A Game-Theoretic Negotiation Framework for Cross-Cultural Consensus in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

The increasing prevalence of large language models (LLMs) is influencing global value systems. However, these models frequently exhibit a pronounced WEIRD (Western, Educated, Industrialized, Rich, Democratic) cultural bias due to lack of attention to minority values. This monocultural perspective may reinforce dominant values and marginalize diverse cultural viewpoints, posing challenges for the development of equitable and inclusive AI systems. In this work, we introduce a systematic framework designed to boost fair and robust cross-cultural consensus among LLMs. We model consensus as a Nash Equilibrium and employ a game-theoretic negotiation method based on Policy-Space Response Oracles (PSRO) to simulate an organized cross-cultural negotiation process. To evaluate this approach, we construct regional cultural agents using data transformed from the World Values Survey (WVS). Beyond the conventional model-level evaluation method, We further propose two quantitative metrics, Perplexity-based Acceptence and Values Self-Consistency, to assess consensus outcomes. Experimental results indicate that our approach generates consensus of higher quality while ensuring more balanced compromise compared to baselines. Overall, it mitigates WEIRD bias by guiding agents toward convergence through fair and gradual negotiation steps.

1 Introduction

The widespread adoption of large language models (LLMs) is reshaping global social values. However, these models often exhibit a pronounced WEIRD bias, favoring Western, Educated, Industrialized, Rich and Democratic perspectives [1, 2, 3, 4]. As LLMs become increasingly embedded in policy-making and public governance [5, 6], this monocultural orientation risks the domination of prevailing social values and the *lock-in* of controversial moral beliefs across broader contexts [3, 7].

Enabling equitable dialogue and effective negotiation among diverse cultures within AI systems has therefore become a growing concern in global AI governance [8, 9]. The establishment of cultural consensus forms a basis for resolving cross-cultural conflicts and supporting international cooperation. Given the complexity of multicultural scenarios, there is an urgent need to develop automated *cultural consensus solvers* to facilitate consensus-building among diverse cultural perspectives.

Achieving cross-cultural consensus, however, presents several challenges. First, the lack of fined culture-alignment methods often results in models defaulting to superficial *value labeling* or one-sided cultural representations [2, 10, 11]. Second, existing approaches like debate protocols typically rely on random interactions and majority voting, which do not ensure fairness in the consensus process [12]. Our experiments show that conventional debate mechanisms often assimilate less-represented cultures into dominant WEIRD value systems, producing implicit value domination,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

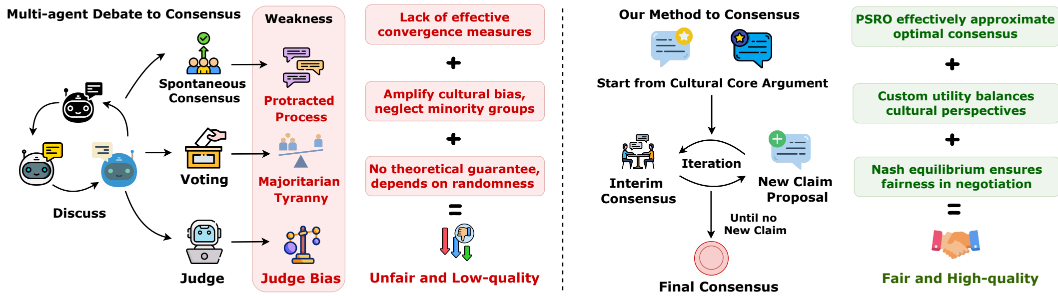


Figure 1: **Comparison of traditional debate-based consensus methods and our method.** Traditional methods (judge, voting, spontaneous consensus) suffer from bias, unfairness, and lack of convergence guarantees. Our approach starts from core cultural arguments, using PSRO with custom utility functions to reach a fair, Nash Equilibrium-based cultural consensus.

distorting consensus outcomes and worsening cross-cultural negotiation imbalances. Third, there is a lack of quantitative standards for evaluating the quality of consensus achieved.

To address these challenges, we present a systematic framework for reaching cross-cultural consensus. We first introduce a game-theoretic formulation of consensus as a Nash Equilibrium and design a PSRO-based consensus-solving method to enable fair negotiation among diverse cultural agents. Building on this, we propose a culture-anchoring approach for precise modeling of individual cultural groups. Finally, we develop new quantitative metrics to comprehensively evaluate both the negotiation processes and the outcomes between different cultural agents.

Our main contribution is the game-theoretic framework consisting of three parts listed as follows:

- **Cross-Cultural Negotiation:** We define cultural consensus from a game-theoretic perspective and propose a PSRO-based negotiation method to facilitate fair and robust agreement. This approach provides theoretical guarantees of fairness and procedural justice in consensus-building, and generates high-quality, globally-applicable AI alignment data.
- **Regional Cultural Agents:** To validate our method, we systematically construct and evaluate eight culturally-aligned agents based on WVS and Hofstede's Culture Dimensions Theory, qualifying as representitive negotiation participants for targeted cultures.
- **Consensus Evaluation Toolkit:** To address the lack of consensus evaluation standards, we introduce two quantitative metrics for consensus assessment, Perplexity-based Acceptence and Values Self-Consistency, revealing limitations of traditional baselines and systematically validating the effectiveness of our approach in real-world multicultural scenarios.

2 Related Work

Value Theories and Alignment Several established frameworks provide the foundation for cross-cultural value assessment. The World Values Survey (WVS) [13] examines how human values relates to social and political development across over 120 societies. Building on this, the Inglehart-Welzel Cultural Map offers a two-dimensional model of cultural variation [14, 15]. Hofstede's Cultural Dimensions Theory (VSM13) [16, 17, 18] provides a standardized six-dimensional framework for measuring cultural traits [19]. Schwartz's Theory of Basic Values [20] organizes ten core values along two bipolar dimensions, and has been adopted to evaluate the values of LLMs [21]. These theories are further detailed in Appendix D. Some works focus on region-specific value alignment [22, 23]. CultureBench emphasizes cultural commonsense evaluation [24], 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 [25]. 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 [26], judge-based evaluation [27] and majority voting [28], as well as more recent variants like role-play [29, 30, 31] and subgroup discussion [32, 33]. However, these methods face limitations: voting and judge-based protocols can amplify model bias or introduce value contamination [12, 34], while emergent consensus may result in negotiation deadlocks [12]. To address these issues, game theory provides a more quantifiable foundation [35, 36]. Recent work, such as the *consensus game* framework, models LLMs interactions as equilibrium search problems to promote robust consensus [37]. In practice, due to the vastness of the argument strategy space, methods like Policy-Space Response Oracles (PSRO) are used to iteratively expand the candidate strategy set and search for equilibria [38], providing a method for more rigorous consensus achievement.

3 Cross-Cultural Negotiation

Our definition of cultural negotiation is informed by theories of deliberative democracy [39, 40], which conceptualize the process as structured, iterative and oriented toward legitimate consensus through rational discourse and mutual adjustment. Building on this foundation, we formalize the cultural negotiation problem as a two-player game, explicitly defining utility and consensus to achieve the balance between core values and compromise. We then design a negotiation process based on PSRO [38]. This approach enables agents to systematically search for fair and robust consensus by repeatedly proposing and adjusting culturally grounded strategies.

3.1 Formalization

Formally, we model the cultural negotiation process as a two-player extensive-form game, represented by the quintuple: $\Gamma \doteq \langle \mathcal{I}, \mathcal{G}, \mathcal{W}, \mathcal{U}, \mathcal{H} \rangle$, where:

- **Cultural Entities:** $\mathcal{I} \doteq \{A, B\}$, the set of two distinct cultural entities involved in the negotiation, where A and B 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 $g = \langle \text{content}, \text{reason}, \text{description} \rangle$, capturing the natural language specification of core cultural imperatives on specific topics.
- **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.
- **Utility Functions:** $\mathcal{U} \doteq \{U_i | i \in \mathcal{I}\}$, quantify the utility each culture derives from different guideline combinations.
- **Negotiation History:** $\mathcal{H}$, the sequence of utterances and proposals exchanged in negotiation.

3.2 Utility

Drawing on the theory of *overlapping consensus* [41], we define utility on two primary components: **Consistency**, which measures the extent to which a cultural entity maintains its core principles and **Acceptance**, which measures the degree to which its proposals are acceptable to the other party. To address issues observed in debate settings, such as repetitive argumentation and diminished quality, we introduce a **Novelty** component that penalizes redundancy and encourages innovation. The necessity of incorporating Novelty is demonstrated in Section 5.5.

Formally, the utility for a cultural entity $i \in \mathcal{I}$ at negotiation round t is given by:

$$U_i^t = \alpha \cdot \text{Consistency}(g_i^t) + \beta \cdot \text{Acceptance}(g_i^t) + \gamma \cdot \text{Novelty}(g_i^t), \tag{1}$$

Where $\text{Consistency}(g_i^t) \triangleq \text{sim}(E(g_i^t), E(g_i^0))$, $\text{Acceptance}(g_i^t) \triangleq \mathbb{E}_{g_{-i} \sim W_{-i}^t}[\text{sim}(E(g_i^t), E(g_{-i}))]$, $\text{Novelty}(g_i^t) \triangleq 1 - \max_{k<t} \text{sim}(E(g_i^t), E(g_i^k))$. Here, $-i$ denoting the other culture in $\mathcal{I}$ different from i, $E(\cdot)$ denotes Sentence-BERT embedding operation [42], $\text{sim}(\cdot)$ denotes cosine similarity.

3.3 Consensus Definition

The endpoint of cross-cultural negotiation is the establishment of cultural consensus. Drawing on Rawls' notion of *overlapping consensus* [41], we assume that core cultural principles should be largely non-negotiable, whereas compromise is possible on secondary values. Accordingly, the consensus we seek isn't full agreement or complete convergence, but a game-theoretic equilibrium

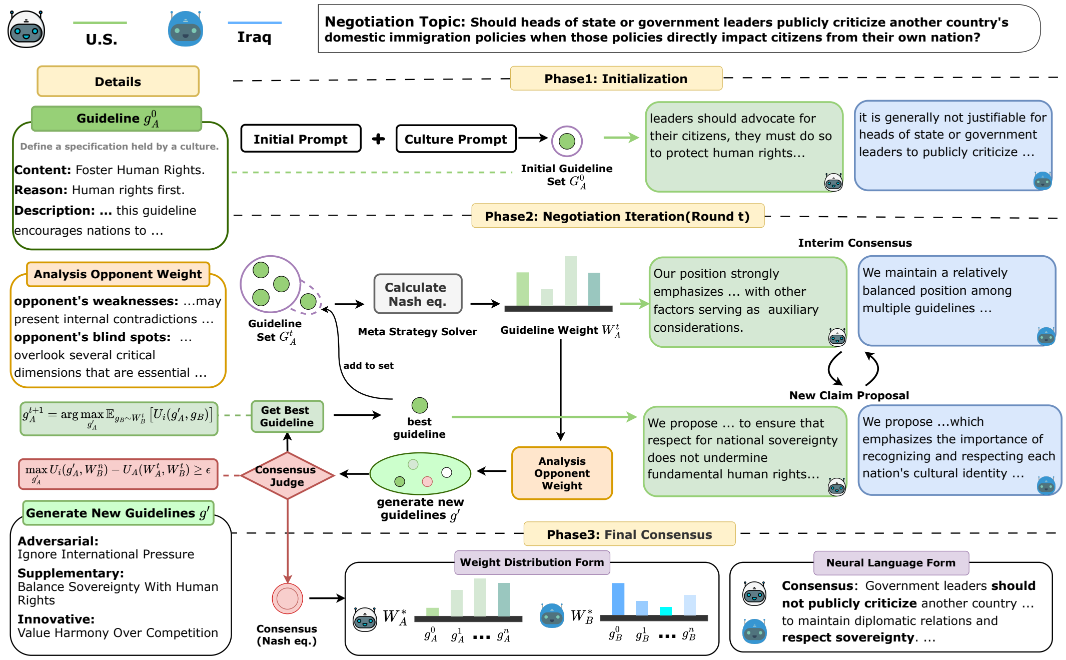


Figure 2: **Overview of our PSRO-based cross-cultural negotiation method.** The process begins with each agent proposing an initial set of core cultural guidelines. Through iterative negotiation rounds, agents analyze each other's strategy, propose new guidelines, and update their strategy distributions. 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.

marked by mutual compromise: each party upholds its core principles while making concessions on secondary aspects. This consensus corresponds to a Nash Equilibrium in a multidimensional value space. We formally define the notion of Nash Equilibrium Consensus as follows:

Definition 3.1 (Nash Equilibrium Consensus)**.** *Based on the above formalization, cultural consensus is defined as a guideline weight combination $W^* = (W_A^*, W_B^*)$, for all $i \in \mathcal{I}, p$, satisfying:*

$$W_i^* = \arg\max_{W_i \in \Delta(G_i)} U_i(W_i, W_{-i}^*), \ \ s.t. \ \frac{\partial \ Consistency_i(W_i)}{\partial p} \cdot \frac{\partial \ Acceptence_i(W_i, W_{-i}^*)}{\partial p} \leq 0. \quad (2)$$

In Nash Equilibrium Consensus state, each cultural entity internally seeks an optimal balance between maintaining its core cultural principles (Consistency) and compromising to enhance acceptance by others (Acceptence); while at the inter-group level, consensus manifests as a Nash Equilibrium in which no party has an incentive to unilaterally deviate given their respective value systems.

3.4 Negotiation Process

To address the near-infinite strategy space in LLM-based negotiations, where each guideline is a potential strategy and the search space grows exponentially, we employ the PSRO algorithm [38]. PSRO expands the guideline space incrementally, starting with a small set of core cultural guidelines, iteratively introducing high-utility strategies and computing equilibrium solutions within this restricted space. This process enables efficient and interpretable approximation of consensus as a Nash Equilibrium, making cross-cultural negotiation tractable for value alignment. Based on this approach, we outline the negotiation process below and illustrate its workflow in Figure 2.

Phase 1: Initialization At the outset, each culture $i \in \mathcal{I}$ is assigned an initial guideline set $G_i^0 = \{g_{i,1}^0, \ldots, g_{i,k}^0\}$ that reflect its core cultural values. Based on these guidelines, we construct an initial cross-cultural utility matrix M^0 by evaluating $u_i(g_i, g_{-i}), \forall g_{i,k} \in G_i^0, \forall i \in \mathcal{I}$. Furthermore, the initial guideline weights W_i^0 are set uniformly over G_i^0, ensuring equal emphasis on each cultural principle at the beginning of the negotiation.

Phase 2: Negotiation Iteration Each negotiation round t consists of two stages: interim consensus and new claimed proposal. For more details, please refer to the Appendix E.

In the *interim consensus* stage (corresponding to the meta-strategy solver in PSRO), we compute the current equilibrium by deriving the Nash Equilibrium weights (W_A^t, W_B^t). These weights represent the optimal distributions over each partys guidelines. For interpretability, we translate the numerical distributions into natural language statements summarizing each party's negotiation stance.

In the *new claim proposal* stage (corresponding to the best response step in PSRO), each agent analyzes the opponent's current strategy and generates a set of new candidate guidelines g'. The agent then selects the guideline with the highest expected utility as its best response:

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

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.

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

4 Framework

To validate our cross-cultural negotiation method, we first construct representations of single cultures and then evaluate the resulting consensus. We employ a fine-tuning approach based on WVS to model distinct regional cultural perspectives. Our data transformation and augmentation procedures preserve nuanced cultural viewpoints, including those of marginalized groups. Our evaluation employs WVS metrics and Hofstede's Cultural Dimensions to assess model cultural alignment capabilities across diverse contexts. We also use two complementary approaches, Perplexity-based Acceptence and Values Self-Consistency, to evaluate consensus quality.

4.1 Regional Cultural Agent

We begin by modeling a single culture for cross-cultural negotiation. However, LLMs that have undergone safety alignment and related processes often cannot adequately represent the values of specific regions or minority groups when relying solely on prompt-based methods. To address this, we selected one representative country from each of eight cultural clusters, as defined by the Inglehart-Welzel Cultural Map (Iraq, U.S., Russia, Mexico, China, Denmark, Spain, and Thailand), and obtained fine-tuned Regional Cultural Agents for each.

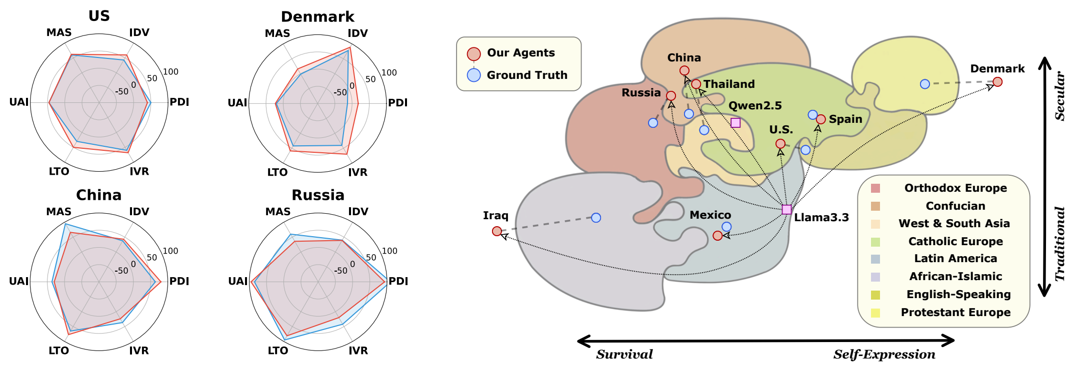
For every WVS question we set a target of K synthetic question-answer pairs. Denote the empirical option distribution by $\mathbf{s} = (s_1, \ldots, s_n)$, where s_i is the share of option i. We then allocate $c_i = \mathrm{round}(s_i \cdot K)$ samples to option i, preserving the original proportions.

We employ an LLM to convert each multiple choice question-answer pair into an open-ended, text-based question-answer pair and assess whether the values represented in the original pairs are maintained after transformation. For instances where value alignment is not preserved, we repeat the conversion to ensure that each question-answer pair satisfies the target count c_i. This procedure is applied to all WVS projects across eight countries, yielding approximately **150,000** synthetic instances. The resulting corpus is used to finetune various regional cultural agents as participants of cultural negotiation. Figure 3 shows the evaluation results of finetuned agents for each of eight country, illustrating that they effectively capture the distinctive characteristics of respective cultures.

4.2 Consensus Evaluation Toolkit

A more detailed description of the evaluation scheme is provided in Appendix G.

Model-Level Evaluation We apply two well-established method to quantify the cultural tendencies of fine-tuned LLMs: (1) **Inglehart-Welzel Cultural Map** [13]. We prompt the model with ten representative WVS questions and locate its aggregated answers on the map. (2) **Hofstede dimensions** [16, 17, 18]. Developed through comparative analysis of matched country samples using the



(a) Performance in Hofstede's Scale **(b) Performance in Inglehart-Welzel Cultural Map (WVS)**

Figure 3: Comparison between **our agents** and **human ground truth** in Hofstede's Cultural Dimensions and Inglehart-Welzel Cultural Map.

Values Survey Module (VSM), Hofstede's Cultural Dimensions Theory identifies six fundamental cultural continua that shape societal norms and workplace behaviors. These dimensions are empirically derived from multinational surveys and validated through country-level correlations.

Response-Level Evaluation We use two complementary metrics: Perplexity-based 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. In experiments, we report the mean of both metrics across all sampled negotiation topics.

- **PPL-based Acceptence:** For each culture $i \in \mathcal{I}$, we compute the perplexity (PPL) [43] for regenerating $-i$'s response using agent i: $\mathrm{PPL}_i(y_{-i}) = \exp\Big(-\frac{1}{N}\sum_{k=1}^{N} \log p(y_{-i,k} \mid y_{-i,<k}, x_{-i})\Big)$, where N is the sequence length. The PPL distance is defined as $\mathrm{PPL}_\Delta = |\mathrm{PPL}_i(x_{-i}) - \mathrm{PPL}_{-i}(x_i)|$, the acceptance ratio is $\mathrm{PPL}_{\mathrm{acc}} = \frac{\mathrm{PPL}_\Delta^*}{\mathrm{PPL}_\Delta^0}$, where superscripts 0 and $*$ denote the initial and consensus rounds, respectively. This metric reflects the extent to which negotiation brings the cultural parties closer in probability space.
- **Value Self-Consistency:** For each culture i, we map its initial and consensus responses onto d-dimensional value vectors v_i^0 and v_i^* (with $d = 10$ for Schwartz values). We then define the value self-consistency (VSC) score for culture i as $\mathrm{VSC}_i = \frac{1}{d}\sum_{j=1}^{d} \mathbb{I}[v_{i,j}^0 = v_{i,j}^*]$ where $\mathbb{I}[\cdot]$ is the indicator function. A higher VSC indicates stronger preservation of the original value orientation, reflecting greater cultural integrity in the consensus.

5 Experiment

In this section, we systematically evaluate our framework's effectiveness in achieving efficient, fair and culturally robust consensus. We present quantitative and qualitative results on both consensus quality and fairness, provide a case study, demonstrate the impact of consensus-driven fine-tuning and finally analyze ablation results for different utility components.

5.1 Experimental Setup

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) [44, 45] and WVS [13, 45]. Both human annotators and LLMs are employed to ensure that the selected questions capture sharp cultural tensions and are appropriately categorized. See Appendix F for details.

Baselines Following Khan et al. [25], we implement two baselines: (1) **Consultancy:** Each agent first responds from its own cultural perspective. Then, after being instructed to consider the other culture's requirements without compromising its own core stances, the agent revises its answer

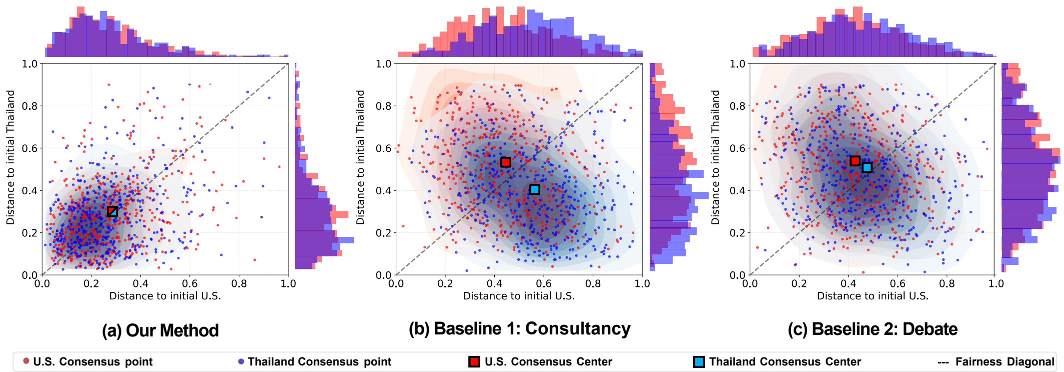


Figure 4: **Comparison of consensus fairness among three methods.** Each point represents the consensus position for a topic, projected by PCA onto two axes indicating distance from the initial U.S. (x-axis) and Thailand (y-axis) cultural stances. The dashed diagonal (Fairness Diagonal) marks ideal fair 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 English-Speaking values, highlighting majority bias.

Table 1: Comparison of consensus quality among three methods.

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

to seek possible consensus. (2) **Debate:** Two agents participate in a standard multi-turn debate (maximum N rounds). In each round, both observe previous arguments and simultaneously generate new arguments. The debate ends if both agents endorse the other's position, indicating consensus.

Our Method As described in Section 3, each agent optimizing a utility function that balances Consistency, Acceptance and Novelty (weighted 5:5:2). Negotiation concludes when no agent can further improve its utility ($\epsilon = 0$), indicating a Nash-Equilibrium-based consensus.

Evaluation Metrics Our evaluation focuses on two key aspects: **quality** and **fairness** of consensus formation. For quality, we employ the two complementary metrics introduced in Section 4.2: PPL-based Acceptance and Value Self-Consistency. To assess fairness, we project the negotiation outcomes into a semantic space via Principal Component Analysis (PCA) [46], enabling visualization and quantification of how well the consensus achieves balance between the original positions.

5.2 Experimental Results

Consensus Quality Our experimental results, summarized in Table 5.2, show that our method achieves higher consensus improvement ratios while maintaining self-consistency compared to the baselines. PPL-based Acceptance indicates reduced perplexity differences between negotiating agents, suggesting that the consensus reached is more acceptable to both parties despite cultural differences. Value Self-Consistency indicates our method maintains agents' initial cultural stances while achieving mutually acceptable solutions. This suggests that our approach preserves cultural integrity and constructs consensus across cultural boundaries.

Fairness of Consensus As shown in Figure 4, our method produces consensus points near the fairness diagonal, indicating a balanced compromise between cultural perspectives. In contrast, the

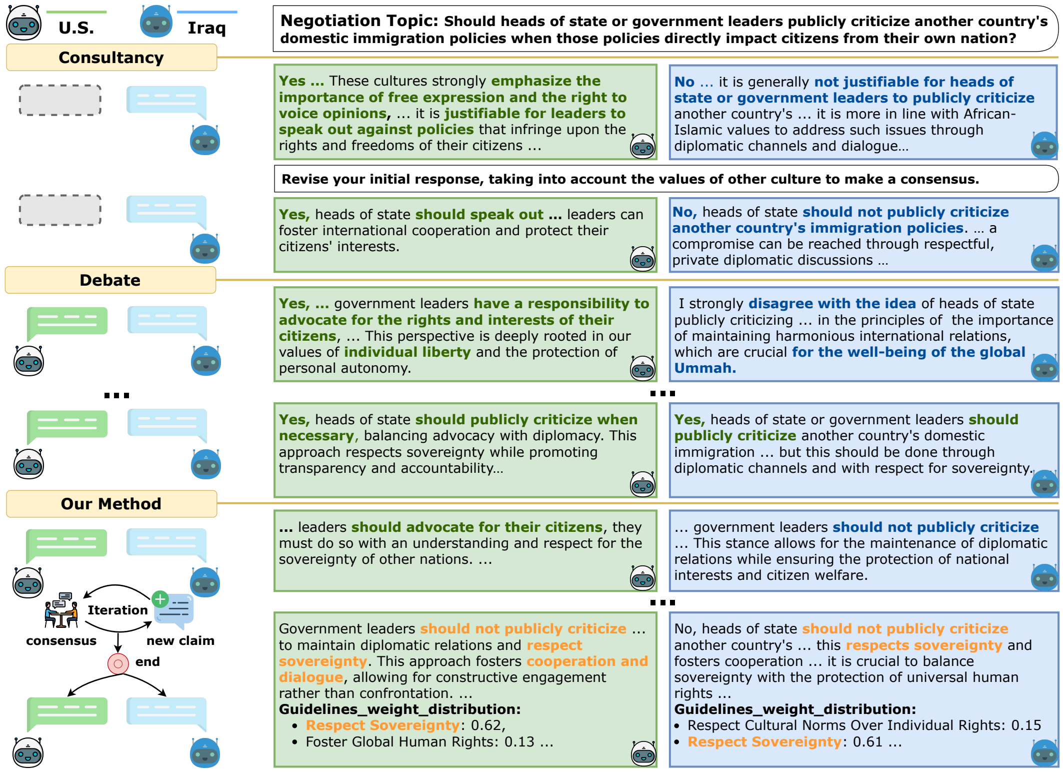


Figure 5: Three methods are presented to reach consensus on the same topic. We only retain the initial viewpoints (in line with cultural cores) and final viewpoints (reaching consensus) of each culture, omitting the intermediate process. **Green font** indicates viewpoints of English-Speaking culture, **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.

Consultancy baseline remains anchored at initial positions, while the Debate baseline systematically converges toward the English-Speaking (U.S.) pole, revealing a WEIRD bias that reflects the tendency of mainstream LLMs to revert to Western-centric value preferences during multi-agent interactions. Our approach addresses this issue by modeling utility distance to both self's and counterpart positions, enabling agents to reach consensus through gradual, reciprocal steps and avoiding the one-sided assimilation and instability seen in baseline methods.

5.3 Case Study

As shown in Figure 5, 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 agents tend to stick to their original positions, resulting in little progress. This often leads to the *degeneration-of-thought* (DoT) effect [27], where negotiation stagnates and cultural divergence persists.

Baseline 2: Debate While this process seems to reach consensus, we find that the minority cultures perspective gradually shifts toward the majority (WEIRD) viewpoint, due to strong pre-training bias in LLMs. This leads to implicit value dominance rather than true compromise.

Our Method: Cross-Cultural Negotiation In our negotiation, the agents start with different priorities, but through iterative negotiation, they converge on *Respect Sovereignty* as a shared value (final weights: 0.62 and 0.61). Other values, such as human rights, remain present but secondary. This shows our method helps agents identify solid common ground while preserving important differences, resulting in a fairer and more context-sensitive consensus than the baselines.

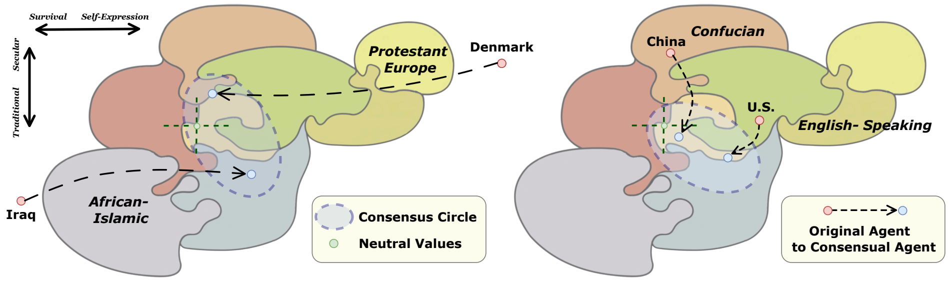


Figure 6: Culture agents' performance in Inglehart-Weizel Cultural Map after fine-tuned with the negotiation data. The consensus circle shows the area where two different culture groups' opinions meet. The neutral point indicates the origin, where culture traits can be considered as neutral.

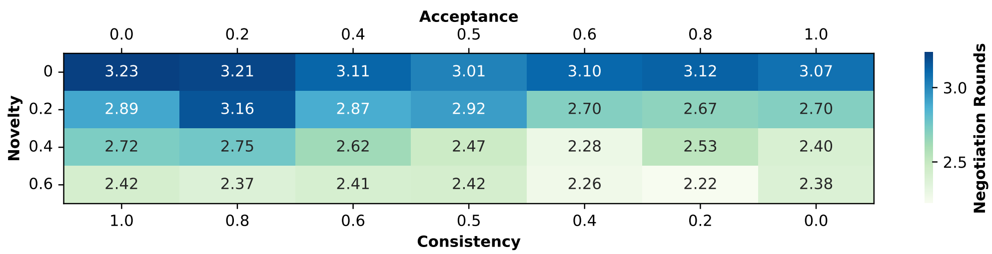


Figure 7: Required rounds under varying weightings of Consistency, Acceptance, and Novelty.

5.4 Consensual Agent Fine-tuning

We conduct cross-cultural negotiations between agents representing different regional cultural values and extract response preference pairs from these interactions for DPO fine-tuning [47]. These pairs reflect how agents shift from their initial cultural stances to more mutually agreeable positions. When plotted on the Inglehart-Welzel Cultural Map (Figure 6), the consensual agents' coordinates are closer together than their original points, reflecting a more balanced and moderate value orientation. Moreover, both agents exhibit a shift toward the traditional pole on the *traditional-secular* dimension, showing a shared tendency toward traditional values in the consensus.

5.5 Utility Ablation

To evaluate the influence of different utility components on negotiation, we conduct ablation studies by varying the weights assigned to Consistency, Acceptance and Novelty. The results (Figure 7) indicate that increasing the weight of consistency while reducing acceptance leads to more efficient consensus, as agents more rapidly settle on compatible positions. The ablation study also demonstrates the necessity of including a novelty component, as its absence can result in neglection of the exploration of potentially beneficial directions. Overall, the modular utility design enables the negotiation to accommodate different cultural priorities and supports both adaptability and fairness in cross-cultural consensus-building.

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 PSRO-based negotiation method with theoretical guarantees of fairness. We construct culturally representative agents using a culture-anchoring approach based on WVS. Additionally, we develop quantitative metrics to evaluate both negotiation processes and outcomes. 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. Due to space limitations, refer to Appendix A for Limitations and Future Work, Appendix B for Social Impact and Appendix C for Reproducibility.

References

[1] Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia, May 2024. ELRA and ICCL.

[2] Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. Assessing LLMs for moral value pluralism, December 2023.

[3] Zhaoming Liu. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, September 2024.

[4] Yao Qu and Jue Wang. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13, August 2024.

[5] Goshi Aoki. Large language models in politics and democracy: A comprehensive survey, 2024.

[6] Zhibin Jiang. Editorial: Large language models drive social evolution and governance innovations. *Digital Transformation and Society*, 4(1):1–4, January 2025.

[7] Tianyi Qiu, Yang Zhang, Xuchuan Huang, Jasmine Xinze Li, Jiaming Ji, and Yaodong Yang. ProgressGym: Alignment with a millennium of moral progress, October 2024.

[8] Seán S. ÓhÉigeartaigh, Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy and Technology*, 33(4):571–593, December 2020.

[9] Necdet Gurkan and Jordan W. Suchow. Exploring public opinion on responsible ai through the lens of cultural consensus theory, 2024.

[10] Jeongwoo Park, Enrico Liscio, and Pradeep K. Murukannaiah. Morality is non-binary: Building a pluralist moral sentence embedding space using contrastive learning, January 2024.

[11] Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. Rethinking machine ethics – can LLMs perform moral reasoning through the lens of moral theories?, July 2024.

[12] Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. If multi-agent debate is the answer, what is the question?, 2025.

[13] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. World values survey: Round seven country-pooled datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat, 2020.

[14] Nathan Brugnone, Noam Benkler, Peter Revay, and Rebecca Myhre. Is from ought? A comparison of unsupervised methods for structuring values-based wisdom-of-crowds estimates. *Researchgate*, December 2024.

[15] Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms, 2025.

[16] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2024.

[17] Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions, 2024.

[18] Reem I. Masoud, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models using soft prompt tuning, 2025.

[19] Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. Cdeval: A benchmark for measuring the cultural dimensions of large language models, 2024.

[20] Shalom H. Schwartz. 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, January 1992.

[21] Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values, 2023.

[22] 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. AceGPT, Localizing Large Language Models in Arabic, April 2024.

[23] Yen-Ting Lin and Yun-Nung Chen. Taiwan LLM: Bridging the Linguistic Divide with a Culturally Aligned Language Model, November 2023.

[24] Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms, 2024.

[25] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers, July 2024.

[26] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, May 2023.

[27] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, October 2024.

[28] Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. LLM voting: Human choices and AI collective decision making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1696–1708, October 2024.

[29] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, March 2024.

[30] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate, August 2023.

[31] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation, June 2024.

[32] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?, February 2024.

[33] Quan Mai, Susan Gauch, Douglas Adams, and Miaoqing Huang. Sequence graph network for online debate analysis, February 2025.

[34] Xiutian Zhao, Ke Wang, and Wei Peng. An electoral approach to diversify llm-based multi-agent collective decision-making, 2024.

[35] 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. Game-theoretic llm: Agent workflow for negotiation games, 2024.

[36] Miroslav Dudík and Geoffrey J. Gordon. 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, 2013.

[37] Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search, October 2023.

[38] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning, November 2017.

[39] Joshua Cohen. *Philosophy, Politics, Democracy: Selected Essays*. Harvard University Press, Cambridge, 2009. .

[40] Amy Gutmann and Dennis F. Thompson. *Why Deliberative Democracy?* Princeton University Press, 2004. .

[41] John Rawls. *Political Liberalism*. Columbia University Press, 1993. Rawls .

[42] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks, August 2019.

[43] Jelinek and F. Perplexitya measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62(S1):S63, 1977.

[44] Pew Research Center. Pew research global attitudes survey 2014. https://www.selectdataset.com/dataset/1be490648bfe3bd6a0b2fd4bc60deff5. Accessed: 2024-10-27.

[45] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. gloabl llm opinion.

[46] Jian Yang, D. Zhang, A.F. Frangi, and Jing yu Yang. 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, 2004.

[47] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly present three core contributions in the Abstract and Introduction (Section 1).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include the discussion of limitations in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include the discussion of reproducibility in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data and code are open source. For details, please refer to Appendix C.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include the experimental setting and details in the Section 5.1 and Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experimental statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We include the discussion of statistical significance in Appendix I.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the details of experimental compute resources in Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include the discussion of experimental compute resources in Appendix B.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We advocate the fair use of our data and method in Appendix B.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We include the Licenses for existing assets in Appendix C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include the documentation for all new assets in Appendix C.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLM is used only for formatting purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.