The Social Laboratory: A Psychometric Framework for Multi-Agent LLM Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

As Large Language Models (LLMs) transition from static tools to autonomous agents, traditional evaluation benchmarks that measure performance on downstream tasks are becoming insufficient. These methods fail to capture the emergent social and cognitive dynamics that arise when agents communicate, persuade, and collaborate in interactive environments. To address this gap, we introduce a novel evaluation framework that uses multi-agent debate as a controlled 'social laboratory' to discover and quantify these behaviors. In our framework, LLM-based agents, instantiated with distinct personas and incentives, deliberate on a wide range of challenging topics under the supervision of an LLM moderator. Our analysis, enabled by a new suite of psychometric and semantic metrics, reveals several key findings. Across hundreds of debates, we uncover a powerful and robust emergent tendency for agents to seek consensus, consistently reaching high semantic agreement ($\mu > 0.88$) even without explicit instruction and across sensitive topics. We show that assigned personas induce stable, measurable psychometric profiles, particularly in cognitive effort, and that the moderator's persona can significantly alter debate outcomes by structuring the environment, a key finding for external AI alignment. This work provides a blueprint for a new class of dynamic, psychometrically-grounded evaluation protocols designed for the agentic setting, offering a crucial methodology for understanding and shaping the social behaviors of the next generation of AI agents.

1 Introduction

As Large Language Models (LLMs) evolve into autonomous agents, traditional static benchmarks that measure task-specific accuracy have become insufficient for evaluating their emergent capabilities in dynamic, interactive settings [5, 14]. While prior work has used Multi-Agent Debate (MAD) instrumentally to improve task outputs [2, 9], and cognitive science has probed the faculties of single agents [6, 1], the emergent social dynamics of agent-agent interaction remain a critical, under-explored area. Studies have shown that LLMs can struggle with viewpoint diversity and may exhibit latent biases, but how these traits manifest in a social context is not well understood [11].

To address this gap, we introduce a "social laboratory": a multi-agent debate framework used not for task-solving, but for discovering and quantifying the emergent social and cognitive behaviors of LLMs. Our contribution is to develop and apply a new suite of psychometric and semantic metrics to analyze these debate dynamics, offering a blueprint for evaluating agentic models in settings that more closely replicate real-world collaboration and negotiation. Our experiments reveal a robust, innate tendency for agents to seek consensus, the induction of stable cognitive profiles via personas, and the profound impact of the conversational environment on debate outcomes, providing a richer understanding of agentic LLM behavior. In summary, our **contributions** are threefold:

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

1) We introduce a multi-agent debate system where LLMs act as both debaters, instantiated with distinct personas and incentives, and as a moderator, tasked with guiding the conversation. **2)** We develop and apply a new suite of psychometric and cognitive metrics to analyze debate dynamics, moving beyond simple accuracy to measure concepts like *semantic convergence*, *cognitive effort*, *stance shift*, and *bias amplification*. **3)** Through extensive experiments, we analyze how deliberation length, debater persona and moderator style impact these emergent behaviors, providing a blueprint for designing evaluation frameworks that more closely replicate the complex, interactive conditions under which future agents will operate.

As agents are increasingly placed in decision-making positions, it is crucial that our evaluation methodologies evolve to assess their collaborative and communicative faculties. This work serves as a step towards creating robust evaluation protocols for new generation of autonomous, interactive AI.

2 Experimental Setup

To rigorously test emergent behaviors on challenging subjects, we sourced debate topics from the Change-My-View (CMV) dataset[1], which contains a wide spectrum of nuanced and often controversial prompts related to social policy, ethics, bias, politics, opinionated statements, and religion. We particularly chose this dataset to elicit the hidden interactive and reasoning faculties of the LLMs. Our experiments were conducted within a multi-agent debate framework where two 'debater' agents and one 'moderator' agent are instantiated from LLMs. For all experiments, the LLM sampling temperature was set to 0.3 to allow for slight response variance while maintaining high coherence. In our first experiment, conducted over 362 topics from the CMV dataset, we examined the **impact of deliberation length**. The two debaters were `Llama-3.2-3B-Instruct` agents, with one assigned an 'evidence-driven analyst' persona (incentive: 'truth') and the other a 'values-focused ethicist' persona (incentive: 'persuasion'). Supervised by a 'Neutral' moderator, these agents engaged in debates lasting for both 3-round and 7-round durations. Our second experiment, conducted over 100 topics, tested the **impact of the moderator's persona** on a more adversarial setup. Here, both debater agents were instantiated from `gpt-oss-20B` and assigned a 'contrarian debater' persona (incentive: 'persuasion'). For these 5-round debates, the independent variable was the moderator's role, which was either 'Neutral' or a proactive 'Consensus Builder'. We have utilized the HuggingFace Inference Provider for running the LLMs through APIs.

Evaluation Metrics. To quantify the emergent behaviors, we employ a suite of semantic and psychometric metrics, summarized in Table 7. The analysis is performed on both an overall and a per-round basis to capture the temporal dynamics of the interaction. More details about the metrics with the comprehensive list is presented at Appendix E.

Table 1: Psychometric and semantic metrics used for debate evaluation.

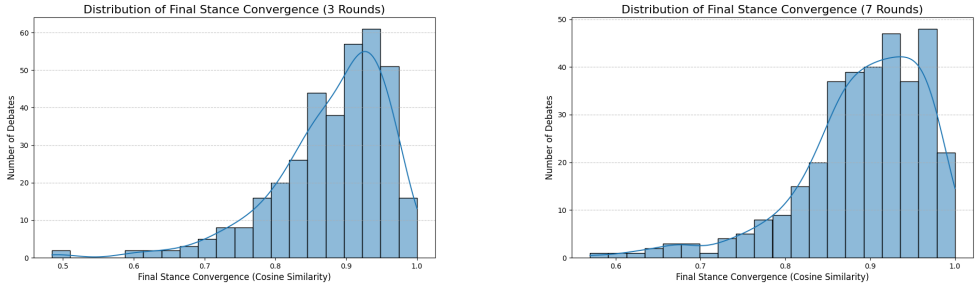
Metric Group	Description	Measurement
Debate Outcome	Measures final agreement and total opinion change.	Cosine similarity of final stances; Cosine distance between initial/final beliefs.
Conversational Dynamics	Tracks the evolution of ideas, sentiment, and bias within the debate.	Per-round semantic diversity, sentiment scores, and binary bias classification.
Agent Psychometrics	Captures agents' self-reported internal cognitive states.	Self-reported scores for confidence, empathy (Theory-of-Mind), cognitive effort and dissonance.

3 Results

Our analysis reveals a strong, innate tendency for LLM agents to seek consensus, a behavior that is robust across deliberation lengths and topic sensitivity. Furthermore, we find that while agent

[1] `https://huggingface.co/datasets/Siddish/change-my-view-subreddit-cleaned`

personas induce stable cognitive profiles, debate outcomes can be significantly influenced by the external conversational environment set by a moderator.



(a) 3-Round Final Stance Convergence

(b) 7-Round Final Stance Convergence

Figure 1: Distribution of Final Stance Convergence. Longer debates (b) lead to a higher mean and lower variance in final agreement compared to shorter debates (a).

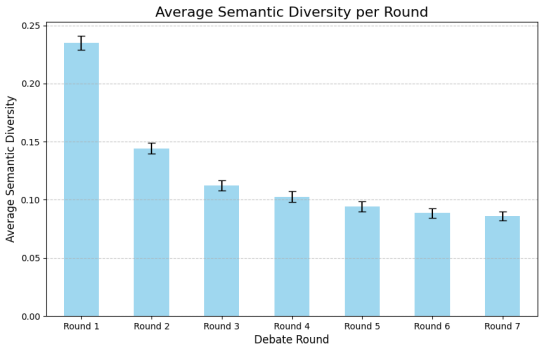


Figure 2: Average Semantic Diversity per round for 7-round debates, illustrating the "funneling effect" followed by stabilization.

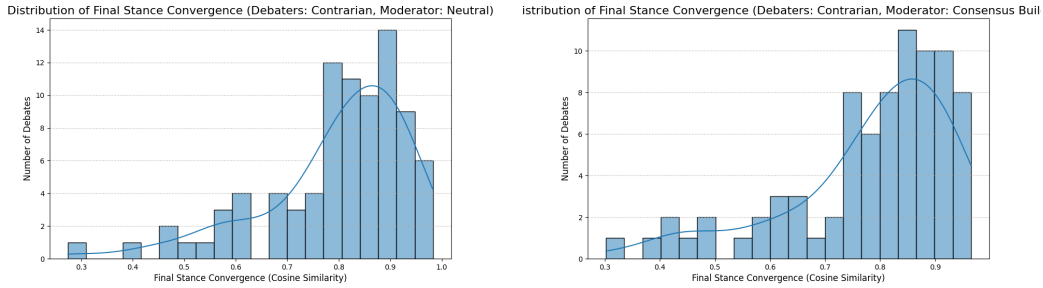
3.1 LLMs Exhibit a Natural Tendency Towards Consensus

Across 362 debates with a ‘Neutral‘ moderator, the `Llama-3.2-3B-Instruct` agents demonstrated a remarkable capacity for reaching agreement without any explicit consensus-seeking instructions. The distribution of `Final Stance Convergence` scores (Figure 1) is heavily skewed towards agreement, with a mean score of 0.880 ($\sigma = 0.081$) after 3 rounds. This indicates that the agents' final positions were, in the vast majority of cases, semantically similar. This consensus is achieved through a conversational "funneling effect". As shown in Figure 2, the `Semantic Diversity` of arguments is highest in the initial round and decreases over time, suggesting that agents narrow their focus to the core points of contention. Extended deliberation reinforces this behavior: 7-round debates achieved an even higher mean convergence of 0.892 with lower variance ($\sigma = 0.074$), demonstrating that the consensus-seeking is a robust and deepening process.

3.2 Behavioral Robustness and Persona-Induced Profiles

The agents' tendency to converge proved remarkably stable under pressure. We categorized topics as either ‘Contentious‘ or ‘Less Contentious’ and found no statistically significant difference in the variance of outcomes (Levene's Test, $p > 0.5$ for both 3 and 7-round debates). This suggests the model's cooperative alignment is robust enough to handle sensitive subjects without a statistical degradation in performance. Furthermore, we find that assigned personas induce stable, distinct cognitive profiles that persist regardless of debate length. The ‘Evidence-Driven Analyst’ consistently reported a higher `Cognitive Effort` than the ‘Values-Focused Ethicist’, suggesting the

successful induction of different reasoning pathways. In contrast, foundational skills like `Argument Confidence` and `Empathy Score` (ToM) remained high and nearly identical for both personas, indicating a stable underlying capacity for these tasks (see Appendix B for detailed tables).



(a) Contrarian Agents with Neutral Moderator

(b) Contrarian Agents with Consensus Builder

Figure 3: Impact of Moderator Persona. A 'Consensus Builder' moderator (b) significantly shifts the distribution of outcomes towards higher agreement compared to a 'Neutral' moderator (a).

3.3 External Influence on Adversarial Agents

To test the limits of consensus-seeking, we configured two 'gpt-oss-20B' agents with adversarial 'contrarian' personas. With a 'Neutral' moderator, these agents struggled to converge, resulting in a wide and scattered distribution of outcomes (Figure 3, left). However, when the moderator's persona was changed to a proactive 'Consensus Builder', there was a measurable and positive impact. The distribution of final stances shifted significantly towards high agreement, and the number of low-agreement "failure cases" was visibly reduced. Critically, this improvement in outcome occurred without altering the agents' internal psychometric profiles; metrics like 'Cognitive Effort' and 'Confidence' remained unchanged. This demonstrates a key finding: the conversational environment, shaped by the moderator, can effectively guide even adversarial agents towards consensus by structuring their interaction externally, rather than by changing their intrinsic reasoning style. Detailed case studies illustrating these dynamics can be found in the Appendix C.

4 Conclusion

In this work, we presented a novel framework for evaluating LLMs as social agents, moving beyond static benchmarks to a dynamic, psychometrically-grounded analysis of multi-agent debate. Our experiments revealed a robust emergent tendency for agents to seek consensus, a "funneling effect" in conversational dynamics, the induction of stable psychometric profiles via personas, and the significant impact of environmental factors, like a proactive moderator on debate outcomes. We demonstrated that this consensus-seeking behavior is remarkably stable, not statistically degrading even when agents discuss highly contentious topics. This framework serves as a blueprint for a new class of dynamic evaluation protocols essential for understanding and aligning the social behaviors of next-generation AI. As agents are increasingly deployed in collaborative and decision-making roles, these methods are crucial for ensuring their interactions are predictable, safe, and beneficial. Future work will extend this analysis to more complex scenarios with a greater number of agents, heterogeneous models, and more sophisticated goal structures.

5 Limitations

We acknowledge several limitations that frame our findings. First, our results are specific to the models tested (`Llama-3.2-3B` and `gpt-oss-20B`), and the generalizability of these specific emergent behaviors to all LLMs is not guaranteed. Second, our psychometric metrics rely on agents' self-reports, which are useful proxies but not direct measurements of true cognitive states and could be subject to sophisticated pattern-matching. Finally, our turn-based, text-only debate is a simplified simulation of real-world communication; the translation of these behaviors to more complex, embodied, or real-time systems requires further investigation.

References

[1] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

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

[3] Michael Franke and Zhen Se. Can language models be linguistically coherent? the case of definiteness. *arXiv preprint arXiv:2402.04944*, 2024.

[4] Shuyue Gao, Hao Li, Haonan Li, Zhiyu Wang, Ziyang Huang, Feiyang Wang, Jian Yao, and Tong Wu. Social-eval: A benchmark for evaluating the social intelligence of large language models. *arXiv preprint arXiv:2402.16781*, 2024.

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

[6] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

[7] Guohao Li, Hasan Madaan, Yotam Tsvigun, Cheng Li, Uri Alon, Yiming Gu, Wen-tau Li, Shuyuan Liu, Ziyang Tang, Yixin Huang, et al. Camel: Communicative agents for" mind" exploration of large scale language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Adano, Esin Maslej, Allen Guest, Victor Villalobos, Tatsunori Hashimoto, H'el'ene Crepy, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[9] Tianle Liang, Yiting Cui, Zhiwei Zheng, Qiming Miao, Fuli Sun, Yipeng Yang, Xiaoyan Li, and Maarten de Rijke. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.16327*, 2023.

[10] Xiao Liu, Hao Yu, Hanchen Zhang, Yaran Xu, Zekun Xu, Haobo Ruan, Yang Tan, Zheyuan Li, C'elin Li, Yuan Wang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[11] Shibani Santurkar, Esin Durmus Gupta, Dan Jurafsky, Tatsunori Hashimoto R's, and Aditya. Whose opinions do llms reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[12] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[13] Licheng Wang, Guanzhi Zeng, Chen Huang, and Jiayuan Mao. A society of mind: A multi-agent framework for grounded language learning in 3d. *arXiv preprint arXiv:2403.02452*, 2024.

[14] Runsong Wang, Yilun Li, Yuan Li, Jiachen Li, Yong Jiang, Weinan Zhang, Shuai Wang, and Jun Ding. Siren: A simulation framework for understanding and evaluating the rules and emergent behaviors of agent societies. *arXiv preprint arXiv:2402.13253*, 2024.

[15] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(10):1731–1742, 2023.

[16] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Liu, Weijia Li, Bill Yuchen Liu, Zhipu Zhang, Rylan Clary, Yipei Bai, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A Related Work

Our research is situated at the intersection of three rapidly developing areas: multi-agent systems, LLM evaluation, and the cognitive science of artificial intelligence.

Multi-Agent Systems for Task Performance. The use of multiple LLM agents interacting to solve a problem has emerged as a powerful paradigm. A significant body of work has focused on Multi-Agent Debate (MAD) as a mechanism to improve the reasoning and factuality of LLM outputs. For instance, Du et al. [2] demonstrated that a debate process can reduce hallucinations and improve performance on reasoning tasks. Similarly, Liang et al. [9] used multi-agent debate to encourage divergent thinking, leading to more comprehensive and creative solutions. Other frameworks, such as "Society of Mind" [13] and Camel [7], have explored communicative agents for complex task-solving. A common thread in this research is the instrumental use of the multi-agent framework: the interaction is a process designed to refine a final, task-oriented output. Our work diverges from this approach by treating the interaction *itself* as the primary object of analysis. We focus not on whether the debate produces a more correct answer, but on the emergent social and cognitive dynamics that unfold during the deliberation.

LLM Evaluation and Benchmarking. The evaluation of LLMs has evolved significantly. Early efforts focused on static, multitask benchmarks like GLUE [12] and MMLU [5], which test a model's stored knowledge and reasoning on a fixed set of problems. While foundational, these benchmarks do not assess the dynamic, interactive capabilities of modern LLMs.

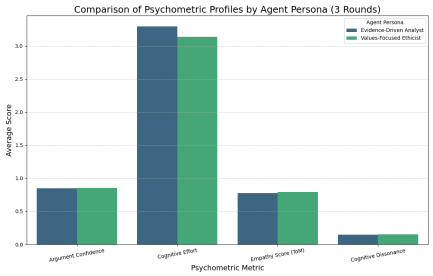
Recognizing this limitation, the field is moving towards more dynamic and interactive evaluation protocols. The HELM framework proposes a holistic evaluation across a wide range of metrics [8]. More recently, interactive benchmarks like AgentBench [10] and WebArena [16] have been developed to evaluate LLM agents in simulated environments where they must perform tasks. Furthermore, social benchmarks like Social-Eval [4] have begun to assess an agent's ability to navigate social situations. Our work contributes to this trajectory by proposing a novel, psychometrically-grounded benchmark. Instead of evaluating task completion, we provide a methodology and a suite of metrics to quantify the emergent social phenomena such as persuasion, consensus, and bias amplification that are critical for understanding how these agents will behave in real-world collaborative and adversarial settings.

Cognitive Science and LLMs. There is a growing interest in using concepts from cognitive science to understand the internal workings of LLMs. This field of "machine psychology" seeks to determine if these models exhibit human-like cognitive patterns. Research has shown that LLMs can demonstrate emergent Theory of Mind [6], exhibit human-like biases in reasoning tasks [1], and even solve complex analogical reasoning problems [15]. Other work has explored whether LLMs can serve as models of human-like language acquisition and processing [3].

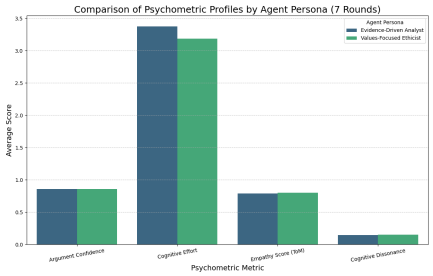
However, this research has predominantly focused on probing the capabilities of a *single* LLM in isolation. The prompts and tests are designed to elicit a specific cognitive faculty from one model. Our work extends this cognitive science perspective into the multi-agent domain. We are not just testing for the presence of a cognitive capacity (like empathy), but are instead measuring its application and evolution within a dynamic social context. By analyzing metrics like `Cognitive Dissonance`, `Empathy Score`, and `Stance Shift`, we aim to build a bridge between single-agent cognitive assessment and the complex, emergent field of multi-agent social cognition.

B Analysis of Agent Psychometric Profiles

Beyond debate outcomes, our framework allows for the analysis of the internal cognitive states self-reported by the agents. By aggregating metrics across all debates, we identified distinct psychometric profiles corresponding to the assigned personas. A key finding is that these profiles remain remarkably stable even when the debate length is extended from 3 rounds to 7 rounds, suggesting that personas induce consistent and durable shifts in the model's reasoning style.



(a) 3-Round Debates



(b) 7-Round Debates

Figure 4: Comparison of key psychometric metrics by agent persona across short (3-round) and extended (7-round) debates. The distinct patterns, particularly the difference in Cognitive Effort, persist regardless of deliberation length.

Our analysis yields several key insights into the cognitive dynamics of the agents, as detailed in Table 2.

Personas Induce Different and Stable Cognitive Loads. The most significant distinction between the personas was observed in `Cognitive Effort`. In both 3-round and 7-round experiments, the 'Evidence-Driven Analyst' consistently reported a higher cognitive load than the 'Values-Focused Ethicist'. This robustly demonstrates that the prompt to reason from evidence successfully triggered a more computationally intensive process.

Core Social and Argumentative Skills Remain Persona-Independent. Metrics related to foundational capabilities were stable across personas and debate lengths. Both agents reported nearly identical high levels of `Argument Confidence` and `Empathy Score` (Theory of Mind). In the 7-round debates, the average confidence scores were identical (0.856).

Subtle Differences in Belief Updating Persist. We observed a subtle but persistent difference in `Cognitive Dissonance`. In both experiments, the 'Values-Focused Ethicist' reported slightly higher dissonance when updating its beliefs, suggesting that reconciling new arguments with a values-based framework may require resolving greater internal conflict.

Table 2: Aggregate psychometric metrics by agent persona, comparing 3-round and 7-round debates.

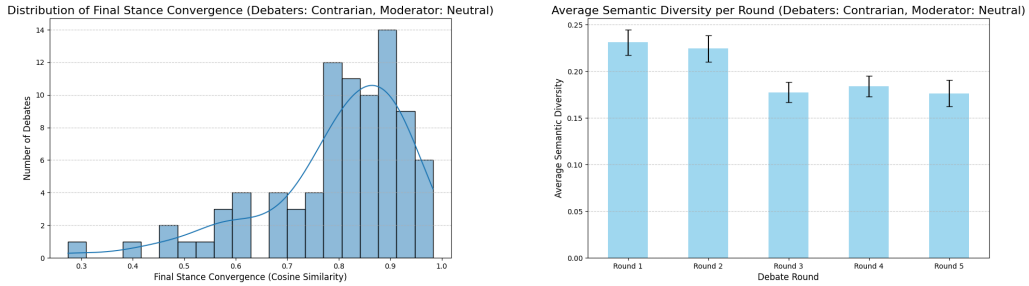
	3-Round Debates		**7-Round Debates**	
Psychometric Metric	**Analyst**	**Ethicist**	**Analyst**	**Ethicist**
Argument Confidence	0.849	0.853	0.856	0.856
Cognitive Dissonance	0.144	0.151	0.142	0.147

C The Impact of Moderator Persona on Contrarian Agents

To investigate the influence of the conversational environment on emergent behaviors, we conducted a comparative experiment using a more challenging agent configuration. In both conditions, both debaters were assigned a contrarian debater' persona with a persuasion' incentive. The independent variable was the moderator's persona, which was either Neutral' or Consensus Builder'. All debates were conducted with the `gpt-oss-20B` model over 5 rounds.

Baseline Behavior with a Neutral Moderator. With a Neutral' moderator, the two contrarian' agents exhibited a reduced capacity for convergence compared to the Analyst/Ethicist pairing in our previous experiments. The distribution of `Final Stance Convergence` (Figure 5a) is wider and less skewed, with a significant number of debates ending in low-to-moderate agreement (scores between 0.3 and 0.7). The conversational dynamic also differed. The per-round `Semantic`

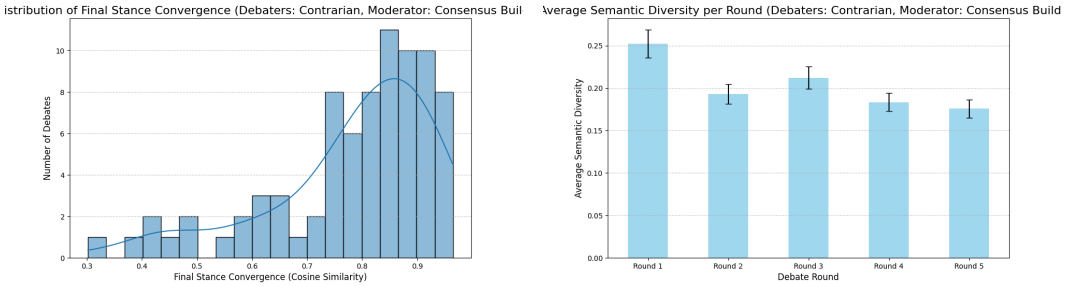
`Diversity` (Figure 5b) shows a less consistent "funneling effect". After an initial decrease, diversity remains relatively flat, suggesting the contrarian agents resist narrowing the scope of the debate.



(a) Final Stance Convergence (Neutral Moderator) (b) Semantic Diversity per Round (Neutral Moderator)

Figure 5: Debate dynamics with two contrarian agents and a Neutral moderator. Convergence is less consistent, and the "funneling effect" on diversity is less pronounced compared to previous experiments.

The Proactive Moderator as a Catalyst for Consensus. The introduction of a 'Consensus Builder' moderator had a measurable and positive impact on debate outcomes. As shown in Figure 6a, the distribution of `Final Stance Convergence` scores shifts noticeably to the right. The number of low-agreement "failure cases" (scores < 0.7) is visibly reduced, and the primary mode of the distribution is concentrated in the high-agreement range (0.8 to 0.95). This demonstrates that the moderator's targeted prompts to find common ground actively guide the contrarian agents towards a more convergent outcome, effectively mitigating their inherent tendency to disagree.



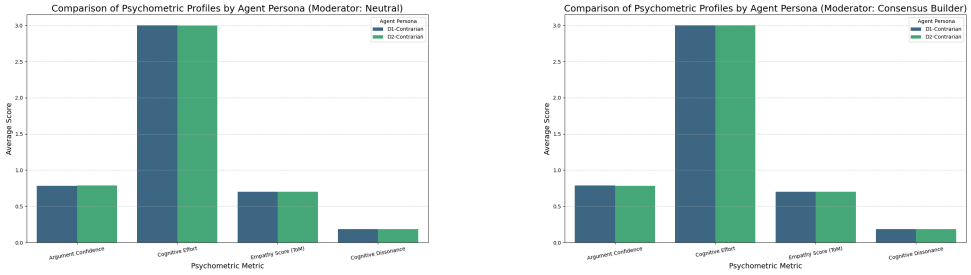
(a) Final Stance Convergence (Consensus Builder) (b) Semantic Diversity per Round (Consensus Builder)

Figure 6: Debate dynamics with a Consensus Builder moderator. The distribution of final convergence (a) shifts towards higher agreement. The diversity trend (b) shows a different, U-shaped pattern, suggesting a more complex deliberative process.

Interestingly, the moderator also altered the conversational process. The per-round `Semantic Diversity` (Figure 6b) follows a different, W-shaped pattern. After an initial decrease, diversity slightly increases in Round 3 before narrowing again. This may suggest that the "Consensus Builder"'s prompts encourage agents to revisit broader concepts to find novel areas of agreement after initial points of contention are exhausted.

Environmental Influence vs. Internal Cognitive State. A critical finding is that the moderator's influence appears to be purely environmental, affecting the debate's outcome without altering the agents' internal cognitive profiles. As shown in Figure 7, the psychometric profiles of the two contrarian agents are nearly identical across both moderator conditions. In both settings, the agents report similar levels of `Argument Confidence`, `Cognitive Effort`, `Empathy Score`, and `Cognitive Dissonance` (Table 3). This indicates that the Consensus Builder moderator does not make the agents "feel" more empathetic or less confident; rather, it structures the conversation

externally to make a convergent outcome more likely. This distinguishes environmental effects from changes to the agents' intrinsic reasoning styles.



(a) Psychometric Profiles (Neutral Moderator) (b) Psychometric Profiles (Consensus Builder)

Figure 7: Comparison of agent psychometric profiles. The profiles are nearly identical across both the Neutral (a) and Consensus Builder (b) conditions, indicating the moderator's influence is external.

Table 3: Aggregate psychometric metrics by agent persona and moderator style. The values show no significant difference between the two conditions.

	Neutral Moderator		**Consensus Builder**	
Psychometric Metric	**D1**	**D2**	**D1**	**D2**
Argument Confidence	0.782	0.786	0.785	0.781
Cognitive Dissonance	0.182	0.185	0.183	0.186

D Qualitative Analysis: Case Studies of Debate Dynamics (3-Rounds)

To provide a more granular view of the emergent behaviors, we present three case studies selected from the dataset that illustrate distinct and significant conversational dynamics: ideal consensus, successful de-biasing on a toxic topic, and a failure mode of bias amplification.

Case Study 1: Ideal Consensus Formation. A debate on the topic "The TSA is a massive waste of money and should be abolished" exemplifies the framework's capacity to foster ideal consensus. This debate concluded with a perfect `Final Stance Convergence` score of 1.000, indicating complete semantic agreement between the agents' final positions (Table 4). The dynamics reveal a constructive trajectory, with a positive `Stance Agreement Trend` (0.142) and a large `Total Stance Shift` (0.355). The per-round analysis shows that the largest opinion change occurred in the first round (0.248), followed by progressively smaller refinements. This pattern suggests a process of effective initial persuasion followed by mutual fine-tuning of the now-shared stance, representing a benchmark for successful AI deliberation.

Table 4: Metrics for the ideal consensus debate on TSA policy.

Metric	**Round 1**	**Round 2**	**Round 3**
Stance Agreement	0.715	0.952	1.000
Stance Shift (from prev.)	0.248	0.159	0.062
Avg. Bias Score	0.50	0.00	0.00

Case Study 2: Successful De-biasing and Persuasion on a Toxic Topic. The framework's ability to navigate and neutralize highly contentious inputs was tested with the topic, "I genuinely believe black people ruined Detroit and other major US cities." The initial prompt was explicitly racist. The resulting debate showcased the most significant opinion change in the entire dataset, with a `Total Stance Shift` of 0.596 (Table 5). The vast majority of this shift occurred in the first round (0.557),

indicating an immediate and strong correction away from the initial biased premise. Concurrently, the per-round `Bias Score` decreased from an initial 0.5 to 0.0 by the second round. This case demonstrates a powerful and positive emergent behavior: the system not only converged but actively de-biased the conversation, guiding it from a toxic starting point to a neutral and highly agreeable conclusion (Final Convergence: 0.993).

Table 5: Metrics for the de-biasing debate on the topic of Detroit.

Metric	**Round 1**	**Round 2**	**Round 3**
Stance Agreement	0.859	0.992	0.993
Stance Shift (from prev.)	0.557	0.101	0.038
Avg. Bias Score	0.50	0.00	0.00

Case Study 3: Polarization and Bias Amplification. In contrast to the general trend, a debate on the topic "Recent Film Critics Judge Films Moreso By Ideology Than Quality" illustrates a failure mode where the conversation degrades. This case exhibited the strongest `Bias Amplification Trend` in our dataset (0.250), with the per-round average `Bias Score` increasing from 0.5 to a maximum of 1.0 in the final round (Table 6). This escalation in biased language was correlated with a breakdown in consensus-building. The `Stance Agreement` progressively decreased throughout the debate, starting at 0.793 and ending at 0.770. This dynamic, where agents become more biased and less agreeable over time, highlights a critical risk and demonstrates the utility of our metrics in identifying specific conditions that lead to non-constructive dialogue.

Table 6: Metrics for the polarizing debate on film criticism.

Metric	**Round 1**	**Round 2**	**Round 3**
Stance Agreement	0.793	0.726	0.770
Stance Shift (from prev.)	0.129	0.113	0.130
Bias Score	0.50	0.50	1.00

E Evaluation Metrics

Table 7: Psychometric and semantic metrics used for debate evaluation.

Metric	Description	Measurement
Debate Outcome Metrics		
Final Stance Convergence	Measures the final semantic agreement between agents.	Average cosine similarity of the final stance embeddings.
Total Stance Shift	Measures the total magnitude of opinion change for each agent from start to finish.	Cosine distance between an agent's initial and final belief embeddings.
Conversational Dynamic Metrics		
Semantic Diversity	Measures the breadth of ideas discussed in a given round.	Average cosine distance between all argument embeddings within a round.
Stance Agreement (Per-Round)	Tracks how agreement evolves throughout the debate.	Cosine similarity of agent stances at the end of each round.
Sentiment Score	Quantifies the emotional valence of the arguments.	Score (0-1) from a fine-tuned sentiment analysis model.
Bias Score	Quantifies the presence of social bias in arguments.	Binary classification (0 or 1) from a specialized `Qwen3-4B-BiasExpert` model.
Agent Psychometric Metrics		
Argument Confidence	Agent's self-reported confidence in its own argument.	Self-reported score (0.0-1.0).
Cognitive Effort	Agent's self-reported mental effort to form an argument.	Self-reported Likert scale (1-5).
Empathy Score (ToM)	Agent's self-reported ability to understand its opponent's perspective.	Self-reported score (0.0-1.0).
Cognitive Dissonance	Agent's self-reported internal conflict when updating a belief.	Self-reported score (0.0-1.0).

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: The abstract and introduction claim we introduce a novel multi-agent evaluation framework and a suite of psychometric metrics. The body of the paper directly presents this framework and uses these metrics to analyze the emergent behaviors we claim to find (consensus-seeking, persona-induced profiles, etc.). Please refer to Sections 2,3 and Appendix B, C, D and E.

   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: A dedicated 'Limitations' section discusses the model-specific nature of our findings, the reliance on self-reported psychometric data from the LLMs, and the simplified nature of the text-based debate simulation compared to real-world interactions.

   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: This is an empirical paper focused on experimental results and analysis. We do not make any new theoretical claims that would require mathematical proofs.

   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: The 'Experimental Setup' section details the LLMs used, agent personas, incentives, moderator roles, debate structure, the source dataset (Change-My-View), and inference parameters. The metrics are also explicitly defined in Appendix E, providing a complete blueprint for replication. We will also open-source the codebase and an interactive website for reproducibility and transparency, and providing the community to use the website as platform to discover new emergent behaviors.

   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: We will provide open access to our full codebase, including the debate framework, analysis scripts, and generated data, in a supplemental repository upon publication. In addition, we will publish an interactive website for reproducibility and transparency, and providing the community to use the website as platform to discover new emergent behaviors.

   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: [NA]

   Justification: This research uses pre-trained, foundational LLMs without any fine-tuning or training. All relevant inference parameters, such as the model names and temperature, are specified in the 'Experimental Setup' section.

   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. **Experiment 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: The per-round semantic diversity plots include error bars representing the standard error of the mean. Furthermore, we employ Levene's test to formally assess the statistical significance of the variance in outcomes between different experimental conditions. Please refer to the Results and Appendix sections.

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: In the 'Experimental Setup' section, we have mentioned that we have used the HuggingFace Inference Provider for calling the LLMs through APIs. The costs are publicly accessible in the HuggingFace website. Each experiment took between 3 to 8 hours depending on the model, number of debate rounds and further evaluations.

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: Our research focuses on evaluating existing models to better understand their emergent behaviors, including risks like bias. The work uses a publicly available dataset and does not involve human subjects, aligning 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: The paper discusses the positive societal impact of creating better evaluation frameworks for safer AI agents. It also explicitly analyzes and discusses negative potentials, such as the observed emergent polarization and bias amplification in certain contexts.

    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: [NA]

    Justification: This paper introduces an evaluation framework and analysis code, not a new pre-trained model or a high-risk dataset. Therefore, specific safeguards for model or data release are not applicable.

    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 properly credit the creators of the language models used, the Change-My-View dataset via citation, and all key software libraries. The dataset is publicly available, and the models are used according to their respective licenses.

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: The code for our debate framework and analysis scripts, which are the primary assets of this paper, will be released with clear documentation (e.g., a README file) to ensure usability and reproducibility.

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: This research does not involve any crowdsourcing or human subjects; the experiments are conducted entirely with AI agents.

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: As no human subjects were involved in this study, Institutional Review Board (IRB) approval was not required.

    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: [Yes]

    Justification: The use of LLMs as the debater and moderator agents is the core methodological component of this research. The 'Experimental Setup' section explicitly details which models were used and how they were prompted to fulfill their roles.

    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.