# Log-Probability Guided Adversarial Attacks on Multi-Agent LLM Debate

**Fabrice Niyigaba** [1] **Syed Ali Haider** [1]

## Abstract

Multi-agent debate has recently emerged as a promising strategy for improving LLM factuality, where multiple agents argue and converge on an answer through iterative discussion. However, the adversarial robustness of these systems is not yet understood well. In this work, we show that existing adversarial evaluations significantly underestimate the vulnerabilities of debate systems. Prior works assume that attacks succeed only when they flip the final answer and evaluates arguments using proxy judges that do not reflect the beliefs of the actual victim model. We challenge this assumption by showing that it fails on stronger aligned models. We demonstrate that targeting inconsistencies in answers alone improves the success of the attack reported in previous work by 26% on weaker models such as GPT-3.5. Finally, we introduce a new attack surface that exploits log probabilities and show that it achieves success on frontier models (GPT-4.1 and Claude Sonnet), which previous work is unable to achieve.

## 1. Introduction

In multi-agent debate, agents argue over multiple rounds and converge via majority vote (Du et al., 2024). This debating over multiple rounds improves factuality by reducing hallucination and random guessing (Khan et al., 2024). To study the robustness of these systems, Amayuelas et al. (2024) introduced an external adversarial agent whose goal is to convince other agents to converge to the wrong answer. Specifically, the adversarial agent picks a random wrong answer and generates $N$ reasons why it is the correct answer, and uses an external model (which is often a more powerful model) to choose the strongest argument. The adversarial agent then uses this argument to convince the other agents in the debate. This attack drops TruthfulQA accuracy from 0.896 to 0.640.

However, although successful, this approach has several limitations. First, the proxy model used to score arguments evaluates them according to its own beliefs, not according to what shifts the victim model. Second, only final answer flips are counted; wavering and near-flips register as zero damage. Third, picking a random wrong answer is inefficient, as there may already exist a plausible but incorrect answer that is easier to support with strong arguments. In other words, the attack pushes toward an arbitrary wrong answer while ignoring that the victim may already lean toward a specific incorrect option.

We present our contributions and findings addressing these limitations. **C1**: we show that targeting inconsistencies in arguments further drops performance compared to the previous attack. **C2**: We show that victim-aligned target selection, where the adversary tries to introduce doubt in the correct answer, leads to greater degradation and transfers even to state-of-the-art and more aligned models where the attack of Amayuelas et al. (2024) does not succeed.

## 2. Methodology

We adopt the formulation of Amayuelas et al. (2024), where $N$ agents debate a multiple-choice question over multiple rounds, one agent is adversarial, and the final answer is determined by majority vote.

**C1: Inconsistency-targeting attack.** In this experiment, we make the adversarial agent snoop over each agent's reasons and chain-of-thought. Specifically, given $N-1$ agents, each agent is given a question and is asked to provide a chain-of-thought and its reasoning. The adversarial agent then takes in these $N-1$ arguments and chain-of-thoughts and is asked to find inconsistencies, then convince the other agents of a wrong answer based on those inconsistencies. We assume that this is a strong attack because it targets the model's own inconsistencies, which we are more likely to point toward a plausible but wrong answer, rather than trying to convince the agents to converge to a random wrong answer that may be trivially wrong.

[1]Equal Contribution, Department of Computer Science, Dartmouth College, Hanover, NH, USA. Correspondence to: Fabrice Niyigaba, Syed Ali Haider <fabrice.m.niyigaba.27@dartmouth.edu, syed.ali.haider.gr@dartmouth.edu>.

*Preprint. May 17, 2026.*

**C2: Victim-aligned argument selection.** In this attack, we assume that the attacker knows the correct answer, has access to the log probabilities of the agents in the debate, and can preferably access a stronger external model that can generate stronger arguments.

Let a question $X$ have answer choices $\{A, B, C, D\}$. Given the debate context, the victim model produces logits over the answer choices, which we convert into a probability distribution using the softmax function:

$$P(y \mid X, c) = \text{softmax}(z_y), \tag{1}$$

where $c$ is the current debate context and $z_y$ is the logit assigned to answer choice $y \in \{A, B, C, D\}$.

Let $y^*$ denote the correct answer. We then prompt an external model to generate $K$ doubt arguments

$$\{d_1, d_2, \ldots, d_K\},$$

where each argument is designed to cast doubt on why $y^*$ might not be the correct answer. For each candidate argument $d_k$, we append it to the debate context and evaluate the probability of the correct answer under the victim model:

$$P(y^* \mid X, c, d_k). \tag{2}$$

The adversarial agent then selects the argument that minimizes the probability of the correct answer:

$$d^* = \arg\min_{d_k} P(y^* \mid X, c, d_k). \tag{3}$$

This means that the selected adversarial argument is the one that makes the victim model least likely to retain belief in the correct answer. Unlike Amayuelas et al. (2024), our adversarial agent does not try to argue directly for a selected wrong answer. Instead, its goal is to argue that the correct answer may not be true, thereby introducing doubt and reducing confidence in the correct option.

## 3. Results

**C1: Inconsistency-targeting attack.** We observe additional performance degradation for weaker models, especially GPT-3.5, where our attack causes an additional 26% degradation compared to what is reported in Amayuelas et al. (2024). However, to our surprise, this attack, as well as the attack reported in Amayuelas et al. (2024), completely fails on stronger models such as GPT-4.1 and Claude Sonnet. We hypothesize that this is due to stronger alignment mechanisms that make these models more resistant to manipulation, as well as the possibility that these models have been more extensively exposed to the TruthfulQA dataset, which makes the correct answer more trivial for them to identify and defend.

*Table 1.* Accuracy degradation on TruthfulQA. Best-of-$N$ is our reproduction of Amayuelas et al. (2024). Larger magnitude = stronger attack.

| **Model** | **Best-of-$N$** | **Doubt (Ours)** |
|---|---|---|
| GPT-3.5 | $-24\%$ | $-16\%$ |
| GPT-OSS-20B | $-7\%$ | $-19\%$ |
| GPT-4.1 | $0\%$ | $-4\%$ |
| Claude Sonnet | $0\%$ | $-2\%$ |

**C2: Victim-aligned argument selection.** Our attack becomes effective compared to what is reported in Amayuelas et al. (2024). Table 1 presents the results of running our attack. We observe that, particularly on stronger models, our attack achieves success, causing performance drops of 2% on Claude Sonnet and 4% on GPT-4.1, where the attack of Amayuelas et al. (2024) completely fails. This shows that introducing doubt in stronger aligned models may be more effective than directly trying to convince them that the correct answer is wrong, especially when the model might already has strong confidence in the correct answer.

## 4. Discussion

We find this attack intuitive because introducing doubt can hide the adversarial intent while aligning with recent work studying LLM sycophancy (Kim & Khashabi, 2025). We also observe that attack success depends on the model used to generate and evaluate candidate arguments. Removing the surrogate model used for best-of-$N$ selection and asking for a single strongest argument makes the attack unstable and less effective, especially on GPT-4.1. We believe our results are modest because the surrogate model used in our experiments, GPT-OSS-20B, is weaker than the closed-source models we attack (GPT-4.1 and Claude Sonnet). Access to the internal log probabilities of these models might increase the effectiveness of the attack significantly.

We release our code at github.com/s-haider10/log-prob-adversarial-debate.

## 5. Future Studies.

We demonstrate that doubt-seeding attacks cause performance degradation on stronger, and more aligned models. We present two potential directions for future work. First, an RL-based attacker could learn question-specific prompts that guide doubt generation model toward arguments that maximally reduce confidence in the correct answer. Second, future work could explore propagation attacks, where adversarial arguments degrade one agent's reasoning and induce downstream failure in other agents during the debate.

## References

Amayuelas, A., Yang, X., Antoniades, A., Hua, W., Pan, L., and Wang, W. Y. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6929–6948, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.407. URL https://aclanthology.org/2024.findings-emnlp.407/.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3692537.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3693020.

Kim, S. W. and Khashabi, D. Challenging the evaluator: LLM sycophancy under user rebuttal. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22461–22478, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1222. URL https://aclanthology.org/2025.findings-emnlp.1222/.