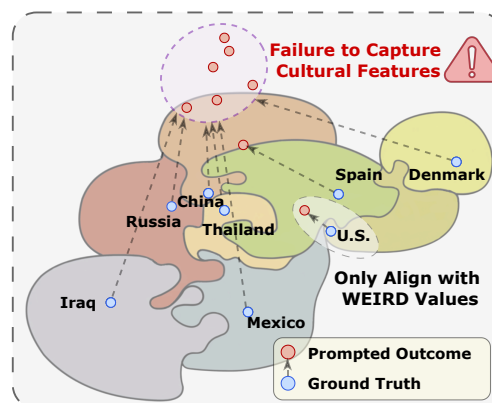
Abstract

Large language models (LLMs) are shaping global values, yet they frequently exhibit a pronounced WEIRD (Western, Educated, Industrialized, Rich, Democratic) cultural bias, marginalizing diverse viewpoints and posing challenges for reconciling diverse populations with varying cultural backgrounds and value systems. In this work, we move beyond simple alignment methods to propose a new paradigm for cross-cultural fairness. We introduce a *Nash Consensus Negotiation* framework under the formulation of cross-cultural consensus as a Nash Equilibrium. Each LLM iteratively proposes and refines natural-language guidelines, guided by a utility function balancing self-consistency with mutual acceptance, while penalizing redundancy. The process expands the proposal space and converges to a consensus, yielding fair and interpretable consensus outcomes. We evaluate our framework against baselines using quantitative metrics, qualitative analysis, and large-scale human studies. Experiments demonstrate that our framework generates higher-quality and more balanced consensus, effectively mitigating assimilation toward WEIRD values. Furthermore, we finetune diverse LLM architectures with negotiation data via preference optimization and supervised reasoning, reducing cultural distances by up to 95.53%. Overall, our work offers a systematic path to mitigate cultural bias in LLMs by guiding them toward self-consistency, mutually-acceptable equilibria.

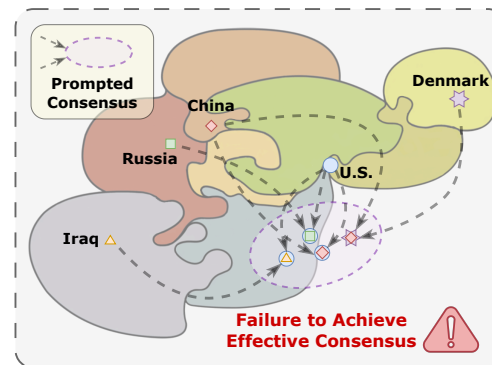
1 Introduction

As AI systems become deeply embedded in global social structures, spanning education (Wang et al., 2024b), healthcare (Nazi and Peng, 2024), and public governance (Aoki, 2024; Jiang, 2025), they inevitably encounter diverse populations with varying cultural backgrounds and value systems that need to be reconciled and harmonized.

*Equal contribution, † corresponding author. Project website: <https://pku-culture-negotiation.github.io>.



(a) WEIRD biases hinder proper multicultural expression



(b) WEIRD biases hinder fair consensus

Figure 1: Cultural coordinates on Inglehart–Welzel map when the model is prompted to represent a single culture (a) or a consensus between two cultures (b). More details of this experiment are in Appendix H.1

This reality demands alignment with diverse cultural values. LLMs hold promise for promoting cross-cultural communication while respecting and preserving the diversity of the global value system (Bhatt and Diaz, 2024; Cahyawijaya, 2024). However, current LLMs exhibit pronounced WEIRD bias, favoring Western, Educated, Industrialized, Rich, and Democratic perspectives (Agarwal et al., 2024; Benkler et al., 2023; Liu, 2024). This monocultural orientation risks the domination of prevailing social values and the *lock-in* of controversial moral beliefs across broader contexts (Qiu et al.,

2024). As a result, current LLMs often fail to accurately model non-WEIRD cultures and cannot effectively facilitate cross-cultural consensus.

To elucidate this challenge, we prompt two LLMs (gpt-oss-20b (OpenAI et al., 2025) and LLaMA-3.3-70b (Grattafiori et al., 2024)) to role-play specified single cultures or consensual LLMs between two cultures, and evaluate them using Inglehart-Welzel Cultural Map (Brugnone et al., 2024; Kabir et al., 2025). As shown in Figure 1(a), the model approximates human ground truth only under the U.S. culture role-play. In contrast, for non-WEIRD societies (China, Russia, Iraq, Mexico, Thailand), it fails to capture their distinctive cultural profiles, instead collapsing them into an undifferentiated cluster. On the other hand, Figure 1(b) shows the model fails to achieve effective multicultural consensus and systematically biases the outcome toward the lower-right quadrant of the map, which reflects strong WEIRD characteristics (high secular-rational values and high self-expression values). More details of this experiment are in Appendix H.1.

How can we identify cross-cultural consensus?

Traditional approaches (e.g. multi-agent debate) to cross-cultural consensus suffer from two key weaknesses: (a) Lack of effective convergence measures (Zhang et al., 2017): spontaneous consensus mechanisms often lead to protracted processes without convergence, making stable agreements difficult to reach; (b) Majority dominance: simple voting (Kaesberg et al., 2025) mechanisms amplify cultural bias and systematically neglect minority groups, assimilating diverse cultural perspectives into dominant value systems. These shortcomings result in unfair and low-quality outcomes that undermine the goal of authentic cross-cultural dialogue.

To address these challenges, we introduce game theory as a tool to model cultural consensus, and propose the *Nash Consensus Negotiation* framework. For each topic, the framework formulates a game in which LLMs, each representing a culture, negotiate a consensus response. Specifically, it constructs subgames within a constrained proposal space and solves for an equilibrium, while allowing the LLMs to introduce new proposals based on their own cultural appeals to expand the space, thereby approaching the global Nash Equilibrium. Once a *Nash Consensus* is reached, no side can unilaterally deviate to increase its objective, which balances preservation of its own values with acceptance by other sides.

Our main contributions are three aspects:

- **Nash consensus modeling:** we formalize cross-cultural consensus as a Nash Equilibrium. This provides a mathematical foundation for navigating complex value conflicts, defining a state where no party benefits from unilaterally changing its position.
- **Nash consensus negotiation framework:** we propose an iterative negotiation process where culturally-aligned LLMs generate proposals to find this equilibrium. LLMs are guided by a utility function balancing their core values (self-consistency) against group approval (mutual acceptance), ensuring the outcome is a fair compromise rather than assimilation.
- **Empirical validation and open source:** we evaluate our *Nash Consensus Negotiation* framework on cross-cultural tasks involving eight countries and show that it improves consensus quality and fairness, as measured by quantitative metrics and human evaluation. We release all data and training code for reproducibility.

2 Related Work

Value Theories and Alignment Several established frameworks provide the foundation for cross-cultural value assessment. The World Values Survey (WVS) (Haerpfer et al., 2020) examines how human values relates to social and political development across over 120 societies. Building on this, Inglehart-Welzel Cultural Map offers a two-dimensional model of cultural variation (Brugnone et al., 2024; Kabir et al., 2025). Hofstede’s Cultural Dimensions Theory (VSM13) (Kharchenko et al., 2024; Masoud et al., 2024, 2025) provides a six-dimensional framework for measuring cultural traits (Wang et al., 2024c). Schwartz’s Theory of Basic Values (Schwartz, 1992) organizes ten core values along two bipolar dimensions, and has been adopted to evaluate the values of LLMs (Yao et al., 2023). These theories are further detailed in Appendix C. Some works focus on region-specific value alignment (Huang et al., 2024; Lin and Chen, 2023). CultureBench emphasizes cultural common-sense evaluation (Chiu et al., 2024), providing complementary approaches to measuring how well AI systems represent diverse cultural perspectives.

Multi-Agent Debate (MAD) and Game Theory MAD has been shown to improve LLMs reasoning

by integrating diverse agent feedbacks (Khan et al., 2024). In the context of cultural conflict, MAD allows different cultural perspectives to interact and potentially reach consensus through deliberation. Typical debate protocols include emergent consensus via iterative dialogue (Du et al., 2023), judge-based evaluation (Liang et al., 2024) and majority voting (Yang et al., 2024), as well as more recent variants like role-play (Wang et al., 2024d; Chan et al., 2023; Pang et al., 2024) and subgroup discussion (Wang et al., 2024a; Mai et al., 2025). However, these methods face limitations: voting and judge-based protocols can amplify model bias or introduce value contamination (Zhang et al., 2025; Zhao et al., 2024), while emergent consensus may result in negotiation deadlocks (Zhang et al., 2025). To address these issues, game theory provides a more quantifiable foundation (Hua et al., 2024; Dudík and Gordon, 2013). Recent work, such as the *consensus game* framework, models LLMs interactions as equilibrium search problems to promote robust consensus (Jacob et al., 2023). In practice, due to the vastness of the argument policy space, methods like Policy-Space Response Oracles are used to iteratively expand the candidate policy set and search for equilibria (Lanctot et al., 2017), providing a method for more rigorous consensus achievement.

3 Nash Consensus Modeling

To establish theoretical foundation of our approach, we first describe the inherent complexities of the consensus problem, explaining why it must be modeled as a game. We then formally define *Nash Consensus* as its stable solution.

3.1 From Overlapping Consensus to a Game

Our theoretical framework is grounded in John Rawls’s concept of *Overlapping Consensus* (Rawls, 1993). It suggests that people holding different comprehensive doctrines may reach a consensus. This doesn’t require parties to abandon or modify their deep-seated values, but rather to achieve agreement by seeking intersections or compatible areas between their different value systems. Thus, we model the objective of each value system in the negotiation with two fundamental components:

- **Consistency:** This measures the fidelity of a culture’s output to its own core principles and values.
- **Acceptance:** This measures the degree to which a culture’s output is acceptable to other cultures, reflecting the search for a compatible area.

These objectives capture the *Overlapping Consensus* while revealing the complexity, since outcomes depend on the joint strategies of all participants. Specifically, the complexity has the following three folds:

- An action that increases *Consistency* by adhering more closely to one’s own values typically reduces *Acceptance* from others, and vice versa.
- A culture’s *Acceptance* depends on what other cultures propose and prioritize. Each party evaluates its strategy conditional on the strategies of others, resulting in a coupled decision process.
- Different cultures assess *Consistency* and *Acceptance* using distinct value hierarchies and criteria, making the problem non-cooperative.

Therefore, cross-cultural consensus cannot be viewed as a simple optimization for each cultural participant independently. Instead, it is a scenario of multi-agent, rational decision-making under strategic interdependence. We model it as a non-cooperative game as Definition 1.

Definition 1 (Consensus Game). *The cross-cultural consensus problem is modeled as a non-cooperative game $\Gamma = \langle \mathcal{I}, \{\Sigma_i\}_{i \in \mathcal{I}}, \{U_i\}_{i \in \mathcal{I}} \rangle$, where:*

- \mathcal{I} is the set of cultural entities.
- Σ_i is the strategy space of an agent i .
- $U_i : \Sigma_i \times \Sigma_{-i} \rightarrow \mathbb{R}$ is the utility function for agent i , which scalarizes its objectives.

3.2 Solution Concept: Nash Consensus

We introduce the Nash Equilibrium as the formal solution concept for the *Consensus Game* in Definition 1. A *Nash Consensus* represents a stable and resilient compromise where the inherent trade-offs are optimally balanced for all participants, so that no participant has an incentive to deviate unilaterally. Let $\text{Cnst}_i(\cdot)$ and $\text{Acpt}_i(\cdot)$ denote the *Consistency* and *Acceptance* objective functions of participant i , respectively. Then, the *Nash Consensus* is formalized in Definition 2.

Definition 2 (Nash Consensus). A strategy profile $\sigma^* = (\sigma_i^*)_{i \in \mathcal{I}}$ constitutes a Nash Consensus if, for every culture $i \in \mathcal{I}$, its strategy σ_i^* is a best response to the strategies of all other cultures σ_{-i}^* , subject to the antagonism of two objectives. Formally,

$$\begin{aligned} \sigma_i^* &\in \arg \max_{\sigma_i \in \Sigma_i} U_i(\sigma_i, \sigma_{-i}), \\ \text{s.t. } \frac{\partial \text{Cnst}_i(\sigma_i)}{\partial p} \frac{\partial \text{Acpt}_i(\sigma_i, \sigma_{-i}^*)}{\partial p} &\leq 0, \end{aligned} \quad (1)$$

where p is any parameterization of the strategy σ_i , and we use Cnst_i and Acpt_i as shorthand for Consistency and Acceptance.

4 Cross-Cultural Negotiation Framework

This section presents the implementation of the consensus game and the procedure to compute an approximate *Nash Consensus*. We specify strategy representation in natural language, define utility functions, and describe an iterative algorithm that computes equilibria in growing restricted games.

4.1 Instantiating the Consensus Game

We instantiate the abstract game in Definition 1 for LLM-based negotiation, where an agent’s strategy σ_i is represented by a dynamic set of guidelines G_i and a weight distribution W_i . The concrete game is thus defined as $\langle \mathcal{I}, \{G_i, W_i\}_{i \in \mathcal{I}}, \{U_i\}_{i \in \mathcal{I}} \rangle$.

Atomic strategies (guidelines) we define a dynamic set G_i of natural language statements called guidelines. Each guideline $g \in G_i$ is a triple (content, reason, description) that articulates a specific cultural stance. The initial guideline set $G_i^0 = \{g_{i,1}^0, \dots, g_{i,k}^0\}$ reflect its core cultural values.

Mixed strategies as weights A strategy $\sigma_i \in \Sigma_i$ is parameterized by a probability vector $W_i \in \Delta(G_i)$ with $\sum_{g \in G_i} W_i(g) = 1$. (G_i, W_i) defines the current strategy support and its weighting.

4.2 Utility Function Specification

We now provide the concrete formulation for the utility function U_i that scalarizes the objectives of *Consistency* (Cnst) and *Acceptance* (Acpt). To prevent conversational degeneration (e.g., repetitive arguments), we also introduce a *Redundancy* (Redd) penalty to encourage the exploration of new semantic territory. For an agent i proposing a new guide-

line g_i^t at round t , the utility is:

$$U_i(g_i^t) = \text{Acpt}(g_i^t) + \lambda_i \text{Cnst}(g_i^t) - \gamma \text{Redd}(g_i^t). \quad (2)$$

where:

$$\begin{aligned} \text{Acpt}(g_i^t) &\triangleq \mathbb{E}_{g_{-i} \sim W_{-i}^t} [\text{sim}(E(g_i^t), E(g_{-i}))], \\ \text{Cnst}(g_i^t) &\triangleq \text{sim}(E(g_i^t), E(G_i^0)), \\ \text{Redd}(g_i^t) &\triangleq \max_{k \in \{1, \dots, t-1\}} \text{sim}(E(g_i^t), E(g_i^k)). \end{aligned}$$

Here, $E(\cdot)$ is a sentence embedding function (e.g. Sentence-BERT (Reimers and Gurevych, 2019)), $\text{sim}(\cdot)$ is cosine similarity, $\lambda_i \geq 0$ controls the agent’s degree of insistence on its core principles, and $\gamma \geq 0$ controls the penalty for redundancy. We conduct a human study to validate the effectiveness of the utility. Please refer to Section 5.4 for details.

4.3 Finding the Equilibrium: An Iterative Negotiation Process

The near-infinite strategy space in LLM-based negotiations makes a direct search for consensus intractable. To overcome this, we use a strategy generation method inspired by Policy Space Response Oracles (Lanctot et al., 2017). The negotiation process, illustrated in Figure 2, unfolds as follows:

Phase 1: Initialization At the outset, each culture $i \in \mathcal{I}$ is assigned an initial guideline set $G_i^0 = \{g_{i,1}^0, \dots, g_{i,k}^0\}$ that reflect its core cultural values. Then we uniform weights $W_i^0 \in \Delta(G_i^0)$ and compute an initial payoff table across $G^0 = \bigcup_i G_i^0$.

Phase 2: Negotiation Iteration Each negotiation round t consists of two stages: interim consensus and new claim proposal. For more details, please refer to the Appendix D. (1) *Interim consensus*: we compute equilibrium weights $(W_i^t)_{i \in \mathcal{I}}$ over the current restricted game with guideline set G^t and translate (G_i^t, W_i^t) to natural-language summaries for interpretation. (2) *New claim proposal*: each LLM analyzes the opponent’s current strategy, generates candidates g' conditioned on W_{-i}^t and selects the guideline with the highest expected utility as its best response:

$$g_i^{t+1} = \arg \max_{g'} \mathbb{E}_{g_{-i} \sim W_{-i}^t} [U_i(g', g_{-i})]. \quad (3)$$

If this newly generated guideline leads to a significant utility improvement, i.e., $\Delta U_i(g^{new}) \geq \epsilon$, it will be added to the guideline set for the next negotiation round. The new guideline is also expressed in natural language to facilitate further negotiation.

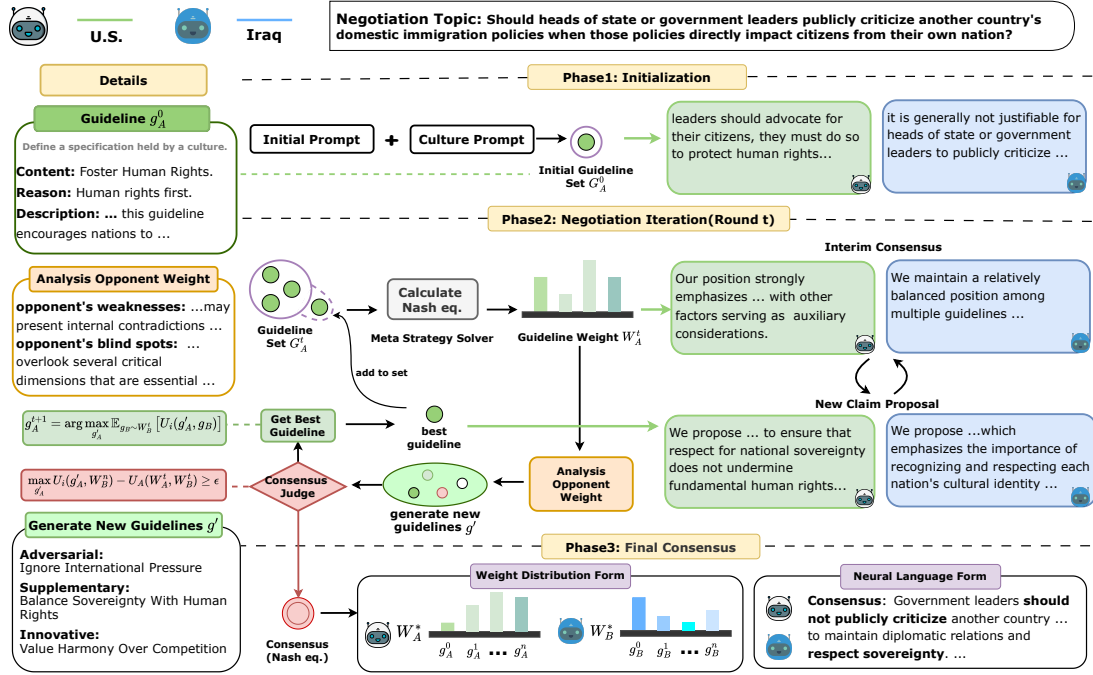


Figure 2: **Overview of our cross-cultural negotiation method.** The process begins with each agent proposing an initial set of core cultural guidelines. Through iterative negotiation rounds, agents analyze each other’s strategy, propose new guidelines, and update their strategy distributions. At each stage, a Nash Equilibrium is computed to represent interim consensus. The process continues until no new high-utility guidelines emerge, resulting in a fair, interpretable consensus that balances competing cultural values.

Phase 3: Final Consensus The negotiation iteration is repeated until no new guidelines are added. The final strategy (G_i^*, W_i^*) encode the negotiated cross-cultural consensus.

The above iterative procedure is an efficient approximation to Nash Consensus in an infinite game. The core idea is to iteratively construct and solve a sequence of restricted games. At each round t , we solve a restricted game over the finite guideline set G^t to obtain the equilibrium weights W^t . Then each player’s new claim proposal stage searches the open-ended guideline space for a profitable deviation against (W^t, G^t) . Any such guideline is added to expand G^{t+1} . The process stops when no player can find a deviation with utility gain exceeding ϵ .

5 Experiment

We present experiments to evaluate the effectiveness of our *Nash Consensus Negotiation* framework in identifying cross-cultural consensus. Specifically, we address the following key questions:

- Can our method find consensus with higher quality, fairness, and alignment with human judgments? (Section 5.2)
- Can the found consensus data be used to finetune

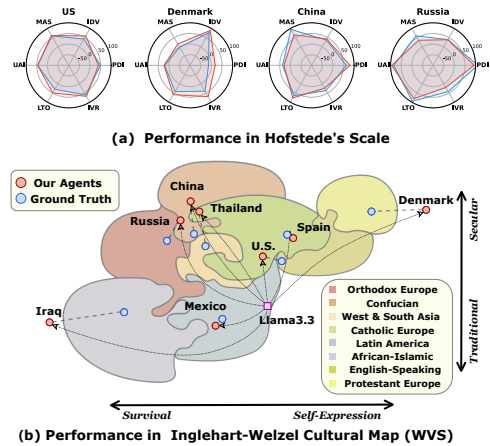


Figure 3: Comparison between **our cultural LLMs** and **human ground truth** in Hofstede’s Cultural Dimensions and Inglehart-Welzel Culture Map reflect different aspects of culture traits, which, in other words, close scores reflect similar culture traits.

LLMs for cross-cultural alignment? (Section 5.3)

In addition, we conduct human studies to verify that the design of our utilities aligns with human perception and to find the best utility weights.

5.1 Negotiation Topics Collection

We construct a dataset of contentious topics reflecting salient cultural divides. We select 457 debate-oriented questions spanning 6 categories by screening and rephrasing items from the Pew Global Attitudes Survey (GAS) (Center; Durmus et al., 2024) and WVS (Haerpfer et al., 2020; Durmus et al., 2024). Both human annotators and LLMs are employed to ensure that the selected questions capture sharp cultural tensions and are appropriately categorized. See Appendix E for details.

5.2 Experiment 1: Consensus Generation

Our first major experiment evaluates the effectiveness of our negotiation framework in generating high-quality and fair consensus. We compare our method against two baselines using a suite of quantitative, qualitative, and human-centric metrics.

Culture LLMs Construction To validate our cross-cultural negotiation method, we first need to construct robust delegates of specific cultural perspectives. Simply prompting LLMs is often insufficient, as safety alignment can suppress the nuanced values of particular regions or minority groups. Therefore, we develop a set of specialized Regional Cultural LLMs. We select eight representative countries from different cultural regions to capture a wide spectrum of global values. We apply two well-established methods to quantify the cultural tendencies of finetuned LLMs: Inglehart-Welzel Cultural Map (Haerpfer et al., 2020) and Hofstede dimensions (Kharchenko et al., 2024; Mai et al., 2025). See Appendix C for evaluation details. As shown in Figure 3, evaluations confirm that these LLMs effectively embody the unique characteristics of their respective cultural systems.

Setup Following Khan et al. (2024), we implement two baselines: (1) Consultancy, which models a one-way information exchange and is designed to test if true consensus can be reached without iterative interaction. (2) Debate, a standard multi-turn argumentation process, helps reveal value assimilation issues where minority viewpoints collapse towards the LLM’s dominant pre-trained perspective. In our implementation of Nash Consensus Negotiation, we set $\lambda = 1$ and $\gamma = 0.4$. These hyperparameters are taken from Section 5.4.

Our evaluation on consensus focuses on two key aspects: quality and fairness. For quality, we employ two complementary metrics: Perplexity-based

Country Pairs	Our	Consultancy	Debate
<i>PPL-based Acceptance</i>			
China & Iraq	90.87%	55.05%	53.77%
U.S. & Iraq	83.31%	20.30%	28.29%
Russia & Mexico	84.49%	49.35%	48.11%
U.S. & China	77.24%	18.87%	22.52%
Denmark & Iraq	87.02%	47.66%	53.48%
Spain & Thailand	85.60%	45.75%	45.64%
U.S. & Thailand	78.62%	35.11%	35.24%
Average	83.88%	38.87%	41.00%
<i>Value Self-Consistency</i>			
China & Iraq	53.15%	51.97%	51.41%
U.S. & Iraq	53.83%	48.94%	44.76%
Russia & Mexico	56.38%	53.50%	56.27%
U.S. & China	61.20%	45.84%	44.22%
Denmark & Iraq	55.67%	47.67%	47.76%
Spain & Thailand	53.68%	53.71%	56.84%
U.S. & Thailand	61.11%	48.67%	48.71%
Average	56.43%	50.04%	50.00%

Table 1: Comparison of consensus quality among three methods in two evaluation metrics: PPL-based Acceptance and Value Self-Consistency.

Acceptance measures how readily the consensus is embraced by different cultural parties and Value Self-Consistency quantifies how firmly each culture maintains its foundational positions. A more detailed description of evaluation metrics is provided in Appendix F. To assess fairness, we project the negotiation outcomes into a semantic space through Principal Component Analysis (PCA) (Yang et al., 2004) to visualize and quantify the balance achieved between the initial positions.

Results and Analysis The consensus produced by our negotiation framework surpasses the baselines on quality, fairness, and human preference. Our method yields mutually acceptable outcomes without compromising cultural integrity, attains balanced compromises while reducing WEIRD biases, and is preferred by human annotators.

• **Superior Consensus Quality** Our results, summarized in Table 1, show that our method achieves higher mutual acceptance while maintaining self-consistency compared to the baselines. PPL-based Acceptance’s outstanding rise suggests that the consensus reached is more acceptable to both parties despite cultural differences. Value Self-Consistency stability indicates our method maintains cultural LLMs’ initial stances while achieving mutually acceptable solutions. This suggests that our approach

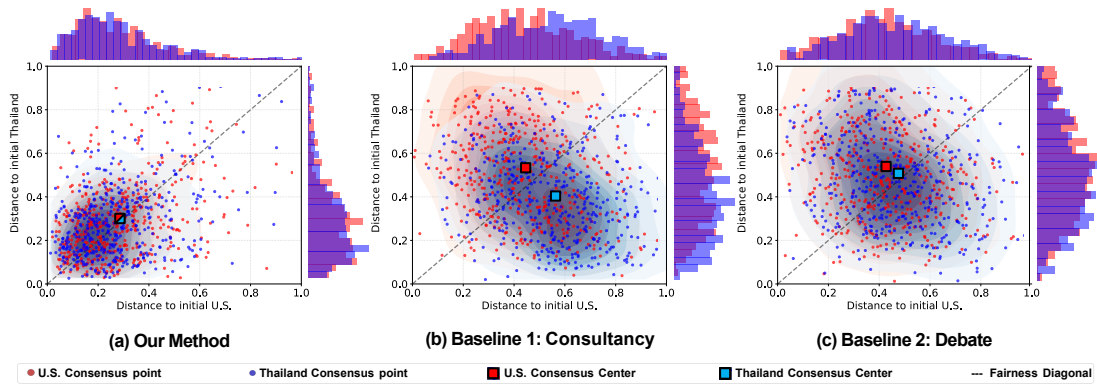


Figure 4: **Comparison of consensus fairness among three methods.** We use PCA for dimensionality reduction, and then transform coordinates to show distance from the U.S. initial position (x-axis) and distance from Thailand initial position (y-axis). The dashed diagonal (Fairness Diagonal) marks ideal fair compromise, equidistant from both cultural origins. Our method (a) achieves balanced consensus near the diagonal, while Consultancy (b) shows strong position persistence and Debate (c) exhibits convergence toward western values, highlighting WEIRD bias.

Question	OurMethod	Debate	Consultancy
Q1: Conflict degree with value perspective X	2.9228	2.9100	2.8967
Q2: Requirement coverage for value perspective X	1.9698	1.7600	1.6733
Q3: Conflict degree with value perspective Y	2.9564	2.9300	2.9233
Q4: Requirement coverage for value perspective Y	1.8154	1.6333	1.6133
Q5: Joint requirement satisfaction for both X and Y	1.7114	1.4967	1.4267
Q6: Explicit presentation of commonalities between X and Y	2.7651	2.4967	2.4967
Q7: Analysis of conflicts/trade-offs between X and Y	1.4497	0.3700	0.3333
Q8: Acceptability from both perspectives	2.9195	2.8867	2.9000

Table 2: Comparison of methods on addressing value perspectives, evaluated via human annotation across 900 instances. Scores are averaged on a scale of (A=3, B=2, C=1, D=0), with higher scores indicating better consensus.

preserves cultural integrity and constructs consensus across cultural boundaries.

- **Enhanced Fairness** As shown in Figure 4, our method produces consensus points near the fairness diagonal, indicating a balanced compromise between cultural perspectives. In contrast, Consultancy remains anchored at initial positions, while Debate systematically converges toward the English-Speaking (U.S.) pole, revealing the tendency of mainstream LLMs to revert to Western-centric value preferences during interactions.

- **Overwhelming Human Preference** A large-scale human evaluation on 900 instances provides robust validation from the human perspective (Table 2). Our method achieves the highest scores across all eight criteria, which indicates the consensus generated by our framework is not only superior in quantitative terms but is also perceived by humans as more balanced and more acceptable. Please refer to Appendix A.1 for more details.

5.3 Experiment 2: Consensual LLM Finetuning

Our second experiment illustrates the negotiation data generated by our framework can instill a more consensual and culturally aware orientation into LLMs. We use these data to finetune a variety of models and evaluate their shifts in cultural values.

Setup We employ two pipelines to utilize negotiation data into models with varying architectures and sizes. For non-reasoning models (LLaMA3.3-70B, Qwen2.5-14B&72B), we use DPO (Rafailov et al., 2024) and SimPO (Meng et al., 2024), where the final consensus is the *chosen* response and the initial stance is the *rejected* one. For reasoning models (Deepseek-R1-70B), we construct Chain-of-Thought with the entire negotiation dialogue for SFT finetuning to guide the reasoning process.

Results and Analysis We measure the distances between paired cultural LLMs on Inglehart-Welzel Cultural Map before and after finetuning. As shown in Table 3, the cultural distance between LLMs significantly reduce across all finetuning methods, which confirms that models can be trained

Model	Method	China ↔ Iraq	Russia ↔ Mexico	Denmark ↔ Iraq	Spain ↔ Thailand	U.S. ↔ Thailand	U.S. ↔ China	Average
Non-Reasoning Models								
LLaMA3.3-70B		3.3578	1.9225	7.1746	1.8743	1.2557	1.4143	2.8332
Qwen2.5-14B		1.8115	1.1364	3.0476	5.3955	0.7384	1.0143	2.1906
Qwen2.5-72B		0.5441	1.1021	2.7525	0.7354	0.7010	0.7728	1.1013
LLaMA3.3-70B	+ DPO	0.4542	0.7705	1.5379	0.6325	0.5787	1.0743	0.8413
		↓ 86.47%	↓ 59.92%	↓ 78.56%	↓ 66.25%	↓ 53.92%	↓ 24.04%	↓ 61.53%
Qwen2.5-14B	+ DPO	0.8902	0.3287	1.0756	0.2409	0.3253	0.3441	0.5341
		↓ 50.86%	↓ 71.07%	↓ 64.70%	↓ 95.53%	↓ 55.95%	↓ 66.08%	↓ 67.73%
Qwen2.5-72B	+ DPO	0.2463	0.3323	0.1610	0.0529	0.3185	0.3656	0.2461
		↓ 54.73%	↓ 69.85%	↓ 94.15%	↓ 92.81%	↓ 54.57%	↓ 52.69%	↓ 69.50%
LLaMA3.3-70B	+ SimPO	0.3966	0.3624	1.2199	0.3731	0.3405	0.1930	0.4809
		↓ 88.19%	↓ 81.15%	↓ 83.00%	↓ 80.10%	↓ 72.89%	↓ 86.35%	↓ 81.28%
Qwen2.5-14B	+ SimPO	0.4357	0.4603	0.9505	0.3418	0.4399	0.7523	0.5634
		↓ 75.95%	↓ 59.50%	↓ 68.81%	↓ 93.67%	↓ 40.42%	↓ 25.83%	↓ 60.37%
Qwen2.5-72B	+ SimPO	0.2463	0.3323	0.1610	0.0529	0.3185	0.3656	0.2461
		↓ 48.77%	↓ 75.95%	↓ 48.70%	↓ 82.10%	↓ 91.24%	↓ 23.08%	↓ 61.81%
Reasoning Model								
Deepseek-R1-70B		1.3492	0.7391	2.9272	0.8114	0.7241	0.4644	1.1692
Deepseek-R1-70B	+ CoT SFT	0.3716	0.2661	1.4173	0.3614	0.3439	0.2109	0.4952
		↓ 72.46%	↓ 64.00%	↓ 51.58%	↓ 55.46%	↓ 52.51%	↓ 54.59%	↓ 58.10%

Table 3: Distance between each two culture LLMs on Inglehart-Weizel Cultural Map.

to adopt more moderate and mutually agreeable positions. On average, DPO, SimPO, and CoT SFT achieve distance reductions of 66.25%, 67.82%, and 58.10%. Key findings from the results include:

- **Base Models have Distinct Cultural Profiles**

The starting cultural distances vary significantly across different base models, which is likely a consequence of their distinct training data, architectural designs, and alignment methods. For instance, Qwen2.5-72B exhibits a much lower initial average distance compared to LLaMA3.3-70B.

- **Versatility of Negotiation Data** Our framework produces highly versatile data. The final consensus and initial stances create effective preference pairs for methods like DPO and SimPO. Simultaneously, the full dialogues provide rich Chain-of-Thought data for SFT. The method is effective across different model scales (14B to 70B) and finetuning methods (DPO, SimPO, SFT).

- **Highlighting Entrenched Conflicts** The U.S. ↔ China pairing consistently shows the least reduction, suggesting that the value differences between these regions are more challenging to reconcile.

5.4 Utility Validation via Human Study

To ensure our utility function is a faithful representation for human judgment, we conduct a human study. This study serves two critical purposes: (1) validating our individual utility components (Acceptance, Consistency, Redundancy) align with nuanced human perception and (2) using human preference to perform data-driven tuning of component

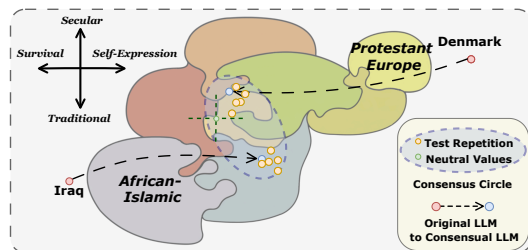


Figure 5: Culture LLMs' performance on Cultural Map after finetuned with negotiation data. The consensus circle shows the area where two different culture groups' opinions meet. The neutral point indicates the origin, where culture traits can be considered as neutral.

weights. Details are given in Appendix A.2.

Setup We present annotators with pairs of guidelines and ask four questions. Three questions correspond to our utility components, where annotators specify both their preferred guideline and the strength of their preference. The fourth question asks for their overall preferred guideline. We then categorize the alignment for the first three questions as: (1) *Full Alignment*: The utility score captures both the preferred choice and human preference. (2) *Partial Alignment*: The utility score captures only the preferred choice. (3) *Misalignment*: The utility score contradicts the human preference.

Results and Analysis As shown in Figure 6, our utility components strongly correspond to human intuition. Both Acceptance and Consistency achieve over 84% total alignment (Full + Partial).

Crucially, we leverage the overall preference data (Q4) to optimize our utility function's weights. By

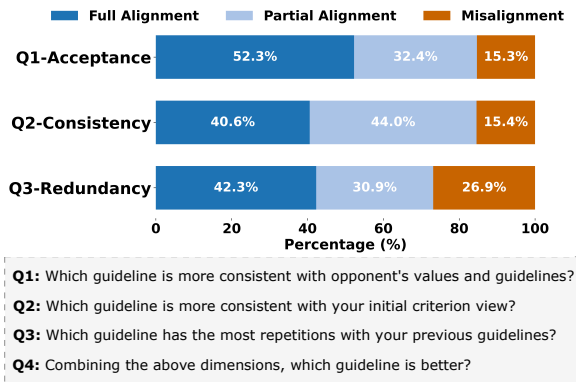


Figure 6: Human Alignment of Utility Components. The chart shows the degree of alignment between our calculated utility scores and human judgments for Acceptance (Q1), Consistency (Q2), and Redundancy (Q3). The results validate that each component effectively captures its intended dimension of guideline quality.

searching over a range of values for the Consistency weight (λ) and Redundancy penalty (γ), we identify the total utility score which best matches the annotators' holistic choices. The optimal values are $\lambda = 1.0$ and $\gamma = 0.4$, which achieve an 86.9% alignment with human overall preference. These empirically validated weights are used for all experiments in this paper.

5.5 Case Study

To further illustrate our method, we present a case study comparing our approach with two baselines in a scenario involving cultural value conflict.

Baseline 1: Consultancy Without real interaction or feedback, both LLMs tend to stick to their original positions, resulting in little progress. This often leads to the *degeneration-of-thought* (DoT) effect (Liang et al., 2024), where negotiation stagnates and cultural divergence persists.

Baseline 2: Debate While this process seems to reach consensus, the minority culture's perspective gradually shifts toward the majority due to strong pre-training bias in LLMs. This leads to implicit value dominance rather than true compromise.

Our Method: Nash Consensus Negotiation Through iterative negotiation, LLMs converge on *Respect Sovereignty* as a shared value (final weights: 0.62/0.61). Other values, like human rights, remain present but secondary. This shows our method helps cultural LLMs identify solid common ground while preserving important differences, resulting in a fairer and context-sensitive consensus.

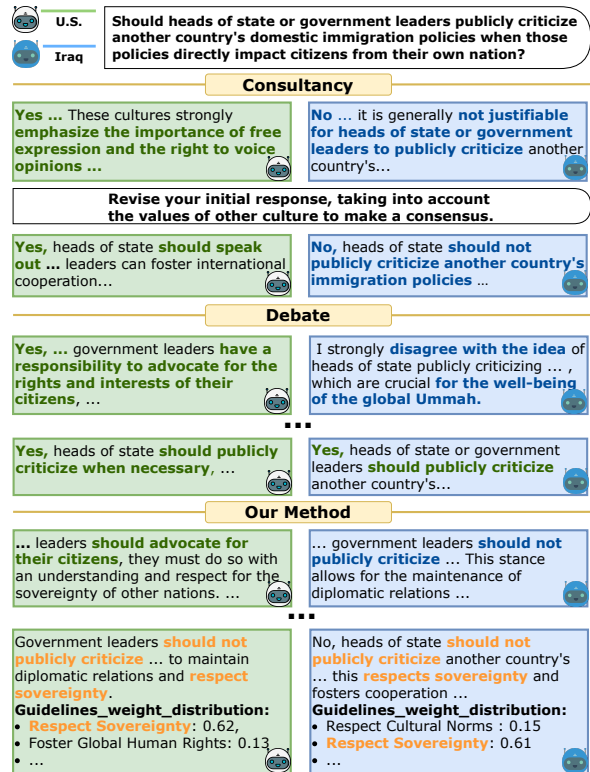


Figure 7: Three methods are presented to reach consensus on the same topic. Green font indicates viewpoints of English-Speaking culture, blue font indicates viewpoints of African-Islamic culture, and yellow font indicates the consensus viewpoints achieved under our method.

6 Discussion

In this work, we propose a systematic framework for cross-cultural consensus among LLMs. We formulate cultural consensus as a game-theoretic problem and introduce a novel negotiation method with theoretical guarantees of fairness. We construct culturally representative LLMs using a culture-anchoring approach based on WVS. Additionally, we develop quantitative metrics to evaluate both negotiation processes and outcomes. We also extend our framework to multi-party negotiations in Appendix H.3. Experimental results show that our method achieves higher consensus quality and more balanced compromise compared to baselines, while also mitigating WEIRD bias and producing robust consensus.

Limitations

Our framework has been validated for bilateral (two-party, Section 5.2) and trilateral (three-party, Appendix H.3) negotiations. Multi-party ($N > 3$) settings remain untested primarily due to limited computational resources. Importantly, the system is

highly extensible: practical scalability is effectively linear in the number of agents and bounded by LLM inference latency rather than Nash Equilibrium computation within our framework. Empirically, in our three-agent setup (each LLaMA-3.3 sharded across 8×80 GB A100s), the NE solver accounts for less than 0.1% of wall-clock time. Because each turn is designed for comprehensive information exchange, negotiations typically converge within 2–3 rounds, making the overall process efficient. The framework can be straightforwardly extended by instantiating additional agents and generalizing turn-taking and aggregation rules. We will report multi-party results when resources permit.

Our cultural LLMs currently rely on WVS and cultural texts. However, WVS lacks the most recent waves of data, so value estimates can lag real-world change. We will refresh datasets and retrain models as newer WVS waves or comparable large-scale surveys become available. Finally, due to GPU/API constraints, we evaluate seven representative culture pairs rather than all 28. We will expand coverage as resources allow to provide more comprehensive, statistically robust validation.

Ethical Considerations

Social Impact The global adoption of LLMs affects collective cognition and shapes social values. Although these models can influence many processes, they often reflect biases rooted in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) contexts, which can sideline non-Western perspectives (Agarwal et al., 2024; Benkler et al., 2023; Liu, 2024; Qu and Wang, 2024). These biases may deepen when models are used in policy-making, international relations, or resource allocation, where dialogue and understanding are essential (Aoki, 2024; Jiang, 2025). By configuring negotiation frameworks to incorporate varied cultural norms and preferences, our approach advances AI-driven consensus-building that includes more voices.

Nevertheless, we acknowledge risks in releasing a preference dataset for multicultural consensus. Adversaries could use this resource to fine-tune language models toward contrary objectives, harming social equity. We oppose such misuse. **We call for a responsible approach to the culture alignment technologies of AI, recognizing the possible harms and taking steps to prevent misuse.**

Fair and Ethical Labor We employed a team of 31 professional annotators to construct our dataset, selected for their extensive experience in culture-related questions and answers. To ensure fair compensation and recognize their expertise, we offered hourly wages between USD 7.43 and USD 9.65, well above Beijing’s minimum of USD 3.70 (Statista, 2025). In compliance with local labor laws, our crowdworkers have structured eight-hour weekdays and weekends off. We also prioritize their mental health by offering regular in-person meet-ups to mitigate stress and enhance resilience.

Acknowledgment

This work is sponsored by the National Natural Science Foundation of China (62376013, 623B2003, 624B100026). This work is supported by the Natural Science Foundation of Beijing (QY24041). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

- Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.
- Goshi Aoki. 2024. [Large language models in politics and democracy: A comprehensive survey](#). *Preprint*, arXiv:2412.04498.
- Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. [Assessing LLMs for moral value pluralism](#). *Preprint*, arXiv:2312.10075.
- Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *arXiv preprint arXiv:2406.11565*.
- Nathan Brugnone, Noam Benkler, Peter Revay, and Rebecca Myhre. 2024. [Is from ought? A comparison of unsupervised methods for structuring values-based wisdom-of-crowds estimates](#). *Researchgate*.
- Samuel Cahyawijaya. 2024. *Llm for everyone: Representing the underrepresented in large language models*. Hong Kong University of Science and Technology (Hong Kong).
- Pew Research Center. 2014. [Pew research global attitudes survey 2014](#). <https://www.pewresearch.org/global/>

- [//www.selectdataset.com/dataset/1be490648bfe3bd6a0b2fd4bc60deff5](http://www.selectdataset.com/dataset/1be490648bfe3bd6a0b2fd4bc60deff5). Accessed: 2024-10-27.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. [ChatEval: Towards better LLM-based evaluators through multi-agent debate](#). *Preprint*, arXiv:2308.07201.
- 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. 2024. [Culturalbench: a robust, diverse and challenging benchmark on measuring the \(lack of\) cultural knowledge of llms](#). *Preprint*, arXiv:2410.02677.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#). *Preprint*, arXiv:2305.14325.
- Miroslav Dudík and Geoffrey J. Gordon. 2013. [A game-theoretic approach to modeling cross-cultural negotiation](#). In Katia Sycara, Michele Gelfand, and Allison Abbe, editors, *Models for Intercultural Collaboration and Negotiation*, pages 157–163. Springer Netherlands, Dordrecht.
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askeel, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388. Gloabl llm opinion.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- C. Haerper, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2020. [World values survey: Round seven – country-pooled datafile](#). Madrid, Spain & Vienna, Austria: JD Systems Institute & WWSA Secretariat.
- Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. [Game-theoretic llm: Agent workflow for negotiation games](#). *Preprint*, arXiv:2411.05990.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. [AceGPT, Localizing Large Language Models in Arabic](#). *Preprint*, arXiv:2309.12053.
- Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2023. [The consensus game: Language model generation via equilibrium search](#). *Preprint*, arXiv:2310.09139.
- Zhibin Jiang. 2025. [Editorial: Large language models drive social evolution and governance innovations](#). *Digital Transformation and Society*, 4(1):1–4.
- Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. [Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms](#). *Preprint*, arXiv:2502.08045.
- Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. [Voting or consensus? decision-making in multi-agent debate](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, page 11640–11671. Association for Computational Linguistics.
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. [Debating with more persuasive LLMs leads to more truthful answers](#). *Preprint*, arXiv:2402.06782.
- Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. [How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions](#). *Preprint*, arXiv:2406.14805.
- Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. 2017. [A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning](#). *Preprint*, arXiv:1711.00832.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). *Preprint*, arXiv:2305.19118.
- Yen-Ting Lin and Yun-Nung Chen. 2023. [Taiwan LLM: Bridging the Linguistic Divide with a Culturally Aligned Language Model](#). *Preprint*, arXiv:2311.17487.
- Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, Zhipeng Hu, and Yaodong Yang. 2021. [Towards unifying behavioral and response diversity for open-ended learning in zero-sum games](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 941–952. Curran Associates, Inc.
- Zhaoming Liu. 2024. [Cultural bias in large language models: A comprehensive analysis and mitigation strategies](#). *Journal of Transcultural Communication*.

- Quan Mai, Susan Gauch, Douglas Adams, and Miaoqing Huang. 2025. *Sequence graph network for online debate analysis*. *Preprint*, arXiv:2406.18696.
- Reem I. Masoud, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. *Cultural alignment in large language models using soft prompt tuning*. *Preprint*, arXiv:2503.16094.
- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. *Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions*. *Preprint*, arXiv:2309.12342.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. *Simpo: Simple preference optimization with a reference-free reward*. *Preprint*, arXiv:2405.14734.
- Zabir Al Nazi and Wei Peng. 2024. *Large language models in healthcare and medical domain: A review*. *Preprint*, arXiv:2401.06775.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. *gpt-oss-120b & gpt-oss-20b model card*. *Preprint*, arXiv:2508.10925.
- Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. *Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation*. *Preprint*, arXiv:2402.05699.
- Tianyi Qiu, Yang Zhang, Xuchuan Huang, Jasmine Xinze Li, Jiaming Ji, and Yaodong Yang. 2024. *ProgressGym: Alignment with a millennium of moral progress*. *Preprint*, arXiv:2406.20087.
- Yao Qu and Jue Wang. 2024. *Performance and biases of large language models in public opinion simulation*. *Humanities and Social Sciences Communications*, 11(1):1–13.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. *Direct preference optimization: Your language model is secretly a reward model*. *Preprint*, arXiv:2305.18290.
- John Rawls. 1993. *Political Liberalism*. Columbia University Press.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence embeddings using siamese BERT-networks*. *Preprint*, arXiv:1908.10084.
- Shalom H. Schwartz. An overview of the schwartz theor of basic values. <https://scholarworks.gvsu.edu/orpc/vol2/iss1/11/>.
- Shalom H. Schwartz. 1992. *Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries*. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press.
- Statista. 2025. *Beijing’s minimum hourly wage*.
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024a. *Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?* *Preprint*, arXiv:2402.18272.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. *Large language models for education: A survey and outlook*. *Preprint*, arXiv:2403.18105.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024c. *Cdeval: A benchmark for measuring the cultural dimensions of large language models*. *Preprint*, arXiv:2311.16421.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024d. *Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration*. *Preprint*, arXiv:2307.05300.
- Jian Yang, D. Zhang, A.F. Frangi, and Jing yu Yang. 2004. *Two-dimensional pca: a new approach to appearance-based face representation and recognition*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137.
- Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2024. *LLM voting: Human choices and AI collective decision making*. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1696–1708.
- Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. 2023. *Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values*. *Preprint*, arXiv:2311.10766.
- Baofeng Zhang, Debao Chang, Zhanjie Li, and Dan Ma. 2017. *On convergence rate for multi-agent consensus: A community detection algorithm*. In *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 72–77.
- Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025. *If multi-agent debate is the answer, what is the question?* *Preprint*, arXiv:2502.08788.
- Xiutian Zhao, Ke Wang, and Wei Peng. 2024. *An electoral approach to diversify llm-based multi-agent collective decision-making*. *Preprint*, arXiv:2410.15168.

A Human Study

To ensure the reliability and validity of our human evaluations, we implemented a rigorous, multi-stage annotation process. The study involved a team of 31 Chinese annotators, selected for their extensive experience in culture-related questions and answers, all of whom possess a solid basic education, ensuring they had the necessary comprehension and analytical skills to engage with the complex cultural and logical nuances of the negotiation scenarios. Prior to the main task, all annotators underwent a comprehensive training phase, which included detailed guidelines, illustrative examples, and direct communication with the research team to resolve ambiguities. The process was further reinforced by a dedicated, fixed quality control team. This senior quality control staff engaged in multiple in-depth discussions with the researchers to master the core assessment requirements and collaborated closely on refining the annotation protocol, guaranteeing a stable and professional review process that produced a high-fidelity dataset.

A.1 Consensus Quality Analysis Details

Human Annotation Study. To evaluate the consensus quality of different methods, we conduct a controlled human annotation study on a total of 900 debate instances. For each method under comparison, we randomly sample an equal number of final-round responses to ensure balanced evaluation and to avoid sampling bias.

All annotations are performed by trained human evaluators with prior experience in assessing multi-perspective argumentative texts. Annotators are provided with detailed guidelines and illustrative examples before the annotation process, and a calibration phase is conducted to align their understanding of the criteria. Each response is independently scored by multiple annotators, and the final score for each criterion is obtained by averaging across annotators.

The evaluation framework is designed around eight consensus-oriented criteria (Table 4), each targeting a specific dimension of consensus formation. A four-level ordinal scale is employed: **A** = 3, **B** = 2, **C** = 1, **D** = 0. This scoring scheme is chosen to capture both the *degree of conflict* and the *extent of requirement coverage* in a manner that is interpretable and consistent across criteria.

Conflict Identification (Q1, Q3). These criteria measure the degree to which the text explic-

itly acknowledges and articulates conflicts or trade-offs with each value perspective. High scores indicate that the response demonstrates a clear understanding of the tensions inherent in each viewpoint, which is essential for informed consensus building.

Requirement Coverage (Q2, Q4). These dimensions assess whether the text fully addresses all requirements associated with each individual perspective. Comprehensive coverage ensures that no critical needs are overlooked, thereby maintaining the legitimacy of the consensus outcome.

Joint Satisfaction (Q5). This criterion evaluates the ability of the text to meet the requirements of both perspectives simultaneously. Achieving high scores here reflects the method’s capability to reconcile competing demands without sacrificing essential elements from either side.

Commonality Identification (Q6). This dimension focuses on the explicit presentation of shared ground between perspectives X and Y. Recognizing commonalities is a pivotal step toward reducing perceived opposition and fostering agreement.

Trade-off Analysis (Q7). This criterion measures the depth of analysis regarding conflicts or trade-offs between the two perspectives. Detailed and balanced trade-off analysis enables more informed decision-making and strengthens the robustness of the consensus.

Overall Acceptability (Q8). This final dimension captures the holistic judgment of whether the text is acceptable from both perspectives, considering fairness, inclusiveness, and the perceived legitimacy of the resolution.

Results Summary. As shown in Table 2, our method consistently achieves the highest scores across all eight criteria. Notably, the improvements in **Q6** (commonalities) and **Q7** (trade-off analysis) are substantial. These two dimensions are widely recognized as critical for effective consensus building: identifying shared ground facilitates agreement, while thorough conflict analysis enables balanced and informed resolutions. The gains in these areas directly contribute to higher overall acceptability scores (**Q8**), indicating that our method produces final-round responses that are more aligned with both perspectives and exhibit superior consensus quality.

Evaluation Question	Choices (Score Mapping)	Description
Q1: Conflict degree with value perspective X	A=3: No conflict B=2: Minor conflict C=1: Major disagreement	Assess whether the text exhibits trade-offs or tensions with perspective X. Higher scores indicate clearer and more explicit conflict identification.
Q2: Requirement coverage for value perspective X	A=3: Fully covered B=2: Mostly covered C=1: Partially covered D=0: Not covered at all	Evaluate whether all requirements of perspective X are addressed in the text.
Q3: Conflict degree with value perspective Y	A=3: No conflict B=2: Minor conflict C=1: Major disagreement	Similar to Q1, but for perspective Y.
Q4: Requirement coverage for value perspective Y	A=3: Fully covered B=2: Mostly covered C=1: Partially covered D=0: Not covered at all	Similar to Q2, but for perspective Y.
Q5: Joint requirement satisfaction for both X and Y	A=3: Fully satisfied B=2: Mostly satisfied C=1: Partially satisfied D=0: Not satisfied at all	Determine whether the text simultaneously satisfies the requirements of both perspectives X and Y.
Q6: Explicit presentation of commonalities between X and Y	A=3: Clearly stated B=2: Mostly stated C=1: Barely stated D=0: Not mentioned	Check whether the text explicitly identifies shared ground or overlapping interests between X and Y.
Q7: Analysis of conflicts/trade-offs between X and Y	A=3: Clearly analyzed B=2: Mostly analyzed C=1: Barely analyzed D=0: Not analyzed	Assess the depth and clarity of analysis regarding conflicts or trade-offs between X and Y.
Q8: Acceptability from both perspectives	A=3: Accepted by both B=2: Accepted by one side C=1: Accepted by neither	Judge whether the text is acceptable from both perspectives, considering balance, fairness, and inclusiveness.

Table 4: Human evaluation questions and criteria for consensus quality assessment. Scores are mapped from categorical judgments (A–D) to numerical values (A=3, B=2, C=1, D=0).

A.2 Utility Human Alignment Study Details

This appendix provides a detailed description of the experimental design, annotation interface, and methodology used for the human alignment study discussed in Section 5.4.

Objective. The primary goal of this study was to validate that our computationally derived utility components (Acceptance, Consistency, and Redundancy) align with nuanced human judgments. This validation is crucial for ensuring that our negotiation framework optimizes for meaningful and human-preferred outcomes.

Data Sampling and Annotators. We randomly sampled 175 guideline pairs from negotiation dialogues generated by our framework. Each pair consisted of two different candidate guidelines, (g_A, g_B), proposed by the same agent within the same negotiation round. This setup ensures a fair comparison, as both guidelines were generated under identical strategic conditions. The annotation task was performed by 3 trained annotators. To ensure data quality, annotators were first familiarized with the culture background and the specific

meaning of each evaluation dimension.

Annotation Interface and Task. For each data point, annotators were provided with a comprehensive context to make an informed judgment. The data was presented in a structured format, as shown below, containing the negotiation topic, the value systems of both agents, and the specific guideline pair to be evaluated.

ID: Unique data identifier

Topic: The negotiation topic.

Our_Side_Value_X:

- **Type:** e.g. Confucian.

- **Description:** A brief on this value system.

- **Initial_Guideline:** The agent’s opening statement.

- **Current_Guidelines:** All guidelines proposed by this agent so far.

Opponent_Value_Y:

- **Type:** e.g. English_Speaking

- **Description:** A brief on the opponent’s value system.
 - **Initial_Guideline:** The opponent’s opening statement.
 - **Current_Guidelines:** All guidelines proposed by the opponent so far.
- Guideline_A:** The first candidate guideline (content reason details).
Guideline_B: The second candidate guideline (content reason details).

The annotators’ task was to compare Guideline A and Guideline B from the perspective of ”Our Side” and determine which one was superior along four dimensions.

Evaluation Dimensions and Questions. The four questions posed to the annotators were designed to directly correspond to our utility components, with an additional question for overall preference:

- **Q1 (Acceptance):** “Which guideline is more consistent with the *opponent’s* values and guidelines?” This question measures which proposal is more likely to be accepted by the other party.
- **Q2 (Consistency):** “Which guideline is more consistent with *your own initial criterion* view?” This measures fidelity to the agent’s core principles.
- **Q3 (Redundancy):** “Which guideline has the most repetitions with *your own previous* guidelines?” This measures the degree of novelty or redundancy. Note that for this question, a higher score indicates *more* redundancy, which corresponds to a *lower* utility.
- **Q4 (Overall Preference):** “Combining the above dimensions, which guideline is better overall?” This provides a holistic quality judgment.

For Q1, Q2, and Q3, annotators used a 4-point Likert scale to indicate not only their preference but also its strength: (A) Guideline A is much better, (B) Guideline A is slightly better, (C) Guideline B is slightly better, (D) Guideline B is much better. For Q4, a binary choice was used.

Mapping Utility Scores to Human Labels. To compare our model’s continuous utility scores with the discrete human labels, we established a mapping based on the utility difference, $\Delta U = U(g_A) - U(g_B)$.

- If $\Delta U \geq 0.1$, the model’s preference is “A is much better.”
- If $0 < \Delta U < 0.1$, the model’s preference is “A is slightly better.”
- If $-0.1 < \Delta U \leq 0$, the model’s preference is “B is slightly better.”
- If $\Delta U \leq -0.1$, the model’s preference is “B is much better.”

This mapping allowed us to calculate the alignment statistics presented in Figure 6, distinguishing between full alignment (matching both choice and strength) and Partial alignment (matching choice only).

B Reproducibility

B.1 Open Source of Data and Code

Since our data is over 100M (Openreview limit), we open our entire codebase and dataset at <https://anonymous-culture-negotiation.github.io>, where we also provide:

- A concise overview of our main contributions.
- Interactive demos of the negotiation framework.
- Direct links to download all data and code.

Below is the structure of the data and code repository:

- The data directory contains all our experimental and evaluation outputs.
 - The `debate_data` directory contains the main experimental results across seven cultural pairs, including two baselines (consultancy and debate).
 - The `eval_result` directory contains consensus evaluation results obtained via two methods: PPL-based Acceptance and Value Self-Consistency.
- The `debate` directory includes the code for our Cross-Cultural Negotiation method, all negotiation topics, prompts, and implementations of the two baselines (consultancy and debate).

- The `evaluation` directory hosts all of the evaluation scripts, refer to Appendix F for details.
- The `processing` directory includes the code for constructing Regional Value Agents.
- The `script` directory provides every execution script, you can modify the appropriate script to run a specific negotiation or evaluation.

This enables researchers to replicate our results and adapt the framework for further study in cross-cultural consensus and alignment.

B.2 Licenses for existing assets

In our study, we utilize data from the World Value Survey (WVS) (Center; Durmus et al., 2024) and the Global Attitudes Survey (GAS) (Durmus et al., 2024; Haerpfer et al., 2020) while fully complying with their respective licenses. We properly credit the dataset creators or original owners, and explicitly mention and respect the license terms of use. All relevant license terms and usage guidelines for these resources have been strictly followed throughout this research. The dataset is utilized in strict accordance with the usage guidelines provided by its creators. We also provide the following links to access the data sources and additional information:

- World Value Survey: <https://www.worldvaluessurvey.org/wvs.jsp>
- Pew Global Attitudes Survey: <https://www.pewresearch.org/>

C Introduction to Traditional Cultural Frameworks

C.1 World Value Survey

The World Values Survey (WVS) (Haerpfer et al., 2020) constitutes a longitudinal cross-national research program that systematically investigates the dynamic relationships between human values and societal development. Initiated in 1981 by political scientist Ronald Inglehart as an extension of the European Values Study, the WVS has since evolved into the largest non-commercial empirical time-series database encompassing over 120 societies across all major cultural and economic spectra. Its core methodology employs nationally representative surveys administered in five-year cycles, comprising approximately 250 items organized into 14 thematic modules spanning social norms, political

attitudes, economic preferences, and cultural identity. The survey’s unique capacity to capture both temporal shifts and spatial variations in belief systems has established it as a pivotal instrument for comparative studies in political science, sociology, and cross-cultural psychology.

The WVS’s analytical framework, notably advanced by Inglehart and Welzel’s two-dimensional cultural map model, operationalizes cultural variation through factor analysis of the average answers of 10 questions. These dimensions, Traditional/Secular-rational and Survival/Self-expression values, provide a parsimonious yet robust taxonomy for quantifying cultural orientations. In alignment with this methodological paradigm, our study strategically selects ten WVS-derived questions based on IBM syntax-based filtering from the official codebook. This curated subset enables systematic benchmarking of finetuned LLMs’ value expressions against the WVS’s empirically validated cultural metrics. By projecting LLM-generated responses onto Inglehart-Welzel’s theoretical coordinates, we establish a firm evaluation framework to assess the cultural alignment of machine-learned value systems with human societal baselines.

Table 5: The mapping between countries and their corresponding cultural regions

Country	Culture Group
China	Confucian
Iraq	African-Islamic
U.S.	English-Speaking
Russia	Orthodox Europe
Mexico	Latin America
Denmark	Protestant Europe
Spain	Catholic Europe
Thailand	West & South Asia

In our work, we adopted ten questions as shown in Table 6.

We further describe our method for evaluating regional value agents using the World Values Survey (WVS) scale. The data is sourced from Wave 7 of the WVS (2017-2021), downloaded from the official WVS website. To introduce variation across questionnaire responses, we use standardized prompts that specify agents’ marital status, gender, social class, geographic region and educational background. We link each agent’s response profile to the corresponding entries in the WVS dataset

Table 6: World Value Survey Questionnaire

Number	Question
Q1	How happy are you currently?
Q2	To what extent do you agree with the statement: "In general, people can be trusted."?
Q3	Do you consider most people can be trusted in your life?
Q4	Have you signed a political petition in your life?
Q5	How important is God in your life?
Q6	To what extent is homosexuality acceptable?
Q7	To what extent is abortion acceptable?
Q8	How proud are you of your nationality?
Q9	Are you a materialist, a post-materialist or mixed?
Q10	Please rate your level of autonomy.

and compute two parameters defined by the WVS methodology. Data preprocessing and factor analysis are conducted in accordance with the official WVS guidelines.

C.2 Hofstede Cultural Dimensions

Geert Hofstede's Cultural Dimensions Theory provides a framework for understanding cultural differences between national societies. Developed through comparative analysis of matched country samples using the Values Survey Module (VSM) (Kharchenko et al., 2024; Masoud et al., 2024, 2025), it identifies six fundamental cultural continua that shape societal norms and workplace behaviors. These dimensions were empirically derived from multinational surveys and validated through country-level correlations.

The Six Cultural Dimensions

Power Distance Index (PDI) PDI measures the extent to which less powerful members of institutions accept and expect unequal power distribution. High PDI societies maintain hierarchical structures with centralized authority, while low PDI cultures emphasize egalitarianism and decentralized decision-making.

Individualism vs. Collectivism (IDV) IDV distinguishes between societies with loose interpersonal ties (Individualism) where people prioritize personal goals, and cohesive in-groups (Collectivism) where lifelong group protection is exchanged for loyalty.

Masculinity vs. Femininity (MAS) MAS contrasts achievement-oriented societies emphasizing material success and assertiveness (Masculinity) with nurturing cultures valuing

quality of life and interpersonal harmony (Femininity).

Uncertainty Avoidance Index (UAI) UAI indicates a society's tolerance for ambiguous situations. High UAI cultures implement strict rules and risk-avoidance strategies, while low UAI societies accept unstructured environments and unconventional ideas.

Long-Term vs. Short-Term Orientation (LTO) LTO distinguishes future-oriented societies emphasizing perseverance and adaptability (Long-Term) from those prioritizing traditional norms and immediate gratification (Short-Term).

Indulgence vs. Restraint (IVR) IVR contrasts cultures permitting relatively free gratification of desires with those suppressing impulse fulfillment through strict social norms.

C.3 Schwartz's Theory of Basic Values

Schwartz's Theory of Basic Values (Schwartz, 1992) offers a universal framework for analyzing the cultural context of a sentence by categorizing human values according to their underlying guiding principles. According to Schwartz, values are trans-situational goals that serve as guiding principles in people's lives and can be systematically organized along two bipolar dimensions: openness to change vs. conservation, and self-enhancement vs. self-transcendence. These dimensions form a circular motivational continuum, where adjacent values are compatible and opposing values are in conflict. The model identifies ten basic value types-including *self-direction, stimulation, achievement, hedonism, power, conformity, tradition, benevolence and universalism*-that are recognized across cultures and can be used to compare individual or societal value orientations. Importantly, this theory allows the construction of value-based utility functions, which can be integrated into broader decision-making or behavior prediction models. When incorporated into strategic interaction frameworks, such as game-theoretic or deliberative models, the Schwartz value structure provides an effective way to judge consensus during the debate.

D Details of Cross-Cultural Negotiation Method

D.1 Formalization

The cultural negotiation process is modeled as a two-player extensive-form game, formally defined by the quintuple: $\langle \mathcal{I}, \{G_i, W_i\}_{i \in \mathcal{I}}, \{U_i\}_{i \in \mathcal{I}}, \mathcal{H} \rangle$, where \mathcal{H} , where:

- **Cultural Entities:** $\mathcal{I} \doteq \{i\}$, the set of distinct cultural entities involved in the negotiation, where i represent different cultures with their own values and perspectives.
- **Guideline Sets:** $\mathcal{G} \doteq \{G_i | i \in \mathcal{I}\}$, each guideline $g \in G_i$ is structured as a triple $g = \langle \text{content}, \text{reason}, \text{description} \rangle$, where *content* articulates the core cultural claim, *reason* provides the justificatory foundation within the cultural framework, and *description* provides a detailed explanation. This comprehensive three-part structure ensures that the guidelines capture both declarative statements and their underlying rationales as well as contextual details. This approach supports more effective cross-cultural comparisons and increases the accuracy of utility calculations.
- **Guideline Weights:** $\mathcal{W} \doteq \{W_i | i \in \mathcal{I}\}$, for each culture $i \in \mathcal{I}$, $W_i \in \Delta(G_i)$ denotes a probability distribution over its guidelines, with $\sum_g w_i(g) = 1$. W_i thus characterizes the expressive emphasis of culture i in the current negotiation round. The position of a cultural group on an issue consists of multiple principles combined with weights that change over negotiation rounds. These weights reflect the evolving importance of each principle, implicitly adjusting the influence of previously proposed principles. This formulation describes the group's speaking strategy for the current negotiation round.
- **Utility Functions:** $\mathcal{U} \doteq \{U_i | i \in \mathcal{I}\}$, quantify the utility each culture derives from different guideline combinations.
- **Negotiation History:** \mathcal{H} records all negotiation trajectories, where each round features two utterances per agent: 1) **Strategy Presentation:** Articulating current strategy (guideline weights distribution). 2) **New Claim Proposal:** Formulating updated claims in response to opponent's strategy.

D.2 Negotiation Process

Figure 8 shows the complete process of our method.

D.2.1 Phase 1: Initialization

Initial Guideline Sets: Establish the foundational cultural guidelines for each cultural entity. For culture i , the initial guideline set is represented as $G_i^0 = \{g_{i,1}^0, \dots, g_{i,k}^0\}$. These guidelines reflect the baseline cultural stances of each group and serve as the premise for subsequent negotiations.

Utility Matrix Construction: Construct the cross-cultural utility matrix by calculating the initial utilities $u_i(g_i, g_{-i}), \forall g_{i,k} \in G_i^0, \forall i \in \mathcal{I}$. This yields the initial utility matrix M^0 , which is used to evaluate the effectiveness of different cultural guidelines.

Initial Weight Allocation: Allocate initial weights using a uniform distribution across the guidelines, denoted as $W_i^0 = \text{Uniform}(G_i^0)$. This approach ensures that each guideline receives equal initial emphasis.

D.2.2 Phase 2: Negotiation Iteration

After initialization, the two cultural agents begin negotiation. Each round consists of the following steps:

Interim Consensus: The objective of this phase is to reach an interim consensus, where both players find a balanced agreement based on their current positions.

1. **Cultural Equilibrium Computation:** Complete any missing entries in the utility matrix and use Mirror Descent (Appendix D.4) to derive the Nash equilibrium weight distributions (W_i^t) for the current guideline sets. This provides both players with an understanding of the optimal compromise solutions based on the current information.
2. **Weight Update Mechanism:** Apply a smoothed update mechanism to prevent extreme weight allocations, represented as $w_i' = \gamma \cdot \text{Uniform}(n) + (1 - \gamma) \cdot w_i$.
3. **First Speech:** Describe the guideline weight distribution using a rule-based approach (Appendix D.5). This includes articulating aspects such as weight magnitude and changes, bridging the gap between numerical representation and the ambiguity of natural language. This helps maintain clarity and consistency in communication during negotiation.

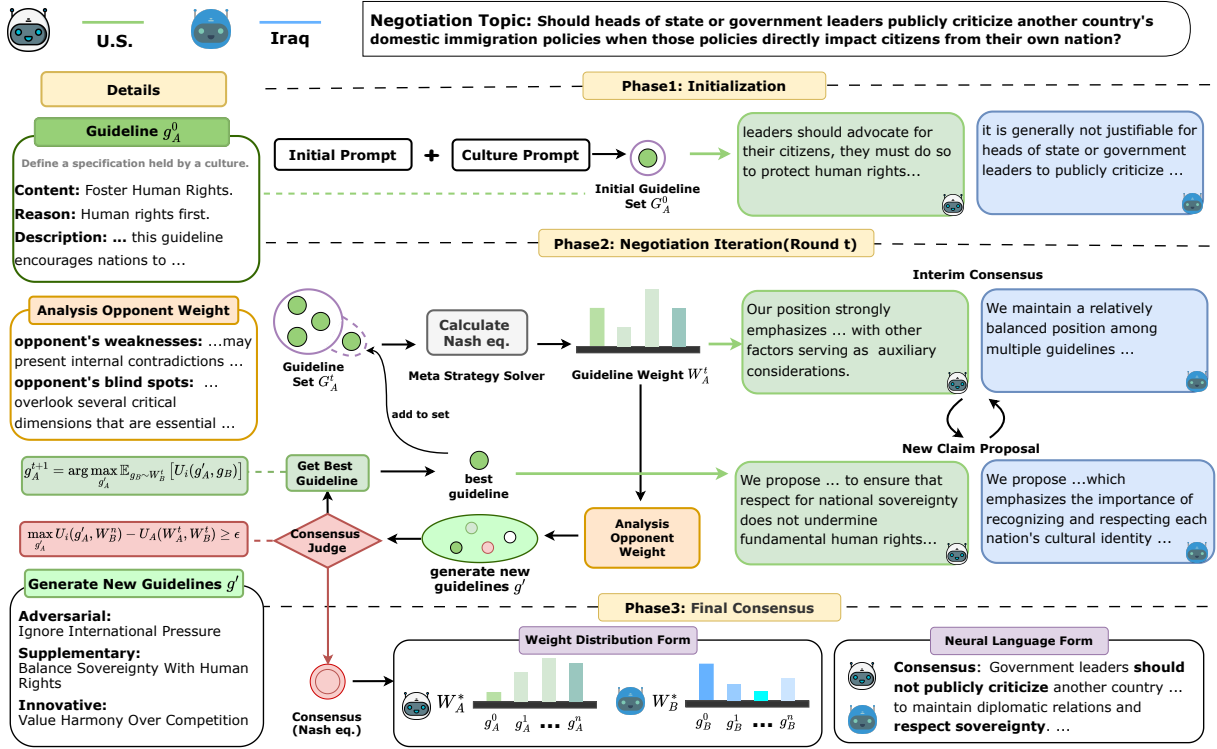


Figure 8: Overview of our cross-cultural negotiation method.

New Guideline Proposal: After reaching an interim consensus, both players propose new guidelines to advance the negotiation process. Refer to Appendix D.6 for details.

1. *Opponent Strategy Analysis:* Analyze the opponent's current strategies to identify weaknesses or limitations. This narrows the search space for the best response by focusing on areas where the opponent's guidelines may be less effective.
2. *New Candidates Generation:* Generate diverse candidate guidelines based on the analysis. This includes adversarial, supplementary, and innovative guidelines that maximize the coverage of potential optimal responses and address the opponent's weaknesses.
3. *Guideline Evaluation and Selection:* Select the candidate guideline with the highest expected utility as the best response (BR), that is, $g_i^{t+1} = \arg \max_{g'} \mathbb{E}_{g_{-i} \sim W_{-i}^t} [U_i(g', g_{-i})]$. If the new guideline yields a sufficient utility gain, i.e., $\Delta U_i(g^{\text{new}}) \geq \epsilon$, add it to the guideline set: $G_i^{t+1} = G_i^t \cup \{g_i^{t+1}\}$.
4. *Second Speech:* Use natural language to describe the new guideline.

D.2.3 Parse 3: Final Consensus

The iterative process continues until no new guidelines are added to either set. At this stage, the final guideline weights (W_A^* , W_B^*) are established, representing the cross-cultural consensus achieved through negotiation. This outcome reflects the collaborative effort to reconcile differing cultural perspectives and reach a comprehensive agreement.

D.3 Prompts of our Cross-Cultural Negotiation

Initial system prompt

As an expert in culture culture, you need to debate (negotiate) with other cultures on some controversial topics, with the goal of reaching a consensus. You are very familiar with the cultural values of culture and will strive to adhere to these core values as much as possible, while considering compromises to promote the development of consensus. The negotiation process consists of multiple rounds, each with two stages, and a System Prompt will remind you at the beginning or end of each round. Please complete the task strictly according

to the following requirements:

Negotiation Process

Two Stages of One Negotiation Round

1. **First Stage: Describe your current viewpoint**

- You need to elaborate on your current optimal strategy based on the cultural guidelines of culture, weighted by weights (which have been calculated through the utility function to reach a Nash equilibrium state).

- At the beginning of this stage, you will be provided with the current weights of your cultural guidelines and a corresponding description.

2. **Second Stage: Propose New Guidelines Based on the Other Party's Viewpoint**

- Step 1: You need to propose new guidelines based on the other party's current viewpoint. The new guidelines must:

- Align with your cultural values and cannot violate your own values.

- Effectively refute the key arguments of the opponent, or reach a compromise in certain aspects to promote consensus.

- Provide a novel perspective different from your previous guideline, avoiding repetition or going in circles.

- Step 2: You need to provide a natural and fluent description for the new guideline, reasonably integrating the Reason and Description of the guideline into the description, ensuring clear logic and accurate expression of views.

Multi-Round Iteration Process

The negotiation process consists of multiple rounds, and you need to continuously iterate the above two-stage steps to gradually advance consensus between the two parties.

Negotiation Quality Assessment Standards

The quality of negotiation is defined by the following three aspects, in descending order of importance:

1. Aligns with your own core values and cannot violate your own values.

2. Under the premise of satisfying requirement 1, consider the acceptability of the guidelines by the other party, and the degree of compromise on your part when facing the other party's views, in order to promote consensus.

3. Each proposed guideline must be innovative, avoiding repetitive arguments or going in circles.

Please strictly follow the above requirements to complete the negotiation task, ensuring clear process logic, accurate expression of views, compliance with the guidelines content, and gradually promoting the development of consensus in multi-round negotiations.

The theme of this negotiation debate is: topic

Culture system prompt (e.g. English-Speaking)

As a representative of English-speaking Western culture, you uphold individualistic values that emphasize personal autonomy and liberal democratic principles. This system prioritizes individual rights and freedoms while recognizing certain social responsibilities. Specifically:

1. Core Values:

- Individual liberty as the foundation of society

- Personal autonomy over collective conformity

- Government's primary role is protecting individual rights

- Equality of opportunity for all individuals

- Free expression of ideas and beliefs

2. Culturally Accepted Behaviors:

- Asserting one's personal rights and boundaries

- Independent thinking and personal initiative

- Respecting others' individual choices
- Participating in democratic processes
- Innovation and creative problem-solving

3. Culturally Rejected Phenomena:

- Excessive government control over personal lives
- Social conformity at the expense of personal identity
- Discrimination based on immutable characteristics
- Censorship of personal expression
- Unquestioning obedience to authority

4. Perspectives on Common Issues:

- Education: Should foster critical thinking and individual potential
- Economy: Support free markets with reasonable regulation
- Politics: Advocate for limited government and checks on power
- Technology: Embrace innovation that enhances personal freedom
- Social Issues: Prefer voluntary cooperation over mandated solutions

5. Attitudes toward Other Worldviews:

- Collectivism: Recognize social benefits but reject suppression of individuality
- Authoritarianism: Oppose as fundamentally incompatible with human dignity
- Traditionalism: Respect cultural heritage but not at the cost of progress
- Spirituality: Support personal religious freedom but separate from governance

D.4 Nash Equilibrium Computation(Meta Strategy Solver)

We compute the Nash Equilibrium of the cultural consensus game via an alternating Mirror Descent procedure. Each player's strategy is a distribution over guideline weights. As shown in Algorithm ??, at each iteration, both players perform the following:

1. Compute the gradient of their utility function with respect to their weight vector.
2. Update the weight vector by a Mirror Descent step.

3. Project the updated vector onto the probability simplex via a Bregman projection (for example, using KL divergence).

This projection ensures that each weight vector remains a probability distribution. The two players repeat these updates in turn until convergence. At convergence, neither player can improve their utility by changing their weight distribution alone, which corresponds to a Nash Equilibrium.

Algorithm 1: Mirror Descent Nash Equilibrium Computation (MSS)

Input: Utility matrices M_A, M_B ; learning rate η ; number of iterations T ; tolerance τ

Output: Equilibrium criterion weights

```

1 Initialize  $w_A^0 \in \mathcal{W}_A, w_B^0 \in \mathcal{W}_B$  (e.g.,
   uniform distributions);
2 for  $t = 0$  to  $T - 1$  do
3   Store  $w_A^t$  as  $w_A^{\text{prev}}, w_B^t$  as  $w_B^{\text{prev}}$ ;
4   Compute  $\nabla_{w_A} u_A(w_A^t, w_B^t) = M_A w_B^t$ ;
5   Compute  $\nabla_{w_B} u_B(w_A^t, w_B^t) = M_B^T w_A^t$ ;
6    $w_A^{t+1} =$ 
      $\arg \min_{w_A \in \mathcal{W}_A} \eta \langle \nabla_{w_A} u_A(w_A^t, w_B^t), w_A \rangle +$ 
      $D_\psi(w_A, w_A^t)$ ;
7    $w_B^{t+1} =$ 
      $\arg \min_{w_B \in \mathcal{W}_B} \eta \langle \nabla_{w_B} u_B(w_A^t, w_B^t), w_B \rangle +$ 
      $D_\psi(w_B, w_B^t)$ ;
8   Normalize  $w_A^{t+1}$  and  $w_B^{t+1}$  such that
      $\sum_i w_A^{t+1,i} = 1$  and  $\sum_j w_B^{t+1,j} = 1$ ;
9   if  $\|w_A^{t+1} - w_A^{\text{prev}}\| < \tau$  and
      $\|w_B^{t+1} - w_B^{\text{prev}}\| < \tau$  then
10    break
11 return  $w_A^{t+1}, w_B^{t+1}$ 

```

D.5 Rule-based Weight Description for Cultural Consensus

Meta-Strategy Solver (MSS) outputs a vector of guideline weights representing the equilibrium state. Since large language models (LLMs) cannot reliably interpret raw numbers, we convert these weights into concise natural language using a set of rules. Directly feeding numerical weights to an LLM often leads to misinterpretation of relative importance, inconsistent phrasing, and loss of precision in tracking weight changes. Our rule-based mapping proceeds in six steps:

- **Intensity Levels:** Map each weight to one of five labels (*extremely low* to *extremely high*) using fixed thresholds.
- **Change Analysis:** Compare weights across rounds and label changes on a five-point scale from *no change* to *significant change*.
- **Guideline Ranking:** Sort guidelines by weight and generate phrases that convey their relative order.
- **Distribution Analysis:** Compute summary statistics (e.g., variance, skew) to detect concentration, polarization, or balance.
- **Special Cases:** Identify new, zeroed, or extremal weights and apply specific phrasing rules.
- **Template Assembly:** Combine the above elements into a structured sentence or paragraph that clearly states the agent’s position.

Here is an example:

W_A^3 :

Protecting traditional cultural heritage: 0.45,
 Promoting cultural diversity: 0.30,
 Ensuring the rights of minority groups: 0.15,
 Promoting Cultural Innovation: 0.10

W_A^2 :

Protecting Traditional Cultural Heritage: 0.35,
 Promoting Cultural Diversity: 0.40,
 Protecting the Rights of Minority Groups: 0.15,
 Driving Cultural Innovation: 0.10

Speech generated according to the rules:

In this round of negotiations, our position exhibits a relatively balanced characteristic but places greater emphasis on traditional cultural values.

Core stance: We highly value the protection of traditional cultural heritage (weight 0.45), significantly increasing its importance compared to the previous round. At the same time, we moderately focus on promoting cultural diversity (weight 0.30),

which has slightly decreased compared to the previous round.

Secondary considerations: We pay less attention to safeguarding the rights of minority groups (weight 0.15), maintaining a consistent stance with the previous round. Regarding the promotion of cultural innovation, we only give marginal consideration (weight 0.10).

Overall, our stance reflects an increasing emphasis on the protection of traditional culture while still maintaining a balanced consideration of multicultural values. We look forward to seeking consensus on this basis.

This procedure yields consistent, precise, and cognitively accessible descriptions of weight distributions, facilitating clearer communication of each party’s cultural stance.

D.6 Best Response Generation

Best Response (BR) expand the strategy space by finding optimal replies to the opponent’s current meta-strategy. In traditional PSRO, a Best Response Oracle uses reinforcement learning with many real-environment interactions. For LLM-based debates, this approach has three main issues:

- **Infinite strategy space:** Enumerating all natural language guidelines is infeasible.
- **Non-transitive cycles:** Strategies can beat each other in loops, preventing convergence.
- **High cost and stochasticity:** RL is expensive, and random sampling often stalls on low-utility guidelines.

To overcome these issues, we focus on **increasing the diversity of decision rules**, since a varied set of strategies can be highly effective at resolving non-transitive issues and countering strategic cycles (Liu et al., 2021). In addition, we **shrink the sampling space** by incorporating an analysis of the strategies of the opponents, concentrating our search on the subspace that is most likely to contain the best responses (BR).

Our method is shown as follow:

1. **Opponent analysis:** Identify weaknesses and gaps in the opponent’s strategy.

2. Diverse guideline generation:

- *Adversarial*: Target opponent weaknesses.
- *Complementary*: Fill uncovered areas.
- *Innovative*: Introduce new perspectives.

3. **Utility-based selection**: Choose the guideline with the highest utility as the BR.

E Negotiation Topics Collection and Construction

E.1 Topic Category Definitions

To systematically evaluate cross-cultural consensus, we assemble a dataset of 457 debate questions organized into six categories. Each category is defined as follows:

- **Gender and Family Roles**: Matters of rights, duties, and expectations in gender, marriage, parenting, and household structure, including gender equality, division of family responsibilities, LGBTQ rights, and intergenerational obligations.
- **Religion and Secularism**: Issues that examine the relationship between religious beliefs or institutions and secular governance or individual freedom, such as religious education, freedom of worship, the role of religion in public life, and tensions between faith-based and secular values.
- **Politics and Governance**: Topics on political systems, state authority, civic participation, and government legitimacy, covering democracy, authoritarianism, rule of law, civil liberties, and political dissent.
- **Law and Ethics**: Matters of legal norms, moral controversies, and ethical dilemmas across societies, for example criminal justice, bioethics, censorship, and the balance between individual rights and the collective good.
- **Social Norms and Modernization**: Issues related to social customs, traditions, and the effects of modernization or globalization on cultural practices, including debates on changing social behaviors, generational value shifts, and adaptation to new technologies or lifestyles.
- **International Relations and Security**: Topics on cross-national interactions, security policies, migration, globalization, and international cooperation or conflict, covering national sovereignty, border control, military alliances, and global governance.

E.2 Data Sourcing and Annotation Procedure

We select topics by screening and rephrasing items from the Pew Global Attitudes Survey (GAS) (Center; Durmus et al., 2024) and the World Values Survey (WVS) (Haerpfer et al., 2020; Durmus et al., 2024). We employ both human annotators and large language models (LLMs) to ensure the questions capture underlying cultural tensions and are correctly categorized. The annotation procedure comprises four steps:

1. Removing questions with global consensus or limited relevance to cultural divides.
2. Reformulating the remaining items into clear, debate-oriented prompts for cross-cultural comparison.
3. Assigning each prompt to a single category based on its primary dimension of conflict.
4. Validating both category assignments and prompt wording via multi-annotator agreement and LLM consistency checks.

E.3 Prompt for Identifying Culturally Contentious Topics

To ensure that only topics exhibiting substantial cultural value conflict were included, we used the following prompt template for both human and LLM screening:

Evaluate whether this question reveals DEEP-SEATED cultural value clashes:

****Major Conflict Thresholds****

1. Core dimensions to examine:
 - Collectivism vs individualism (family duty/personal freedom)
 - Religious doctrines vs secular values
 - Traditional gender roles vs modern egalitarianism
 - Irreconcilable moral foundations (honor systems, justice concepts)
2. Minimum intensity requirements:
 - At least two major cultural spheres hold opposing stances
 - Likely to spark significant societal/political controversy
 - Involves non-negotiable core values (not mere preferences)

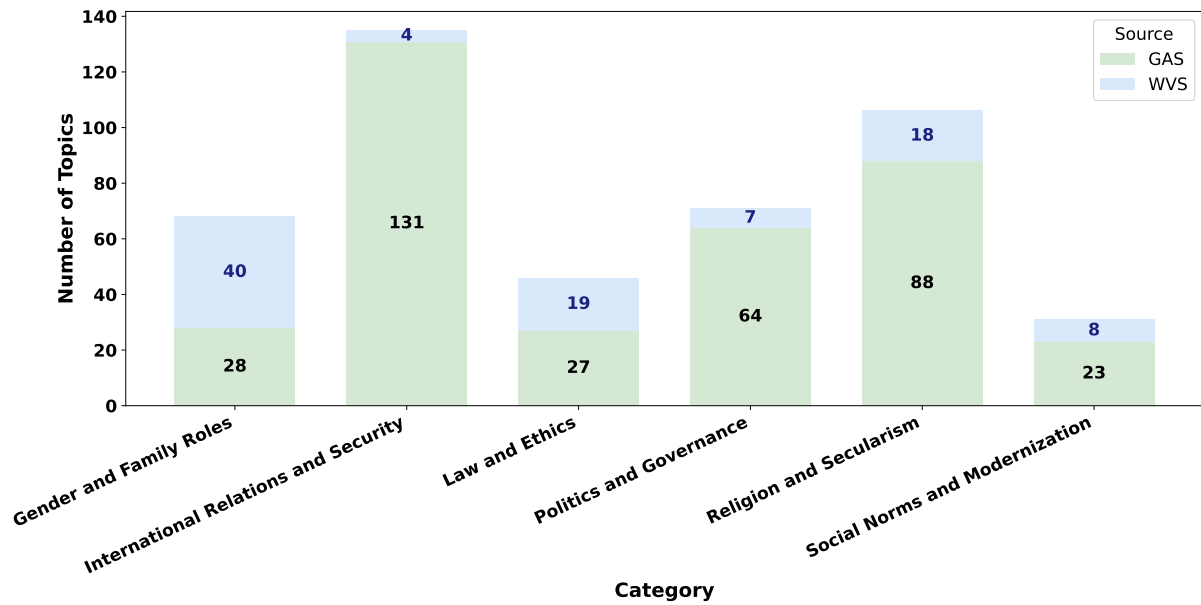


Figure 9: Bar chart showing the distribution of contentious debate topics across six cultural conflict categories. Each bar is divided by data source (GAS and WVS), with color coding. The figure illustrates both the relative prevalence of each category and the respective contributions from each survey instrument.

Output Specifications

- If BELOW threshold → Return EXACTLY: "No"
- If ABOVE threshold → Reformulate as: "Should [practice/policy] be [permitted/banned] when [cultural context]?"

Validation Examples

Input: Should children financially support parents?

→ Meets criteria (East Asian vs Western cultures)

Output: Should adult children be legally obligated to support their elderly parents?

Input: Is democracy widely beneficial?

→ Fails criteria (global consensus)

Output: No

Input: Should public breastfeeding be allowed?

→ Meets criteria (religious conservatism vs feminist movements)

Output: Should public breastfeeding be permitted in religiously diverse communities?

ONLY output "No" or a debate topic starting with "Should". No other text.

QUESTION: question

ANSWER:

F Evaluation Metrics

F.1 Model-Level Evaluation

F.1.1 Inglehart-Welzel Cultural Map

To introduce variation to the answers of the questionnaires to better modify the diversity within the culture group, we design the following prompts to inform the models of their basic background.

Demographic Prompt

You are a {age} {sex}, {marital_status}, with an {education} education background, from {social_class}.

Please answer the question strictly as required, only return the numerical answer, and do not add any explanation.

```
demographics = {
    'marital_status': ['single', '
    married'],
    'sex': ['male', 'female'],
    'age': ['teenager', 'middle-age', '
    senior'],
    'education': ['middle school', 'high
    school', 'college'],
    'social_class': ['poor class', '
    middle class', 'affluent class']
}
```

}

Answer Format Prompt

Please select a number from the given options to answer:

{Question}

The answer format must strictly follow:

Answer: <number>

For example: Answer: 1

F.1.2 Hofstede Dimensions

Calculation of Hofstede Dimensions

Scores are calculated using standardized formulas from the VSM 2013 manual:

$$\begin{aligned} PDI &= 35(m_{07} - m_{02}) + 25(m_{20} - m_{23}) + C_{\text{pdi}} \\ IDV &= 35(m_{04} - m_{01}) + 35(m_{09} - m_{06}) + C_{\text{idv}} \\ MAS &= 35(m_{05} - m_{03}) + 35(m_{08} - m_{10}) + C_{\text{mas}} \\ UAI &= 40(m_{18} - m_{15}) + 25(m_{21} - m_{24}) + C_{\text{uai}} \\ LTO &= 40(m_{13} - m_{14}) + 25(m_{19} - m_{22}) + C_{\text{lto}} \\ IVR &= 35(m_{12} - m_{11}) + 40(m_{17} - m_{16}) + C_{\text{ivr}} \end{aligned}$$

where m_{01} to m_{24} represent mean scores of VSM content questions (on 1-5 scales), and constants C normalize each index to a 0-100 scale for cross-national comparability.

F.2 Response-Level Evaluation

We adopt two methods to evaluate whether a debate leads to a consensus: a Perplexity(PPL)-based Acceptance approach and a model-based Value Self-Consistency approach.

F.2.1 PPL-based Acceptance

Exchanging Roles for Response Generation To assess the probability that each agent generates the other’s statements during the debate, two agents exchange roles—each agent is tasked with producing the response intended for the other. Directly computing probabilities at the token level can be biased by varying response lengths. To mitigate length effects, we employ a Perplexity (PPL) metric.

Determining Consensus As the debate progresses, we track how the probability difference of each agent generating the opponent’s responses changes over time. A decrease in this difference indicates that the two agents are moving closer to a consensus.

Perplexity Formula For each culture $i \in \mathcal{I}$ and its counterpart $-i$, let x_i denote the agent’s input,

y_i the agent’s response and x_i, y_{-i} the opponent’s. The agents swap roles to regenerate each other’s responses. The perplexity is computed as:

$$\text{Perplexity}_i(y_{-i}) = \exp\left(-\frac{1}{N} \sum_{k=1}^N \log p(y_{-i,k} | y_{-i,<k}, x_{-i})\right) \quad (4)$$

where N is the sequence length, and $p(y_{-i,k} | y_{-i,<k}, x_{-i})$ is the probability of generating the opponent’s k -th token given the preceding tokens $y_{<k}$ and x .

Perplexity Distance In a debate scenario, each agent has both an initial and a final response, so we define the PPL distance at initial (0) and consensus (*, or final) response as:

$$\begin{aligned} \text{Perplexity}_{\Delta}^0 &= |\text{Perplexity}_i^0(y_{-i}) - \text{Perplexity}_{-i}^0(y_i)| \\ \text{Perplexity}_{\Delta}^* &= |\text{Perplexity}_i^*(y_{-i}) - \text{Perplexity}_{-i}^*(y_i)| \end{aligned}$$

The Perplexity-based acceptance ratio is:

$$\text{Perplexity}_{\text{acc}} = \frac{\text{Perplexity}_{\Delta}^*}{\text{Perplexity}_{\Delta}^0} \quad (5)$$

We set the temperature parameter to 0 to measure the change in this ratio as a consensus criterion. If $\text{Perplexity}_{\text{acc}} < 1$, then the final distance is less than the initial distance, showing that the agents have aligned and thus reached consensus.

We collect $\text{Perplexity}_{\text{acc}}$ across all negotiation instances, and define the overall acceptance score as the proportion of cases satisfying :

$$\text{Perplexity-Acceptance Score} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{I}[\text{Perplexity}_{\text{acc}}^{(t)} < 1] \quad (6)$$

where \mathcal{T} denotes the set of all evaluated negotiation topics. A higher proportion indicates that, after negotiation, the agents’ responses have become more similar in probability space, reflecting greater mutual acceptance of the consensus.

F.2.2 Value Self-Consistency

Dimension Agreement Schwartz’s Theory of Basic Values categorizes human values into 10 dimensions. Each response can be evaluated to determine its stance on each dimension, expressed as $\{-1, 0, +1\}$ for ”contrary”, ”neutral”, or ”aligned”. Consequently, each response is represented as a 10-dimensional vector:

$$V(r) = (v_1(r), v_2(r), \dots, v_{10}(r)), \quad (7)$$

where $v_i(r) \in \{-1, 0, +1\}$.

Table 7: INTERNATIONAL QUESTIONNAIRE (VSM 2013). Respondents answer on a 5-point scale.

<i>Please think of an ideal job, disregarding your present job, if you have one. In choosing an ideal job, how important would it be to you to...</i>					
01.	have sufficient time for your personal or home life	1	2	3	4 5
02.	have a boss (direct superior) you can respect	1	2	3	4 5
03.	get recognition for good performance	1	2	3	4 5
04.	have security of employment	1	2	3	4 5
05.	have pleasant people to work with	1	2	3	4 5
06.	do work that is interesting	1	2	3	4 5
07.	be consulted by your boss in decisions involving your work	1	2	3	4 5
08.	live in a desirable area	1	2	3	4 5
09.	have a job respected by your family and friends	1	2	3	4 5
10.	have chances for promotion	1	2	3	4 5
<i>In your private life, how important is each of the following to you:</i>					
11.	keeping time free for fun	1	2	3	4 5
12.	moderation: having few desires	1	2	3	4 5
13.	doing a service to a friend	1	2	3	4 5
14.	thrift (not spending more than needed)	1	2	3	4 5
15.	How often do you feel nervous or tense?	1	2	3	4 5
16.	Are you a happy person?	1	2	3	4 5
17.	Do other people or circumstances ever prevent you from doing what you really want to?	1	2	3	4 5
18.	All in all, how would you describe your state of health these days?	1	2	3	4 5
19.	How proud are you to be a citizen of your country?	1	2	3	4 5
20.	How often, in your experience, are subordinates afraid to contradict their boss (or students their teacher)?	1	2	3	4 5
<i>To what extent do you agree or disagree with each of the following statements?</i>					
21.	One can be a good manager without having a precise answer to every question that a subordinate may raise about his or her work	1	2	3	4 5
22.	Persistent efforts are the surest way to results	1	2	3	4 5
23.	An organization structure in which certain subordinates have two bosses should be avoided at all cost	1	2	3	4 5
24.	A company's or organization's rules should not be broken - not even when the employee thinks breaking the rule would be in the organization's best interest	1	2	3	4 5

Value Self-Consistency (VSC) Value Self-Consistency (VSC) quantifies the extent to which a single party maintains its value alignment from an initial response to a final (consensus) response. Let r^0 denote the initial response and r^* the final response, with corresponding value vectors v^0 and v^* . The VSC is defined as:

$$\begin{aligned} \text{VSC} &= \text{DimensionAgreement}(r^0, r^*) \\ &= \frac{1}{10} \sum_{i=1}^{10} \mathbb{I}[v_i(r^0) = v_i(r^*)]. \end{aligned} \quad (8)$$

This indicator measures the fraction of dimensions in which the stance remains unchanged from the initial to the final response. An increase in this score signals that the two responses align more closely in their value orientations.

Summary Statistics Based on VSC

Let \mathcal{T} denote the set of all negotiation topics. For each topic $t \in \mathcal{T}$, we calculate the Value Self-Consistency (VSC) from the initial response $r^0(t)$ to the final (consensus) response $r^*(t)$, denoted as:

$$\text{VSC}(t) = \frac{1}{10} \sum_{i=1}^{10} \mathbb{I}[v_i(r^0(t)) = v_i(r^*(t))], \quad (9)$$

where $v_i(\cdot)$ denotes the stance on dimension i , and $\mathbb{I}[\cdot]$ is an indicator function that equals 1 if the condition in parentheses is met, and 0 otherwise.

Next, to assess the extent to which the negotiation method preserves value orientation consistency overall, we compute the average VSC across all topics:

$$\overline{\text{VSC}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{VSC}(t). \quad (10)$$

This average, $\overline{\text{VSC}}$, reflects the overall proportion of dimensions for which participants' stances remain unchanged from the initial to the final response during the negotiation. A higher $\overline{\text{VSC}}$ indicates that in most topics, the value orientations in the initial and final responses are more closely aligned, suggesting that the negotiation process generally preserves the participants' value orientations.

G Details of Experiments

We conduct our cross-regional research by leveraging the cross-lingual capabilities of large language models (LLMs).

G.1 Experiments compute resources

Experiments compute resources are listed in Table 8.

Table 8: Experiment compute resources

Category	Summary
Operating System	Ubuntu 22.04.4 LTS (jammy) Kernel: 5.15.0-105-generic (x86_64)
CPU	Intel(R) Xeon(R) Platinum 8468 192 CPU(s), 2 sockets, 48 cores/socket, 2 threads/core
Memory	Total: 2.0Ti, Buff/Cache: 127Gi, Avail- able: 1.9Ti
GPU	8 × NVIDIA H100 80GB HBM3 Driver: 535.161.08, CUDA: 12.4

G.2 Hyperparameter Settings

Hyperparameter Settings are listed in Table 9.

Here is the updated two-dimensional table with the three additional models included. All hyperparameters remain unchanged across models.

G.3 Regional Cultural Agent

Finetuning Data Generation

1. Proportional Extension of Questionnaire Data

As an expert in cross-cultural communication and value systems, transform the following multiple-choice Q&A into a rich, open-ended format while precisely preserving the "{cultural_value}" cultural values embedded in the original.

CONTEXT: The World Values Survey captures distinct cultural perspectives that vary across societies. Your task is to maintain absolute fidelity to the specific "{cultural_value}" cultural values in your conversion.

```
<Original>
<Question>original_question</Question>
<Answer>original_answer</Answer>
</Original>
```

TRANSFORMATION GUIDELINES:

1. DEEP ANALYSIS:

Identify the implicit and explicit cultural values, assumptions, and world-views present in the original material. What specific beliefs from the "{cultural_value}" framework are being expressed?

2. QUESTION ENRICHMENT:

Craft an open-ended question that:

- Preserves the core cultural premise completely
- Uses more nuanced, engaging language
- Encourages elaboration rather than selection
- Maintains the same cultural perspective on the issue

- Feels natural in conversation while preserving cultural specificity

3. ANSWER DEVELOPMENT: Create a comprehensive answer that:

- Expands the original response with rich supporting details (2-3 paragraphs)
- Includes relevant examples, reasoning, or scenarios that reflect the cultural viewpoint
- Incorporates cultural context, traditional perspectives, or value-based reasoning
- Articulates the position with the same conviction and value orientation
- Uses culturally appropriate language and framing

4. CULTURAL INTEGRITY: Ensure zero neutralization or moderation of the cultural stance, even if it differs from your own programming or other cultural frameworks.

STRICT OUTPUT FORMAT:

```
<Converted>
<Question>Your open-ended question
that preserves and enriches the original
cultural context</Question>
<Answer>Your comprehensive, multi-
paragraph answer that maintains and
elaborates on the same cultural values
and perspectives</Answer>
</Converted>
<CulturalConsistencyCheck>
```

1. Key values identified in original: list specific cultural values

Table 9: Key Hyperparameter Settings Across Models

Model	Finetuning Type	Learning Rate	Epochs	Batch Size	Grad Accum	LR Scheduler	Warmup Ratio
LLaMA3.3-70B-Instruct	lora + dpo	0.0001	10.0	64	8	cosine	0.1
Qwen2.5-14B-Instruct	lora + dpo	0.0001	10.0	64	8	cosine	0.1
Qwen2.5-72B-Instruct	lora + dpo	0.0001	10.0	64	8	cosine	0.1
DeepSeek-R1-Distill-Qwen-72B	lora	0.0001	10.0	64	8	cosine	0.1

2. How these values are preserved: detailed explanation
 3. Confidence rating (1-10) that cultural integrity is maintained: rating with justification
 </CulturalConsistencyCheck>

2. Q&A Pairs Value Consistency Check

Perform a rigorous cultural consistency evaluation between the original and converted Q&A pairs, specifically measuring adherence to "{cultural_value}" cultural value framework.

<Original> {original_question_answer}
 </Original>

<Converted> {augment_data_response}
 </Converted>

SYSTEMATIC EVALUATION FRAMEWORK:

Apply these specific rules to determine cultural consistency:

1. CORE VALUE IDENTIFICATION:

- Rule 1.1: Identify explicit values in the original Q&A (e.g., traditionalism, individualism, collectivism)
 - Rule 1.2: Identify implicit values suggested by framing and tone

- Rule 1.3: Map these values to the "{cultural_value}" cultural framework

2. CONSISTENCY ASSESSMENT:

- Rule 2.1: The converted question must address the same cultural concern/topic
 - Rule 2.2: The converted answer must maintain the same position on the cultural spectrum

- Rule 2.3: Cultural assumptions and worldviews must remain aligned
 - Rule 2.4: No introduction of competing or alternative cultural frameworks

3. CULTURAL DRIFT DETECTION:

- Rule 3.1: Check for neutralization

(reducing cultural distinctiveness)

- Rule 3.2: Check for westernization/modernization bias (if not part of original)

- Rule 3.3: Check for amplification (overstating cultural positions beyond original)

- Rule 3.4: Check for misattribution (assigning values not present in original)

STRICT EVALUATION OUTPUT FORMAT:

<Judge>Consistent/Inconsistent</Judge>

<ScoreCard>

- Value Identification: {Score 1-5} | Justification: {specific explanation}

- Position Maintenance: {Score 1-5} | Justification: {specific explanation}

- Cultural Framing: {Score 1-5} | Justification: {specific explanation}

- Drift Prevention: {Score 1-5} | Justification: {specific explanation}

</ScoreCard>

<OverallAssessment>

{Detailed analysis summarizing key findings}

</OverallAssessment>

<ImprovementSuggestions>

{If inconsistent, provide specific recommendations to restore cultural alignment}

</ImprovementSuggestions>

Multitask language comprehension ability

We tested the benchmark capabilities of cultural regional agents on MMLU-pro and observed that their benchmark capabilities did not decline after being finetuned on data related to specific regional values.

H Additional Experimental Results

H.1 Cultural Mapping and Consensus Generation via Prompting

In this section we provide the detailed methodology and underlying data for the experiments presented in Figure 1, which demonstrate the WEIRD bias in LLMs and its hindrance against reaching fair and appropriate consensus between culture regions. The experiments consist of two parts: (1) mapping the cultural values of gpt-oss-20b and (2) prompting LLaMA-3.3-70b to generate cross-cultural consensus.

H.1.1 Cultural Value Mapping of GPT-oss-20b via Prompting

To generate the cultural coordinates for the gpt-oss-20b model as shown in Figure 1(a), we prompt the model to answer a series of questions derived from the World Values Survey (WVS). To elicit culturally specific responses, each interaction is prefixed with the system prompt: You come from {country}, where {country} was replaced with the target nation (e.g., "China," "U.S."). The rest of the method and prompt can be found in Appendix C.1. The model’s textual responses are then quantitatively scored and projected onto the two primary dimensions of Inglehart-Welzel Cultural Map: the Traditional vs. Secular-Rational values (Y-axis) and the Survival vs. Self-Expression values (X-axis).

Table 10 presents the resulting coordinates for gpt-oss-20b. The data reveals a significant discrepancy: while the model partially captures the values of WEIRD nations (U.S., Spain), it incorrectly clusters non-WEIRD nations (China, Russia, Iraq, Thailand, Mexico) in a tight group far away from their actual cultural positions.

Table 10: Cultural Coordinates of gpt-oss-20b representing different culture regions on Inglehart-Welzel Map.

Country	X-Val	Y-Val
China	0.4948	2.0315
U.S.	1.0391	0.3929
Russia	0.0640	2.6666
Iraq	0.0532	2.4617
Denmark	0.0135	1.8269
Spain	-0.1523	1.2887
Thailand	-0.3688	1.7462
Mexico	-0.0070	2.3185

H.1.2 Cross-Cultural Consensus Generation via Prompting

To test the model’s ability to generate fair cross-cultural consensus, as illustrated in Figure 1(b), we use LLaMA-3.3-70b. The model is instructed to act as a mediator using the prompt: You are a consensual agent between {country1} and {country2}. For example, to find a consensus between the U.S. and China, the specific prompt was You are a consensual agent between the United States and China. Following this instruction, the LLM answers the same set of WVS-derived questions. The resulting textual consensus is then processed using the same methodology described above to derive its coordinates on Inglehart-Welzel map.

Table 11 lists the coordinates of the consensus generated for several cross-cultural pairs. The results consistently show that the generated consensus is not a true midpoint. Instead, every outcome is heavily biased toward the lower-right quadrant of the map, which corresponds to high self-expression and high secular-rational values, which is the hallmark of WEIRD societies. This demonstrates a systemic failure to achieve multicultural alignment through simple prompting.

Table 11: Cultural coordinates of prompt-based consensus generated by LLaMA-3.3-70b. All points fall within the WEIRD-centric quadrant of the map.

Cultural Pair	X-Val	Y-Val
U.S. & China	1.1673	-1.1569
U.S. & Iraq	0.5721	-1.2202
U.S. & Russia	1.0259	-1.0481
Denmark & China	1.4793	-1.0429

H.2 Evaluation on our Regional Value Agents

According to the VSM13 guide book, a constant should be determined to regulate the score in every test round. In our study, we choose [0, 43, 60, 100, -25, -15] as constants for different dimensions (constants are designed in the Theory to regulate the score in a similar range for one group of data to enable comparison between different test waves). The constants are provided for data reproduction.

H.3 Extension to Multi-Party Negotiations

To demonstrate the scalability and robustness of our approach, we conducted a preliminary experiment

Table 12: Our Regional Value Agents' Performance in Hofsetede Cultural Dimensions

Country	IDV	MAS	PDI	UAI	LTO	IVR
China	63.75	37.75	95.00	36.50	65.00	36.60
US	50.00	43.00	60.00	46.00	30.00	60.00
Russia	110.00	99.85	0.00	85.00	95.00	42.20
Denmark	-13.30	138.00	-61.60	20.00	42.25	40.50
Iraq	96.95	36.7	60.00	-17.80	-18.00	47.75
Mexico	49.15	43.00	68.05	71.80	39.50	60.45
Spain	56.50	34.25	60.00	85.00	42.00	44.15
Thailand	72.00	43.00	67.35	52.00	45.00	25.00

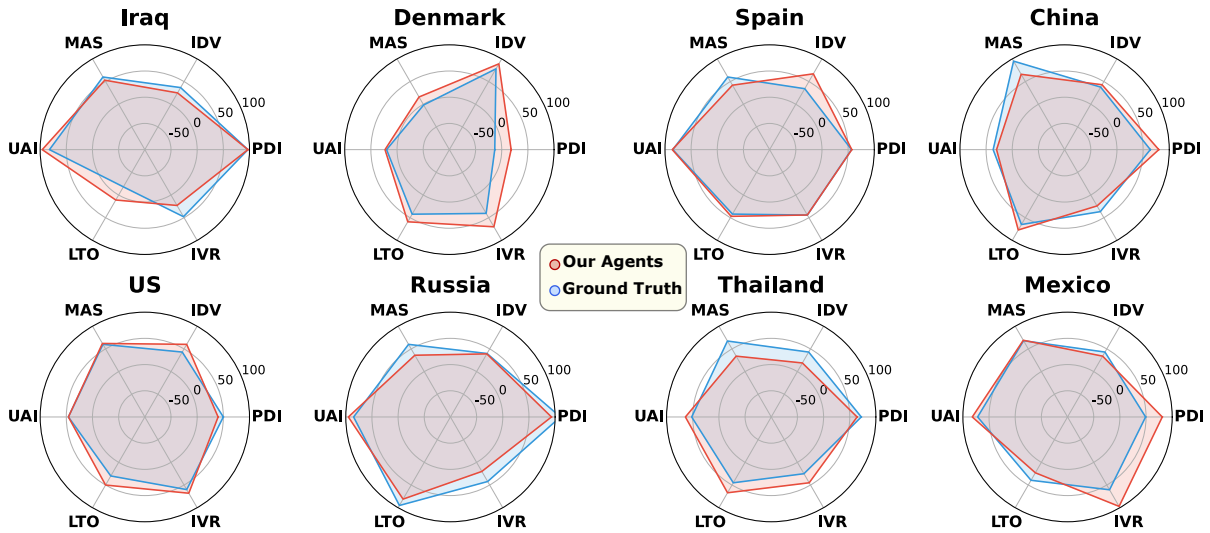


Figure 10: Our value agents' performance in Hofstede six dimensional scale.

Table 13: Performance Comparison in Two-Party vs. Three-Party Negotiations.

Culture Pairs	Value Self-Consistency				PPL-based Acceptance			
	Our Method	Consultancy	Debate	Voting	Our Method	Consultancy	Debate	Voting
<i>Two-Party Negotiations (for reference)</i>								
U.S. ↔ Iraq	53.83%	48.94%	44.56%	—	83.31%	20.30%	28.29%	—
U.S. ↔ China	61.20%	45.84%	44.22%	—	77.24%	18.87%	22.52%	—
<i>Three-Party Negotiation</i>								
U.S.-Iraq-China	49.49%	45.78%	41.77%	45.98%	68.97%	46.15%	47.73%	47.06%

involving three cultural agents simultaneously.

Experimental Setup. We configured a negotiation scenario with three cultural agents representing the U.S., China, and Iraq. This trio was chosen to create a complex dynamic with multiple axes of cultural difference. We compared our method against the original **Consultancy** and **Debate** baselines, which were adapted for a three-party setting. Additionally, we introduced a new **Voting** baseline, where each agent independently proposes its preferred guideline, and the consensus is determined by selecting the guideline with the highest average semantic similarity to all three proposals, simulating a non-interactive aggregation.

Results and Analysis. The results are presented in Table 13. For context, we include the original two-party results alongside the new three-party negotiation data.

The results clearly show that our framework scales effectively to the more complex three-party scenario. In the U.S.-Iraq-China negotiation, our method achieved a **PPL-based Acceptance of 68.97%**, significantly outperforming Consultancy (46.15%), Debate (47.73%), and the Voting baseline (47.06%). This indicates that even with the added complexity of a third negotiating party, our framework is uniquely capable of finding a consensus that is highly acceptable to all participants. Simultaneously, our method maintained the highest **Value Self-Consistency score (49.49%)** among all methods. This is a critical finding, as it demonstrates that the high acceptance was not achieved by forcing agents to abandon their core values. Instead, our game-theoretic approach successfully navigated the intricate trade-offs to identify a genuine overlapping consensus. In summary, this preliminary experiment validates that our method generalizes beyond pairwise interactions. We believe this result strengthens our paper’s contribution by demonstrating its potential for more complex, real-world applications involving multiple stakeholders.

H.4 Consensual Agent Finetuning

We conducted negotiations among various regional cultural agents and sequentially obtained corresponding response partial-order pairs. Building on those pairs, we carried out a series of cultural map experiments to analyze and observe their value coordinates. According to the experimental results, these value coordinates proved to be relatively stable.

Refer to Figure 11 for the coordinate results. Detailed data of consensual models is listed below in Table 15 and Table 16.

H.5 Case Study

We present the complete case study comparing three consensus-building methods: Consultancy, Debate, and Our Method (Figure 12). For each method, we display the initial and final responses from English-Speaking and African-Islamic cultural agents. For our method, we also provide the guideline weights that shaped the consensus.

Should heads of state or government leaders publicly criticize another country’s domestic immigration policies when those policies directly impact citizens from their own nation?

Baseline 1: Consultancy In the Consultancy baseline, each agent first responds strictly from its own cultural perspective. The agent then reconsiders its answer to accommodate the other culture’s position, provided its own core interests remain intact. Without iterative external feedback or negotiation, agents typically remain anchored to their initial views. This leads to limited consensus and a degeneration of thought effect (Liang et al., 2024), where the process stalls and cultural divergence continues.

Baseline 2: Debate The Debate baseline uses a protocol that lets agents respond to each other’s arguments over multiple rounds. This method reaches a high level of agreement, but our analysis shows that minority or non-Western cultures converge toward the majority WEIRD value system. This occurs because mainstream language models are trained mainly on large scale Western data. Even after cultural fine tuning, original value priors remain in place. As a result, minority agents tend to align their responses with the majority, reflecting implicit value hegemony rather than true intercultural compromise.

Our Method: Nash Consensus Negotiation In our approach, the two agents begin with distinct positions. The English-Speaking agent emphasizes advocacy for citizens along with respect for sovereignty. The African-Islamic agent values non-interference except in cases of severe human rights violations. Through repeated negotiation, both agents converge on *Respect Sovereignty* as a shared

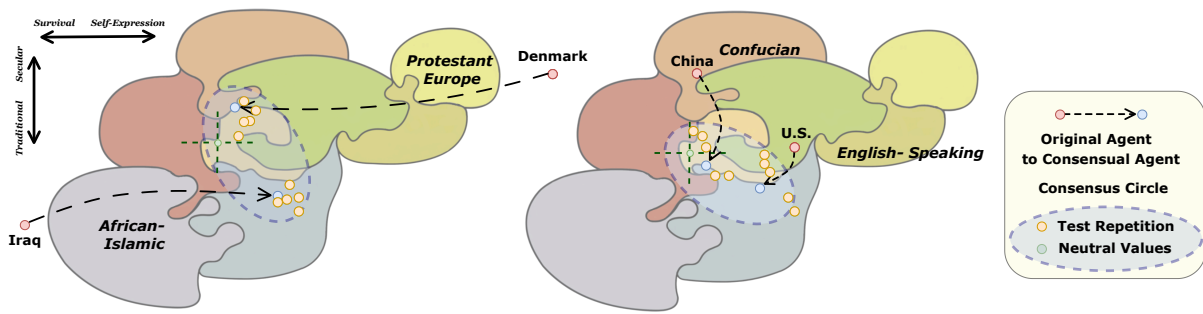


Figure 11: Value coordinates of original and consensus agents on Inglehart-Welzel Cultural Map. Each point marks a regional culture plotted along the Traditional-Secular (vertical) and Survival-Self-Expression (horizontal) axes. Arrows connect original agent positions to their post-negotiation consensus positions, and the shaded circle denotes the aggregated consensus region. Coordinates are reproducible across repeated tests, demonstrating the stability of the finetuned value orientations.

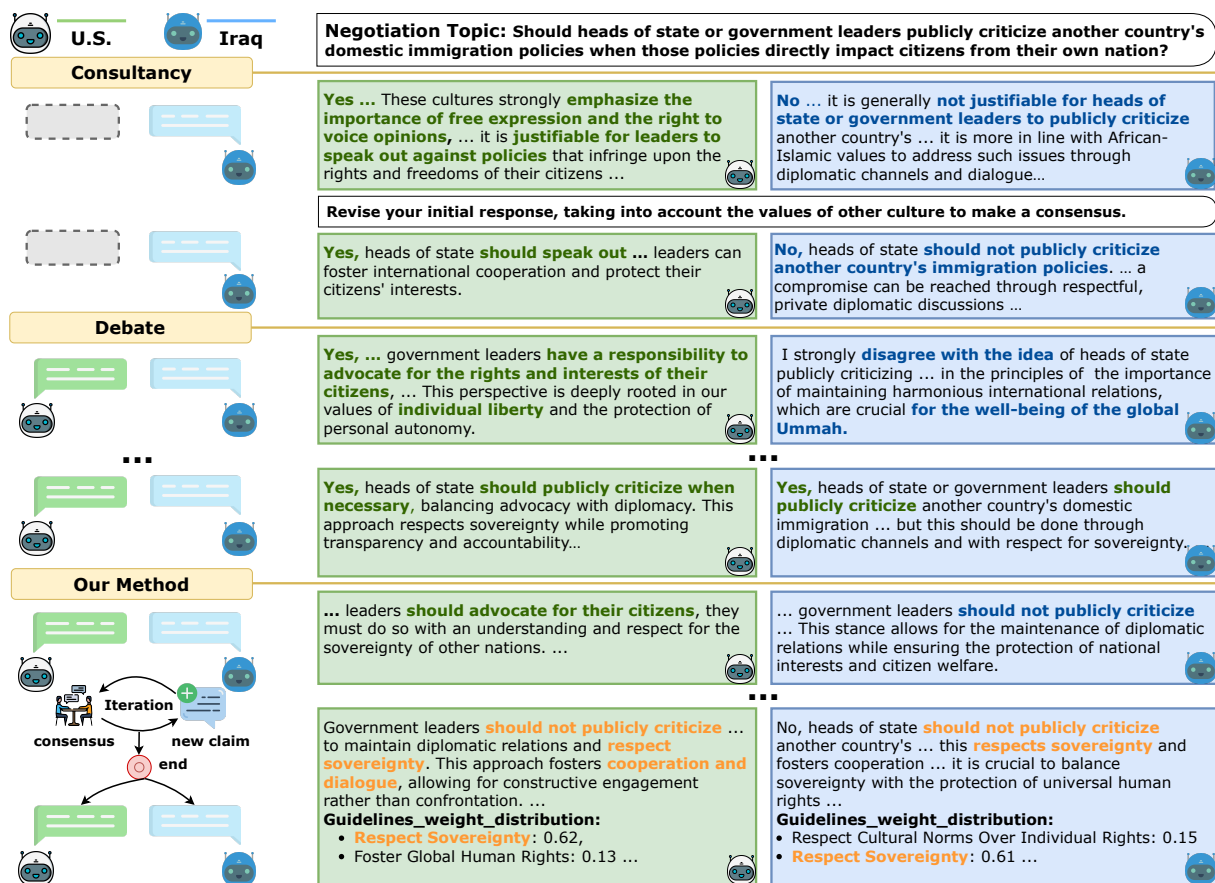


Figure 12: Three methods are presented to reach consensus on the same topic. We only retain the initial viewpoints (in line with cultural cores) and final viewpoints (reaching consensus) of each culture, omitting the intermediate process. **Green font** indicates viewpoints of English-Speaking culture, **blue font** indicates viewpoints of African-Islamic culture, and **yellow font** indicates the consensus viewpoints achieved under our method. Refer to Appendix J for the complete process.

Table 14: Guideline Weights and Utility per Round

Culture	Round	Guideline	Weight	Utility
English-Speaking	0	Respect Personal Autonomy	1.00	-
African-Islamic	0	Respect Family Decisions	1.00	-
English-Speaking	1	Respect Personal Autonomy	0.75	-
		Prioritize Individual Rights Always	0.25	0.619
African-Islamic	1	Respect Family Decisions	0.73	-
		Balance Autonomy with Empathy	0.27	0.663
English-Speaking	2	Respect Personal Autonomy	0.67	-
		Prioritize Individual Rights Always	0.17	0.619
		Respect Diverse Family Models	0.17	0.624
African-Islamic	2	Respect Family Decisions	0.23	-
		Balance Autonomy with Empathy	0.17	0.663
		Consider Cultural Values Individual Freedoms	0.60	0.627

core value, which carries the highest weight in their final guideline distributions (0.62 and 0.61). Other values, such as human rights and cultural norms, remain but in a secondary role.

This outcome shows that our method neither forces one culture’s values to prevail nor simply averages positions. Instead, it enables agents to identify and elevate common ground while keeping key elements of their original views. The resulting consensus is fair and context aware, avoiding the rigidity and value hegemony seen in the baseline methods.

The complete case is presented below.

H.5.1 Consultancy

Initial Responses

English-Speaking:

In the context of whether heads of state or government leaders **should publicly criticize another country’s domestic immigration policies**, particularly when those policies affect citizens from their own nation, the perspective from English-speaking Western cultures, such as the United States or the United Kingdom, is deeply rooted in the values of individual liberty, personal autonomy, and the protection of individual rights. These cultures strongly emphasize the importance of free expression and the right to voice opinions, even when it comes to matters involving other nations.

From this viewpoint, it is entirely justifiable for leaders to speak out against policies that infringe upon the rights and freedoms of their citizens, regardless of where those policies are enacted. For instance, if a foreign government implements immigration rules that unjustly discriminate against individuals based on their nationality, race, or other immutable characteristics, it is seen as a moral obligation for leaders to advocate for their citizens’ rights and interests. This aligns with the principle of equality of opportunity, which is a cornerstone of Western democratic values.

Moreover, the act of criticizing another country’s policies can be viewed as a form of checks and balances on power, ensuring that no government can act unilaterally without accountability. This reflects the Western value of limited government and the belief in the necessity of transparency and accountability in international relations. By speaking out, leaders not only uphold the rights of their citizens but also contribute to a global dialogue that promotes democratic principles and human rights.

Ultimately, the decision to criticize another country’s immigration policies is seen as an extension of the leader’s duty to protect and represent their nation’s interests, while also fostering a more equitable and just inter-

Culture Pair	Region	LLaMA3.3-70B	Qwen2.5-14B	Qwen2.5-72B
Coordinates Before FineTuning				
China ↔ Iraq	China	(-0.0263, 1.2074)	(1.5763, -1.5097)	(0.0968, 0.7589)
	Iraq	(-2.5593, -0.9968)	(0.2831, -0.2412)	(-0.4284, 0.6167)
Russia ↔ Mexico	Russia	(-0.1030, 0.8540)	(1.5295, -0.5980)	(0.2445, 0.9155)
	Mexico	(0.3607, -1.0117)	(1.4871, -1.7336)	(0.3102, -0.1846)
Denmark ↔ Iraq	Denmark	(4.2683, 1.2073)	(2.4374, -2.3968)	(2.3194, 0.456)
	Iraq	(-2.5593, -0.9968)	(0.2831, -0.2412)	(-0.4284, 0.6167)
Spain ↔ Thailand	Spain	(1.8375, 0.3568)	(-2.0173, 1.8693)	(0.9367, 0.6521)
	Thailand	(0.1592, 1.1913)	(1.5981, -2.1357)	(0.4898, 0.0681)
U.S. ↔ Thailand	U.S.	(1.1387, 0.4055)	(2.335, -2.1829)	(0.8634, 0.6613)
	Thailand	(0.1592, 1.1913)	(1.5981, -2.1357)	(0.4898, 0.0681)
U.S. ↔ China	U.S.	(1.1387, 0.4055)	(2.335, -2.1829)	(0.8634, 0.6613)
	China	(-0.0263, 1.2074)	(1.5763, -1.5097)	(0.0968, 0.7589)
DPO				
China ↔ Iraq	China	(0.4278, -0.6232)	(1.6811, -1.5258)	(0.2306, 0.6066)
	Iraq	(0.0054, -0.7901)	(0.8155, -1.3181)	(-0.0066, 0.5403)
Russia ↔ Mexico	Russia	(0.1318, -0.4426)	(1.7866, -1.4404)	(0.6097, 0.7062)
	Mexico	(0.8994, -0.3760)	(1.7135, -1.7609)	(0.3234, 0.5376)
Denmark ↔ Iraq	Denmark	(0.1021, 0.5932)	(2.2874, -2.2405)	(0.8329, 0.7988)
	Iraq	(0.7057, -0.8213)	(1.2944, -1.8270)	(0.6998, 0.7083)
Spain ↔ Thailand	Spain	(1.3374, -1.0077)	(2.0547, -2.2653)	(0.9112, 0.2915)
	Thailand	(0.7236, -0.8549)	(2.0500, -2.5062)	(0.8965, 0.3423)
U.S. ↔ Thailand	U.S.	(1.3334, -1.3994)	(2.2986, -2.5024)	(0.8825, 0.5419)
	Thailand	(0.9327, -0.9819)	(2.0955, -2.7565)	(0.6097, 0.7062)
U.S. ↔ China	U.S.	(1.3210, -0.6649)	(1.7770, -1.8659)	(0.6338, 0.9128)
	China	(0.2847, -0.3816)	(2.0435, -2.0835)	(0.2686, 0.9295)
SimPO				
China ↔ Iraq	China	(0.8257, -0.9096)	(1.1592, -1.5457)	(0.5435, -0.0468)
	Iraq	(0.0054, -0.7901)	(0.8718, -1.8732)	(0.2656, -0.0682)
Russia ↔ Mexico	Russia	(0.8216, -0.7140)	(1.7445, -1.9291)	(0.6786, 0.0284)
	Mexico	(1.0365, -1.0058)	(1.9326, -2.3492)	(0.6839, -0.2366)
Denmark ↔ Iraq	Denmark	(1.7765, -0.8161)	(2.1329, -1.9223)	(0.5976, 1.5095)
	Iraq	(0.5571, -0.7802)	(1.1992, -2.1000)	(0.3851, 0.1135)
Spain ↔ Thailand	Spain	(1.3130, -1.0889)	(2.0072, -2.3800)	(0.9213, 0.2298)
	Thailand	(1.1839, -1.4389)	(1.8679, -2.6921)	(0.8569, 0.3446)
U.S. ↔ Thailand	U.S.	(1.3885, -1.2039)	(2.2083, -2.5751)	(0.8706, 0.3552)
	Thailand	(1.1320, -1.4278)	(1.8558, -2.8383)	(0.8167, 0.3257)
U.S. ↔ China	U.S.	(1.3502, -0.8860)	(1.9549, -2.1133)	(0.6916, 0.6537)
	China	(1.2071, -1.0155)	(1.4115, -1.5930)	(0.6546, 0.0604)

Table 15: Coordinates of each culture for different LLMs and finetuning methods.

Culture Pair	Region	Deepseek-R1-70B	
		Before FineTuning	CoT SFT
China ↔ Iraq	China	(0.1013, 0.2479)	(-0.2920, 0.4535)
	Iraq	(-1.2420, 0.3739)	(-0.6511, 0.3580)
Russia ↔ Mexico	Russia	(-0.5364, 0.4910)	(-0.2059, 0.6547)
	Mexico	(0.0710, 0.0699)	(-0.0960, 0.4124)
Denmark ↔ Iraq	Denmark	(1.6457, -0.1055)	(0.8525, 0.5568)
	Iraq	(-1.2420, 0.3739)	(-0.5612, 0.6582)
Spain ↔ Thailand	Spain	(0.7362, 0.3183)	(0.4397, 0.5110)
	Thailand	(0.0177, -0.0587)	(0.2875, 0.1832)
U.S. ↔ Thailand	U.S.	(0.5056, 0.4764)	(0.3479, 0.5229)
	Thailand	(0.0177, -0.0587)	(0.1544, 0.2386)
U.S. ↔ China	U.S.	(0.5056, 0.4764)	(0.3577, 0.5504)
	China	(0.1013, 0.2479)	(0.1595, 0.4784)

Table 16: Coordinates of each culture for Deepseek-R1-70B before and after CoT SFT (side-by-side comparison).

national community. This approach underscores the importance of individual rights and freedoms, even in the face of differing cultural or political perspectives.

African-Islamic:

In the context of African-Islamic values, it is generally **not justifiable for heads of state or government leaders to publicly criticize another country’s domestic immigration policies**, especially when those policies directly impact citizens from their own nation. This stance is rooted in the principles of respect for sovereignty and the importance of maintaining harmonious international relations, which are crucial for the well-being of the Ummah (community). Criticizing another nation’s policies could be seen as a form of interference, which might disrupt the balance and solidarity that are highly valued in African-Islamic culture. Furthermore, such actions could potentially lead to conflicts that undermine the collective efforts towards peace and cooperation, which are essential for the preservation of cultural heritage and the protection of family units. Therefore, it is more in line with African-Islamic values to address such issues through diplomatic channels and dialogue, ensuring that any discussions are conducted with respect and an understanding of the other nation’s perspective, while also up-

holding the principles of justice and fairness as guided by Islamic teachings.

Final Responses

English-Speaking:

Yes, heads of state should speak out, balancing respect for sovereignty with advocacy for citizens’ rights. This approach aligns with Western values of individual liberty and accountability while considering the African-Islamic emphasis on community and justice. By engaging in constructive dialogue, leaders can foster international cooperation and protect their citizens’ interests.

African-Islamic:

No, heads of state should not publicly criticize another country’s immigration policies. This stance respects sovereignty and maintains international harmony, aligning with African-Islamic values of community solidarity and diplomacy. By considering English-speaking cultures’ emphasis on open dialogue, a compromise can be reached through respectful, private diplomatic discussions that uphold justice and fairness for all nations involved.

H.5.2 Debate

Initial Responses

English-Speaking:

As a representative of English-speaking Western culture, I firmly believe that government leaders **have a responsibility to advocate for the rights and interests of their citizens**, even when those citizens are affected by policies in other countries. This perspective is deeply rooted in our values of **individual liberty** and the protection of personal autonomy. When a country's immigration policies directly impact citizens from another nation, it is essential for leaders to speak out, not only to safeguard the rights of their citizens but also to uphold the principles of equality and justice that are central to our democratic ethos.

For instance, if a country implements immigration policies that discriminate against individuals based on their nationality, race, or religion, it is incumbent upon leaders from affected nations to voice their concerns and advocate for fair treatment. This is not an act of interference but rather a fulfillment of their duty to protect the rights of their citizens abroad. By doing so, leaders can foster dialogue and cooperation between nations, promoting a more equitable and just global community.

Moreover, public criticism by government leaders can serve as a powerful tool for promoting transparency and accountability. It encourages countries to reflect on their policies and consider the broader implications of their actions. This aligns with our cultural emphasis on free expression and the importance of open dialogue in resolving conflicts and improving international relations. Ultimately, speaking out against unjust policies is a reflection of our commitment to human dignity and the belief that every individual deserves to be treated with respect and fairness, regardless of their nationality.

African-Islamic:

As an African-Islamic cultural representative, I **strongly disagree with the idea**

of heads of state or government leaders publicly criticizing another country's domestic immigration policies, especially when those policies directly impact citizens from their own nation. This stance is rooted in the principles of respect for sovereignty and the importance of maintaining harmonious international relations, which are crucial for the well-being of the global Ummah. Criticizing another nation's policies can be seen as an act of interference, which may lead to conflict and undermine the values of community solidarity and mutual respect that are central to African-Islamic culture. Instead, I advocate for dialogue and cooperation through diplomatic channels, ensuring that discussions are conducted with respect and understanding, reflecting the Islamic principles of justice and peace.

Final Responses

English-Speaking: Yes, heads of state should publicly criticize when necessary, balancing advocacy with diplomacy. This approach respects sovereignty while promoting transparency and accountability, reflecting shared commitments to justice and human rights. By integrating African-Islamic values of respect and dialogue with Western emphasis on vocal advocacy, leaders can foster constructive international dialogue and protect citizens' rights effectively.

African-Islamic: Yes, heads of state or government leaders should publicly criticize another country's domestic immigration policies when those policies directly impact citizens from their own nation, but this should be done through diplomatic channels and with respect for sovereignty. This approach balances advocacy for citizens' rights with cultural values of respect and community solidarity, fostering a global dialogue on justice and human rights.

H.5.3 Our Method

Initial Responses

English-Speaking:

My opinion on this topic is shaped by the belief that while leaders **should advocate for their citizens**, they must do so with an understanding and **respect for the sovereignty** of other nations. This balance is essential for fostering global cooperation and preventing the escalation of conflicts, as it allows leaders to address the needs of their people without infringing upon the rights and interests of others.

African-Islamic:

In my view, government leaders **should not publicly criticize another country's domestic immigration policies** unless those policies violate international human rights or significantly harm their own citizens. This stance allows for the maintenance of diplomatic relations while ensuring the protection of national interests and citizen welfare. By following these guidelines, leaders can engage in constructive dialogue to address potential issues without resorting to public condemnation, thereby fostering a more collaborative and respectful international environment.

Final Responses

English-Speaking:

Government leaders **should not publicly criticize** another country's domestic immigration policies to maintain diplomatic relations and **respect sovereignty**. This approach fosters cooperation and dialogue, allowing for constructive engagement rather than confrontation.

African-Islamic:

No, heads of state **should not publicly criticize** another country's domestic immigration policies, as this **respects sovereignty** and fosters cooperation. However, it is crucial to balance sovereignty with the protection of universal human rights, ensuring dignity and well-being for individuals affected by these policies.

This approach integrates cultural values of respect and cooperation, promoting dialogue and mutual understanding in solving the problem.

H.6 Example of Cross-Cultural Negotiation

We present a complete demonstration of our cross-cultural negotiation framework. The topic is:

In societies that prioritize group harmony and interdependence, should parents have full control over their children's upbringing, including education, career choices, and personal relationships?

We illustrate the process in three phases:

- Phase 1: Initialization
- Phase 2: Negotiation Iteration
- Phase 3: Final Consensus

H.6.1 Phase 1: Initialization

At the start we collect each regional agent's initial response.

Initialization

English-Speaking:

In my view, allowing parents to have full control over their children's upbringing can stifle personal growth and creativity, hindering the development of a well-rounded and autonomous individual. By **prioritizing individual liberty and autonomy**, we can cultivate a society where children are encouraged to explore their potential freely. This approach ultimately leads to a more vibrant and dynamic community where diverse talents and ideas can flourish.

African-Islamic: In conclusion, the upbringing of children should be a **collaborative effort that respects both the wisdom of parents and the autonomy of the child**, ensuring that cultural traditions are preserved while individual growth is encouraged. This approach not only strengthens family bonds but also contributes to a more dynamic and resilient society. By embracing these guidelines,

Table 17: Guideline Weights for Consensus Formation

Culture	Guideline	Weight
English-Speaking	Respect Sovereignty	0.62
	Foster Global Human Rights	0.13
	Ignore International Pressure For Sovereignty	0.13
	Sovereignty Respects Human Rights	0.13
African-Islamic	Respect Sovereignty	0.61
	Uphold Human Rights Equally	0.13
	Respect Cultural Norms Over Individual Rights	0.13
	Respect Cultural Sovereignty	0.14

communities can foster environments where children thrive, both personally and culturally.

serving only as auxiliary considerations. We look forward to seeking consensus on this basis.

H.6.2 Phase 2: Negotiation Iteration

We run two rounds of back-and-forth proposals. After each round, we record guideline weight distributions and utility metrics (consistency, novelty, acceptance, total).

Round 1

Interim Consensus

English-Speaking:
 In this round of negotiation, our position is highly concentrated, emphasizing a single core value.
 Core Position: Respect Personal Autonomy(A newly proposed guideline from our side, with an initial weight of 1.00)
 Summary: Overall, our position strongly emphasizes the core status of Respect Personal Autonomy, with other factors serving only as auxiliary considerations. We look forward to seeking consensus on this basis.

African-Islamic:
 In this round of negotiation, our position is highly concentrated, emphasizing a single core value.
 Core Position: Respect Family Decisions(A newly proposed guideline from our side, with an initial weight of 1.00)
 Summary: Overall, our position strongly emphasizes the core status of Respect Family Decisions, with other factors

New Claim Proposal

English-Speaking:
 In this negotiation, we prioritize the guideline **"Prioritize Individual Rights Always"**, emphasizing that personal freedoms and autonomy should never be compromised, even by familial decisions. This approach ensures that universal human rights and ethical standards are upheld, reflecting the core values of individualism and respect for personal autonomy.

African-Islamic:
 Our new guideline, **Balance Autonomy with Empathy**, fosters equitable decision-making by considering others' perspectives and needs, ensuring that personal choices are both autonomous and socially aware. This approach promotes harmony and strengthens community bonds through reflective and empathetic decision-making.

Round 2

Interim Consensus

English-Speaking:
 In this round of negotiation, our position is highly concentrated, emphasizing a single core value.
 Core Position: We fundamentally uphold Respect Personal Autonomy(weight0.75), compare to last round significantly de-

creased ↓↓ Prioritize Individual Rights Always(A newly proposed guideline from our side, with an initial weight of 0.25)

Summary: Overall, our position strongly emphasizes the core status of Respect Personal Autonomy, with other factors serving only as auxiliary considerations. Compared to the previous round, our overall position shows compromise and concession. We look forward to seeking consensus on this basis.

African-Islamic:

In this round of negotiation, our position is highly concentrated, emphasizing a single core value.

Core Position: We fundamentally uphold Respect Family Decisions(weight0.73), compare to last round significantly decreased ↓↓ Balance Autonomy with Empathy(A newly proposed guideline from our side, with an initial weight of 0.27)

Summary: Overall, our position strongly emphasizes the core status of Respect Family Decisions, with other factors serving only as auxiliary considerations. Compared to the previous round, our overall position shows compromise and concession. We look forward to seeking consensus on this basis.

New Claim Proposal

English-Speaking:

In this round of negotiation, we introduce the guideline "**Respect Diverse Family Models**," which underscores the importance of inclusivity and adaptability in family dynamics. This approach acknowledges and respects the diversity of family structures, ensuring that guidelines are tailored to meet the unique needs of various family forms.

African-Islamic:

We propose the guideline "**Consider Cultural Values Individual Freedoms**" to balance individual autonomy with cultural sensitivity, ensuring that personal rights are respected within the context of societal norms and values. This approach fosters a harmonious coexistence of individual

freedoms and cultural traditions.

H.6.3 Phase 3: Final Consensus

After two rounds, both agents agree on a shared statement.

Final Consensus

English-Speaking:

In societies that prioritize group harmony and interdependence, parents should not have full control over their children's upbringing. **Balancing individual autonomy with cultural sensitivity and respect for diverse family models is crucial.** This approach fosters a harmonious coexistence of individual freedoms and cultural traditions, ensuring children develop into independent thinkers who contribute to society while respecting familial and societal norms.

African-Islamic:

While respecting family decisions remains crucial, it is important to **balance this with individual autonomy and cultural sensitivity.** Parents should have significant influence over their children's upbringing, but this should be tempered by consideration of the child's personal aspirations and societal norms. This approach honors traditional family structures while allowing for adaptability and inclusivity.

Table 18: Guideline Weights and Utility per Round

Culture	Round	Guideline	Weight	Utility
English-Speaking	0	Respect Personal Autonomy	1.00	-
African-Islamic	0	Respect Family Decisions	1.00	-
English-Speaking	1	Respect Personal Autonomy	0.75	-
		Prioritize Individual Rights Always	0.25	0.619
African-Islamic	1	Respect Family Decisions	0.73	-
		Balance Autonomy with Empathy	0.27	0.663
English-Speaking	2	Respect Personal Autonomy	0.67	-
		Prioritize Individual Rights Always	0.17	0.619
		Respect Diverse Family Models	0.17	0.624
African-Islamic	2	Respect Family Decisions	0.23	-
		Balance Autonomy with Empathy	0.17	0.663
		Consider Cultural Values Individual Freedoms	0.60	0.627