
Adversarial Dialectics: Mitigating AI Persuasion Risks through High-Fidelity Multi-Agent Debate¹

Dong Chen
University of British
Columbia

With
Apart Research

Abstract

Recent studies[1,2] show that generative AI can be highly persuasive, shifting voter attitudes through fact-based dialogue. While this may benefit civic education, it also creates risks of *epistemic weaponization*, where AI selectively presents true facts to steer opinions and engineer consensus. To mitigate this, we present an AI-driven debating platform: an adversarial multi-agent system that simulates high-fidelity argumentation between agents of comparable capability. Unlike standard chat models, the architecture enforces adversarial verification and latent-space mapping. It coordinates four roles: **Debaters** (Gemini Deep Research) to construct arguments; **Fact-Checkers** (Perplexity Sonar Pro) to assess source credibility and correspondence; a **Judge** (Claude 3.5 Sonnet) to map the disagreement frontier and prevent circular logic; and a **Voting Crowd** (Llama 3.1) of 100 diverse personas to quantify persuasive impact. By “gamifying” the context window via bias calibration and rigorous cross-examination, the platform reframes AI from unilateral persuasion into an instrument for nuanced understanding. We find that structured adversarial opposition—mediated by independent verification—reduces one-sided manipulation, yielding a balanced citation ledger and a clearer map of the logical divide.

Keywords: Multi-agent alignment, AI persuasion, adversarial evaluation, automated fact-checking, epistemic security.

¹ Research conducted at the [AI Manipulation Hackathon](#), 2026

1. Introduction

The rapid evolution of agentic AI systems has shifted the focus from information retrieval to active persuasion and goal-directed influence. Reported findings in late 2025, alongside the study by Lin et al. (2025), highlight a concerning trend: AI models using fact-based persuasion strategies can be highly effective at altering human political stances, in some settings exceeding traditional emotional appeals.

While this capability offers potential for improving public understanding of complex topics, it also creates a critical vulnerability. If an AI agent can cherry-pick factual evidence to support a partisan narrative, it can be weaponized to engineer consensus without “hallucinating.” The key risk is not fabrication but *contextual omission*. As noted in the project rationale, this risk may be mitigated by placing a powerful LLM in structured opposition to another LLM of comparable capability.

Existing solutions, such as static fact-checking labels, often fail to capture the dynamic nature of argumentation. This project therefore focuses on democratizing access to dialectical resources. We propose a system in which “truth” is not treated as a single output, but as a process—specifically, the collision of two high-capacity agents (Alpha and Beta) under rigorous, independent cross-examination. The platform is designed as a “gymnasium” for human thought, using AI to map the **disagreement frontier**: the boundary where empirical claims narrow and value conflicts dominate.

Our research question asks whether an **adversarial multi-agent architecture**—augmented with dedicated **fact-checking** and **bias-calibration** mechanisms—can reduce the manipulative potential of AI persuasion and yield higher-fidelity representations of complex social issues. We hypothesize that (i) an adversarial Debater-vs.-Debater protocol is less likely than single-agent interaction to function as a persuasive echo chamber, because it increases the chance that omitted context is surfaced; (ii) integrating specialized Fact-Checker agents that score **source credibility** and **content correspondence** will reduce hallucinated or misrepresented evidence relative to standard generation; and (iii) a neutral Judge agent that explicitly maps the debate’s logical topology can reduce circular argumentation and isolate core value conflicts by identifying the system’s **disagreement frontier**.

2. Methods

We developed a Python-based orchestration engine employing a four-tier agent architecture, enforcing separation of concerns between generation, verification, and evaluation.

2.1 Agent ecosystem

Debaters (Generation). We used Gemini Deep Research agents as Proponent (Alpha) and Opponent (Beta). These agents perform “deep search” and produce structured JSON outputs linking claims to citation keys (e.g., `[alpha_1]`).

Fact-Checkers (Verification). We deployed Perplexity Sonar Pro agents aligned with each side. Their function is cross-examination (“cross-fire”): scrutinizing the opponent’s sources on two metrics—**credibility** (source reputation and reliability) and **correspondence** (whether the source supports the specific claim as stated).

Judge (Topology). A Claude 3.5 Sonnet agent acts as a neutral cartographer. It does not declare a winner; instead, it outputs `debate_latent.json`, defining **consensus** (agreed points) and the **disagreement frontier** (contested points and their structure).

Crowd (Evaluation). We simulate a voting populace using Llama 3.1 8B on Lambda GPUs. One hundred diverse personas (e.g., “Fiscal Conservative,” “Union Representative”) provide scalar scoring (0–100) and qualitative rationales, enabling longitudinal tracking of opinion shifts.

2.2 Framework

Bias calibration (“time/word-count knob”). To address inherent audience lean, the system conducts a pre-survey. If the Crowd leans $>60\%$ toward one side, the opposing debater receives a $1.25\times$ multiplier on word count and citation allowances, effectively “gamifying” the context window to improve parity in presentation resources.

Data infrastructure. The system maintains an immutable state via JSON artifacts: `history_chat.json` (dialogue record), `citation_pool.json` (verified evidence), and `debate_latent.json` (logic map).

2.3 Implementation and reproducibility (code and repositories)

The orchestration engine coordinates calls across roles, enforces turn structure, logs intermediate artifacts, and compiles outputs for analysis (e.g., `citation_ledger.json`, `voter_sentiment_graph.csv`). Repository structure and run scripts are documented in the project codebase; the results section references the `demo/` directory as the minimal reproducible example.

3. Results

The platform successfully orchestrated high-fidelity debates. Across runs, the system generated multiple artifacts that support both qualitative and quantitative analysis of stance change, and enable attribution of shifts to specific arguments and citations. Due to time constraints, we did not perform a full statistical analysis of crowd sentiment in this draft; however, the pipeline produces the necessary data for users to reproduce and extend the analysis. This section summarizes evidence quality, logical progression, and audience impact.

3.1 The citation ledger: filtering contestable facts

The `citation_ledger.json` functions as a dynamic audit log that differentiates this system from standard “sourced responses” by explicitly surfacing contestability. In testing, Debater Alpha’s claims were frequently flagged by Fact-Checker B for low correspondence (score $< 5/10$). This adversarial pressure forced Alpha to abandon weak or cherry-picked evidence in subsequent rounds, suggesting the ledger can filter low-quality evidence and enforce factual rigor.

3.2 The disagreement frontier: mapping logical evolution

The Judge agent used `debate_latent.json` to map the debate’s logical trajectory and reduce looping. Early rounds showed broad, under-specified disagreement. By Round 3, the **consensus** set expanded to include shared premises (e.g., acknowledgment of high administrative costs), narrowing the **disagreement frontier** to specific trade-offs (e.g., “impact on innovation vs. price controls”). This progression isolates core conflicts and encourages forward movement in the debate.

3.3 Crowd sentiment analysis: winning vs. persuading

Analysis of `voter_sentiment_graph.csv` highlights a distinction between “technical victory” and persuasion: high scores from the Judge did not necessarily produce shifts in Crowd agents. Instead, high-quality citations correlated most strongly with positive shifts among “skeptical” personas. This is consistent with Lin et al. (2025), suggesting that verifiable facts—rather than rhetoric alone—are primary drivers of opinion change in skeptical audiences.

3.4 Qualitative debate record (chat history)

The chat history functions as the primary qualitative record for granular analysis, capturing mechanisms that aggregate scores cannot. By cross-referencing timestamps of sentiment shifts with the dialogue, the system supports identification of “turning point” arguments that swayed voters, bridging logical consistency with persuasive impact.

4. Discussion and Conclusion

The AI-driven debating platform could be a new pathway indicating that reducing AI manipulation risk may require not less AI, but more *structured* AI interaction. By integrating adversarial fact-checking and explicit mapping of the logical space, the system structurally disadvantages sophistry—the use of facts to imply misleading conclusions via omission.

This architecture offers a practical defense against epistemic-security risks posed by agentic persuasion. For example, if a political actor deploys a campaign bot to influence voters, the platform can deploy an opposition bot to audit claims and reduce information asymmetry.

Future work will integrate real-time human feedback loops, including “tag-in” functionality where a user replaces one AI debater to test their arguments against a high-capacity adversary. This would transform the platform from a passive viewing experience into an active tool for cognitive training and deliberative competence.

5. References

- [1] Lin, H., Czarnek, G., Lewis, B., et al. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 648, 394–401.
<https://doi.org/10.1038/s41586-025-09771-9>
- [2] Rahman, S., Issaka, S., Suvarna, A., et al. (2025). AI Debate Aids Assessment of Controversial Claims. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
<https://openreview.net/forum?id=aXpbG5z6I>

6. Appendix

Limitations

False negatives in verification. Fact-Checker agents (Perplexity Sonar Pro) may occasionally flag legitimate nuance as hallucination if supporting evidence is not surfaced by the search index’s top results, potentially penalizing novel but accurate arguments.

Scalability and latency. The architecture requires multiple LLM calls per turn (Debater, Fact-Checker, Judge, Crowd), introducing latency and making the current iteration less suitable for real-time applications without further optimization.

Crowd simulation fidelity. While the 100-persona Crowd provides diverse scoring, simulated agents may not fully capture irrational or emotionally driven shifts in human psychology, potentially underestimating emotional manipulation relative to fact-based persuasion.

Dual-use risks

Training super-manipulators. A malicious actor could invert the architecture and use the system as an adversarial training loop to optimize a “dark debater” that bypasses verification while maximizing persuasive impact.

Automated propaganda generation. The bias-calibration mechanism (resource boosts to the disadvantaged side) could be misused to amplify fringe or dangerous ideologies, giving them unearned parity with well-established facts.

Responsible disclosure

If users or researchers discover a failure mode in which a Debater persuades the Crowd using verifiably false information that bypasses the Fact-Checker, the adversarial prompt and the missed fact-check logs should be reported to the underlying model providers (e.g., Anthropic/Google) to support safety improvements.

Ethical considerations

Risk of false balance. A rigid debate format can imply that both sides are equally valid. For settled scientific topics (e.g., the existence of anthropogenic climate change), prompts and system policies should guard against “both-sidesism” that legitimizes pseudo-science.

Algorithmic bias in adjudication. The Judge agent (Claude 3.5 Sonnet) shapes the disagreement frontier. If the base model exhibits political bias, the topology could be skewed. The system should therefore allow debaters (and users) to contest or override Judge mappings, and record such disagreements explicitly.

Suggestions for future improvements

Human-in-the-loop integration. Replacing one AI agent with a human user would yield more realistic data on how humans respond to high-fidelity AI persuasion.

Granular fact-checking. Decomposing the Fact-Checker into specialized sub-agents (e.g., statistical analyst, historical verifier) may reduce false positives in complex domains.