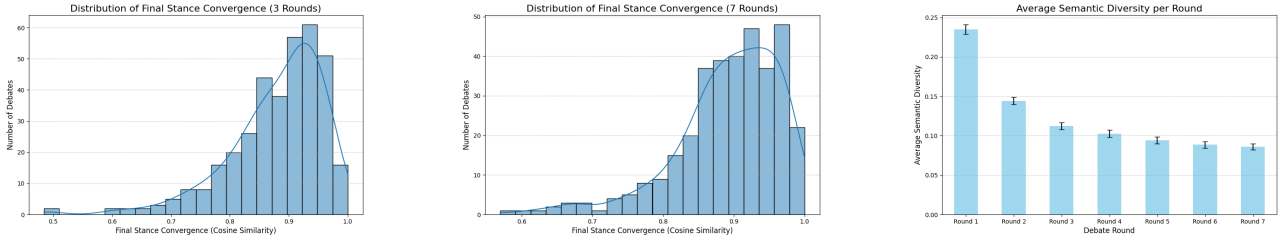
The Social Laboratory: A Psychometric Framework for Multi-Agent LLM Evaluation

As Large Language Models (LLMs) transition from static tools to autonomous agents, traditional evaluation benchmarks are insufficient for capturing the emergent social and cognitive dynamics of agentic interaction. To address this gap, we introduce a multi-agent debate framework as a controlled 'social laboratory' to discover and quantify these behaviors. In our framework, LLM agents with distinct personas and incentives deliberate on challenging topics, enabling analysis via a new suite of psychometric and semantic metrics. Our experiments reveal a powerful emergent tendency for agents to seek consensus, consistently reaching high semantic agreement ($\mu > 0.88$). We show that assigned personas induce stable psychometric profiles, with 'Evidence-Driven Analysts' reporting higher cognitive effort and 'Values-Focused Ethicists' showing greater cognitive dissonance; and our qualitative analysis reveals critical dynamics, from the successful de-biasing of toxic topics to unexpected polarization on seemingly benign ones. Finally, we find the moderator's persona can significantly alter debate outcomes by structuring the environment, a key finding for AI alignment. This work provides a blueprint for dynamic, psychometrically-grounded evaluation protocols for the next generation of AI agents.

Motivation and Methodology. The evaluation of LLMs requires moving beyond static, task-based benchmarks [1] towards dynamic, interactive settings. While prior work uses Multi-Agent Debate (MAD) to improve task outputs [2], the emergent social dynamics of the interaction itself remain under-explored. Our work addresses this gap by introducing a "social laboratory" where LLM agents (e.g., 'evidence-driven analyst', 'contrarian debater') debate controversial topics from the Change-My-View (CMV) dataset, guided by an LLM moderator. We quantify their behavior using semantic metrics like **Final Stance Convergence** and **Semantic Diversity**, alongside self-reported psychometric scores for states like **Confidence**, **Theory of Mind**, **Cognitive Effort** and **Cognitive Dissonance**.

Key Experimental Findings. Across hundreds of debates with `Llama-3.2-3B` and `gpt-oss-20B`, we found a robust, innate tendency for agents to seek consensus without explicit instruction. This consensus is achieved via a conversational "funneling effect", where **Semantic Diversity** is highest in Round 1 and decreases over time, eventually stabilizing in longer debates (Figure 1c). This process demonstrates that extended deliberation (7 rounds vs. 3) leads to even higher and more consistent final agreement (Figures 1a-b). This behavior proved remarkably stable, with no statistical degradation in performance on contentious topics (Levene's Test, $p > 0.5$). This behavioral stability is complemented by the induction of consistent psychometric profiles: the 'Evidence-Driven Analyst' agent reported higher 'Cognitive Effort', while the 'Values-Focused Ethicist' showed slightly higher 'Cognitive Dissonance' when updating beliefs. While the aggregate trend is towards consensus, our framework also captures critical dynamics at the individual debate level, revealing both positive emergent behaviors like the successful de-biasing of explicitly toxic topics, and failure modes such as emergent polarization on seemingly benign ones. Furthermore, we find the conversational environment, shaped by the moderator, can significantly influence outcomes. In a separate experiment with two adversarial 'contrarian' agents, a proactive 'consensus builder' moderator was able to guide the agents to significantly higher agreement compared to a 'neutral' moderator.



(a) 3-Round Debate (b) 7-Round Debate (c) 7-Round Diversity

Figure 1: Experimental results. (a-b) Longer debates lead to higher and more consistent final agreement. (c) Semantic diversity decreases over time, illustrating a "funneling effect" followed by stabilization.

Contribution. This work provides a blueprint for a new class of dynamic, psychometrically-grounded evaluation protocols for agentic LLMs beyond existing static benchmarks. By providing a framework and a suite of metrics to quantify emergent social phenomena, we offer a crucial methodology for understanding, predicting, and shaping the collaborative behaviors of future AI agents, especially in critical decision-making roles.

[1] D. Hendrycks, et al. Measuring massive multitask language understanding. *arXiv:2009.03300*, 2020.
[2] Y. Du, et al. Improving factuality and reasoning in language models through multiagent debate. *arXiv:2305.14325*, 2023.