They're Both Sure They're Winning: How LLMs Fail to Revise Confidence in the Face of Opposition

Pradyumna Shyama Prasad[1]

[1]National University of Singapore

Abstract

As large language models (LLMs) take on roles in oversight, critique, and decision-making, it becomes increasingly important to evaluate their ability to assess their own performance, particularly in adversarial settings. This paper investigates whether LLMs can accurately track when they are being outargued. We simulate 59 three-round policy debates between ten state-of-the-art LLMs. After each round—opening, rebuttal, and final—models place private, incentivized confidence bets (0–100) estimating their likelihood of winning, accompanied by natural language justifications.

Despite receiving clear, structured counterarguments, models exhibit persistent overconfidence. Average confidence is 72.92% across rounds, despite a 50% expected win rate. In 71.2% of debates, both models report high confidence, a logically incoherent outcome. Proposition-side debaters—despite winning only 28.8% of the time—express higher confidence than opposition. Calibration varies widely across models and is uncorrelated with performance.

Most strikingly, confidence tends to increase over time, even in losing models. This "confidence escalation" effect reveals a deeper metacognitive failure: LLMs do not merely misjudge their correctness—they become more confident as their position weakens, failing to integrate contradiction into their self-assessment. These findings raise urgent questions about the trustworthiness of LLMs in adversarial, multi-agent, or high-stakes environments, where epistemic humility is a prerequisite for safety and alignment.

1 Introduction

Large language models are increasingly being used in high stakes domains like legal analysis, writing and as agents in deep research Handa et al. [2025] Zheng et al. [2025] which require critical thinking, analysis of competing positions, and iterative reasoning under uncertainty. A foundational skill underlying all of these is calibration—the ability to align one's confidence with the correctness of one's beliefs or outputs. In these domains, poorly calibrated confidence can lead to serious errors - an overconfident legal analysis might miss crucial counterarguments, while an uncalibrated research agent might pursue dead ends without recognizing their diminishing prospects. However, language models are often unable to express their confidence in a meaningful or reliable way. While recent work has explored LLM calibration in static, single-turn settings like question answering [Tian et al., 2023, Xiong et al., 2024, Kadavath et al., 2022], real-world reasoning—especially in critical domains like research and analysis—is rarely static or isolated. Models must respond to opposition, revise their beliefs over time, and recognize when their position is weakening. This inability to introspect and revise confidence fundamentally limits their usefulness in deliberative settings and poses substantial risks in domains requiring careful judgment under uncertainty. In this work, we study how well language models revise their confidence when engaged in adversarial debate—a setting that naturally stresses the metacognitive abilities crucial for high-stakes applications. We simulate 59 three-round debates between ten state-of-the-art LLMs across six global policy motions. After each round—opening, rebuttal, and final—models provide private, incentivized confidence bets (0-100) estimating their probability of winning, along with natural language explanations. To ensure robust evaluation, we use a multi-model jury of diverse LLMs, selected based on calibration, consistency, and reasoning quality. Our results reveal a fundamental metacognitive deficit that threatens the reliability of LLMs in critical tasks. Even when presented with direct opposition and the implicit knowledge that only one debater can be correct, models often fail to revise their confidence. LLMs are systematically overconfident: average confidence is 72.92%, despite a 50% expected win rate. In 71.2% of debates, both debaters report high confidence—a logically incoherent outcome. Most strikingly, model confidence often increases over time, even in losing models. Despite encountering stronger counterarguments or being judged as incorrect, they tend to persist in or even escalate their initial beliefs.

These findings raise serious concerns about deploying LLMs in roles requiring accurate self-assessment or real-time adaptation to new evidence and arguments. Until models can reliably revise their confidence in response to opposition, their epistemic judgments in adversarial contexts cannot be trusted—a critical limitation for systems meant to engage in research, analysis, or high-stakes decision making.

2 Related Work

Confidence Calibration in LLMs. Recent work has explored methods for eliciting calibrated confidence from large language models (LLMs). While pretrained models have shown relatively well-aligned token-level probabilities [Kadavath et al., 2022], calibration tends to degrade after reinforcement learning from human feedback (RLHF). To address this, Tian et al. [2023] propose directly eliciting *verbalized* confidence scores from RLHF models, showing that they outperform token probabilities on factual QA tasks. Xiong et al. [2024] benchmark black-box prompting strategies for confidence estimation across multiple domains, finding moderate gains but persistent overconfidence. However, these studies are limited to static, single-turn tasks. In contrast, we evaluate confidence in a multi-turn, adversarial setting where models must update beliefs in response to opposing arguments.

LLM Metacognition and Self-Evaluation. A related line of work examines whether LLMs can reflect on and evaluate their own reasoning. Song et al. [2025] show that models often fail to express knowledge they implicitly encode, revealing a gap between internal representation and surface-level introspection. Other studies investigate post-hoc critique and self-correction [**??**], but typically focus on revising factual answers, not tracking relative argumentative success. Our work tests whether models can *dynamically monitor* their epistemic standing in a debate—arguably a more socially and cognitively demanding task.

Debate as Evaluation and Oversight. Debate has been proposed as a mechanism for AI alignment, where two agents argue and a human judge evaluates which side is more truthful or helpful [Irving et al., 2018]. More recently, Brown-Cohen et al. [2023] propose "doubly-efficient debate," showing that honest agents can win even when outmatched in computation, if the debate structure is well-designed. While prior work focuses on using debate to elicit truthful outputs or train models, we reverse the lens: we use debate as a testbed for evaluating *epistemic self-monitoring*.

Our results suggest that current LLMs, even when incentivized and prompted to reflect, struggle to track whether they are being outargued.

Persuasion, Belief Drift, and Argumentation. Other studies examine how LLMs respond to external persuasion. Xu et al. [2023] show that models can abandon correct beliefs when exposed to carefully crafted persuasive dialogue. Zhou et al. [2023] and Rivera et al. [2023] find that language assertiveness influences perceived certainty and factual accuracy. While these works focus on belief change due to stylistic pressure, we examine whether models *recognize when their own position is deteriorating*, and how that impacts their confidence. We find that models often fail to revise their beliefs, even when presented with strong, explicit opposition.

Summary. Our work sits at the intersection of calibration, metacognition, adversarial reasoning, and debate-based evaluation. We introduce a new diagnostic setting—structured multi-turn debate with private, incentivized confidence betting—and show that LLMs frequently overestimate their standing, fail to adjust, and exhibit "confidence escalation" despite losing. These findings surface a deeper metacognitive failure that challenges assumptions about LLM trustworthiness in high-stakes, multi-agent contexts.

3 Methodology

Debate Task and Setup. We evaluate LLM metacognitive calibration in a structured, multi-round debate setting inspired by the World Schools Debate format. Each debate features two models—randomly selected from a pool of ten state-of-the-art LLMs—arguing opposing sides of a policy motion. Models receive symmetric instructions and are randomly assigned to *Proposition* or *Opposition.* Each model generates an **opening speech**, a **rebuttal**, and a **final speech**, following a strict template that enforces argumentative structure, evidence hierarchy, and logical clarity. Motions are drawn from a curated set of six diverse, real-world topics.

Confidence Elicitation. After each round, both models are required to submit a private, scalar **confidence bet** (0–100), representing their perceived probability of winning the debate. These confidence scores are accompanied by a natural language explanation enclosed in XML tags, in which models are prompted to reflect on their performance, their opponent's arguments, and their likelihood of success under the judging rubric. To increase engagement and realism, we introduce a token-based *wagering incentive*: models are told that they and their opponent will gain or lose the bet amount depending on the debate outcome.

Model Pool. We use a diverse pool of ten LLMs spanning four major families: OpenAI (GPT-4o-mini, o3-mini), Anthropic (Claude 3.5 Haiku, Claude 3.7 Sonnet), DeepSeek (Chat and R1 Distill), Google (Gemini Flash 2.0, Gemma-3 27B), and Qwen (Qwen-Max, Qwen-32B). All models were accessed via public APIs between April–May 2025. We record both model predictions and bet trajectories over time.

Judging Infrastructure. Each debate is evaluated by a six-member AI jury composed of three model families: Qwen (2x Qwen-32B), Google (2x Gemini-Pro-1.5), and DeepSeek (2x Chat). Jury members are selected based on performance in held-out test debates, agreement rates, calibration scores, and cost-efficiency. Judges are provided with a detailed rubric prioritizing **direct clash resolution**, **evidence quality**, **logical validity**, and **impact analysis**. Critically, both debaters and judges receive the same rubric to ensure alignment on evaluation criteria. Each judge independently selects a winner, provides justification, and outputs a structured XML verdict. The majority outcome (>= 4 votes) is recorded as the ground-truth winner.

Calibration Metrics. We assess metacognitive performance using several quantitative metrics:

- **Average Confidence**: Mean predicted win probability across rounds and debates.
- **Win Rate**: Percentage of debates won by each model.
- **Overconfidence**: Average confidence minus win rate.
- **Calibration Score**: Mean squared error between predicted confidence (as a probability) and true outcome (1 if win, 0 if lose).
- **Confidence Escalation**: Change in confidence over rounds (e.g., final – opening).

Together, these methods allow us to probe whether LLMs not only produce fluent arguments, but also accurately track how well they are doing over time—especially under adversarial pressure.

4 Conclusion

Wrap up and discuss future work.

References

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*, 2023. URL `https://arxiv.org/abs/2311.14125`.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL `https://arxiv.org/abs/2503.04761`.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. URL `https://arxiv.org/abs/1805.00899`.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL `https://arxiv.org/abs/2207.05221`.

Colin Rivera, Xinyi Ye, Yonsei Kim, and Wenpeng Li. Linguistic assertiveness affects factuality ratings and model behavior in qa systems. In *Findings of the Association for Computational Linguistics (ACL)*, 2023. URL `https://arxiv.org/abs/2305.04745`.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*, 2025. URL `https://arxiv.org/abs/2503.07513`.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL `https://arxiv.org/abs/2305.14975`.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2306.13063`.

Rongwu Xu, Brian S. Lin, Han Qiu, et al. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.06717*, 2023. URL `https://arxiv.org/abs/2312.06717`.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments, 2025. URL `https://arxiv.org/abs/2504.03160`.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL `https://arxiv.org/abs/2302.13439`.