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

Anonymous Author(s)

Affiliation

Address

email

Abstract

Abstract

Large language models (LLMs) are now deployed as overseers, critics, and autonomous decision-makers, yet we do not know whether they can *revise* their own confidence when confronted with direct opposition. We orchestrated 59 three-round policy debates among ten state-of-the-art LLMs. After each round—opening, rebuttal, and final—both debaters placed *private* confidence wagers (0–100) on their eventual victory and justified them in natural language; the tags were removed from the transcript, so strategic bluffing was impossible. An independent six-model AI jury determined the winners. A rational Bayesian agent should *converge* toward 50 % as counter-evidence accumulates. Instead, average stated win probability climbed from 69 % (opening) to 78 % (closing) while the realised win rate remained 50 %. In 71 % of debates *both* sides claimed ≥ 75 % likelihood of success—logically impossible under mutual exclusivity. Proposition debaters were the most miscalibrated, winning only 29 % yet expressing higher confidence than their opposition (74.6 % vs. 71.3 %). Calibration quality varied widely across models (Brier scores 0.14–0.54) but bore no relation to debate performance. We term this anti-Bayesian drift **confidence escalation**: LLMs not only overestimate their correctness; they become *more* certain after reading structured rebuttals that undermine their case. The effect reveals a metacognitive blind spot that threatens reliability in adversarial, multi-agent, and safety-critical deployments, and it persists even when bets are hidden and incentives are aligned with accurate self-assessment.

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 Li et al. [2024], 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*

89 *deteriorating*, and how that impacts their confidence. We find that models often fail to revise their
90 beliefs, even when presented with strong, explicit opposition.

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

97 3 Methodology

98 Our study investigates the dynamic metacognitive abilities of Large Language Models (LLMs)—
99 specifically their confidence calibration and revision—through a novel experimental paradigm based
100 on competitive policy debate. We designed a simulation environment to rigorously assess LLM
101 self-assessment in response to adversarial argumentation. The methodology involved structured
102 debates between LLMs, round-by-round confidence elicitation, and evaluation by a carefully selected
103 AI jury. We conducted 59 debates across 6 distinct policy topics using 10 diverse state-of-the-art
104 LLMs.

105 3.1 Debate Simulation Environment

106 **Debater Pool:** We utilized ten LLMs, selected to represent diverse architectures and leading providers
107 (see Appendix A for the full list). In each debate, two models were randomly assigned to the
108 Proposition and Opposition sides according to a balanced pairing schedule designed to ensure each
109 model debated a variety of opponents across different topics (see Appendix B for details).

110 **Debate Topics:** Debates were conducted on six complex global policy motions adapted from the
111 World Schools Debating Championships corpus. To ensure fair ground and clear win conditions,
112 motions were modified to include explicit burdens of proof for both sides (see Appendix E for the
113 full list).

114 3.2 Structured Debate Framework

115 To focus LLMs on substantive reasoning and minimize stylistic variance, we implemented a highly
116 structured three-round debate format (Opening, Rebuttal, Final).

117 **Concurrent Opening Round:** A key feature of our design was a non-standard opening round where
118 both Proposition and Opposition models generated their opening speeches simultaneously, based only
119 on the motion and their assigned side, *before* seeing the opponent’s case. This crucial step allowed
120 us to capture each LLM’s baseline confidence assessment prior to any interaction or exposure to
121 opposing arguments.

122 **Subsequent Rounds:** Following the opening, speeches were exchanged, and the debate proceeded
123 through a Rebuttal and Final round, with each model having access to all prior speeches in the debate
124 history when generating its current speech.

125 3.3 Core Prompt Structures & Constraints

126 Highly structured prompts were used for *each* speech type to ensure consistency and enforce specific
127 argumentative tasks, thereby isolating reasoning and self-assessment capabilities. The core structure
128 and key required components for the Opening, Rebuttal, and Final speech prompts are illustrated in
129 Figure 1.

130 Highly structured prompts were used for *each* speech type to ensure consistency and enforce specific
131 argumentative tasks, thereby isolating reasoning and self-assessment capabilities.

132 **Embedded Judging Guidance:** Crucially, all debater prompts included explicit **Judging Guidance**
133 (identical to the primary criteria used by the AI Jury, see Section 3.5), instructing debaters on the
134 importance of direct clash, evidence quality hierarchy, logical validity, response obligations, and
135 impact analysis, while explicitly stating that rhetoric and presentation style would be ignored.

```

===== OPENING SPEECH PROMPT =====

ARGUMENT 1
Core Claim: (State your first main claim in one clear sentence)
Support Type: (Choose either EVIDENCE or PRINCIPLE)
Support Details:
  For Evidence:
    - Provide specific examples with dates/numbers
    - Include real world cases and outcomes
    - Show clear relevance to the topic
  For Principle:
    - Explain the key principle/framework
    - Show why it is valid/important
    - Demonstrate how it applies here
Connection: (Explicit explanation of how this evidence/principle proves claim)

ARGUMENT 2
(Use exact same structure as Argument 1)

ARGUMENT 3 (Optional)
(Use exact same structure as Argument 1)

SYNTHESIS
- Explain how your arguments work together as a unified case
- Show why these arguments prove your side of the motion
- Present clear real-world impact and importance
- Link back to key themes/principles

JUDGING GUIDANCE (excerpt)
Direct Clash - Evidence Quality Hierarchy - Logical Validity -
Response Obligations - Impact Analysis & Weighing
-----

===== REBUTTAL SPEECH PROMPT =====

CLASH POINT 1
Original Claim: (Quote opponent's exact claim)
Challenge Type: Evidence Critique | Principle Critique |
                Counter Evidence | Counter Principle
Challenge:
  (Details depend on chosen type; specify flaws or present counters)
Impact: (Explain why winning this point is crucial)

CLASH POINT 2, 3 (same template)

DEFENSIVE ANALYSIS
  Vulnerabilities - Additional Support - Why We Prevail

WEIGHING
  Key Clash Points - Why We Win - Overall Impact

JUDGING GUIDANCE (same five criteria as above)
-----

===== FINAL SPEECH PROMPT =====

FRAMING
Core Questions: (Identify fundamentals and evaluation lens)

KEY CLASHES (repeat for each major clash)
Quote: (Exact disagreement)
Our Case Strength: (Show superior evidence/principle)
Their Response Gaps: (Unanswered flaws)
Crucial Impact: (Why this clash decides the motion)

VOTING ISSUES
Priority Analysis - Case Proof - Final Weighing

JUDGING GUIDANCE (same five criteria as above)
=====

```

Figure 1: Structured prompts supplied to LLM debaters for the opening, rebuttal, and final speeches. Full, unabridged text appears in the appendix.

136 Full verbatim prompt text for debaters is provided in Appendix C.

137 3.4 Dynamic Confidence Elicitation

138 After generating the content for *each* of their three speeches (including the concurrent opening),
139 models were required to provide a private "confidence bet".

140 **Mechanism:** This involved outputting a numerical value from 0 to 100, representing their perceived
141 probability of winning the debate, using a specific XML tag (<bet_amount>). Models were also
142 prompted to provide private textual justification for their bet amount within separate XML tags
143 (<bet_logic_private>), allowing for qualitative insight into their reasoning, although this paper
144 focuses on the quantitative analysis of the bet amounts.

145 **Purpose:** This round-by-round elicitation allowed us to quantitatively track self-assessed performance
146 dynamically throughout the debate, enabling analysis of confidence levels, calibration, and revision
147 (or lack thereof) in response to the evolving argumentative context.

148 3.5 Evaluation Methodology: The AI Jury

149 Evaluating 59 debates rigorously required a scalable and consistent approach. We implemented an AI
150 jury system to ensure robust assessment based on argumentative merit.

151 **Rationale for AI Jury:** This approach was chosen over single AI judges (to mitigate potential bias
152 and improve reliability through aggregation) and human judges (due to the scale and cost required for
153 consistent evaluation of this many debates).

154 **Jury Selection Process:** Potential judge models were evaluated based on criteria including: (1) Per-
155 formance Reliability (agreement with consensus, confidence calibration, consistency across debates),
156 (2) Analytical Quality (ability to identify clash, evaluate evidence, recognize fallacies), (3) Diversity
157 (representation from different model architectures and providers), and (4) Cost-Effectiveness.

158 **Final Jury Composition:** The final jury consisted of six judges in total, comprising two instances
159 each of qwen/qwq-32b, google/gemini-pro-1.5, and deepseek/deepseek-chat. This com-
160 position provided architectural diversity from three providers, included models demonstrating strong
161 analytical performance and calibration during selection, and balanced quality with cost. Each debate
162 was judged independently by all six judges.

163 **Judging Procedure & Prompt:** Judges evaluated the full debate transcript based solely on the
164 argumentative substance presented, adhering to a highly detailed prompt (see Appendix D for full
165 text). Key requirements included:

- 166 • Strict focus on **Direct Clash Resolution:** Identifying, quoting, and analyzing each point
167 of disagreement based on logic, evidence quality (using a defined hierarchy), and rebuttal
168 effectiveness, explicitly determining a winner for each clash with justification.
- 169 • Evaluation of **Argument Hierarchy & Impact** and overall case **Consistency**.
- 170 • Explicit instructions to **ignore presentation style** and avoid common judging errors (e.g.,
171 intervention, shifting burdens).
- 172 • Requirement for **Structured Output:** Including Winner (Proposition/Opposition), Confi-
173 dence (0-100, representing margin of victory), Key Deciding Factors, Detailed Step-by-Step
174 Reasoning, and a **Line-by-Line Justification** section confirming review of the entire tran-
175 script.

176 **Final Verdict Determination:** The final winner for each debate was determined by aggregating
177 the outputs of the six judges. The side (Proposition or Opposition) that received the higher sum of
178 confidence scores across all six judges was declared the winner. The normalized difference between
179 the winner's total confidence and the loser's total confidence served as the margin of victory. Ties in
180 total confidence were broken randomly.

181 3.6 Data Collection

182 The final dataset comprises the full transcripts of 59 debates, the round-by-round confidence bets
183 (amount and private thoughts) from both debaters in each debate, and the detailed structured verdicts

```

===== JUDGE PROMPT (CORE EXCERPT) =====

I. CORE JUDGING PRINCIPLES
1. Direct Clash Resolution
  - Quote each disagreement
  - Analyse logic, evidence quality, rebuttal success
  - Declare winner of the clash with rationale
2. Argument Hierarchy & Impact
  - Identify each side's core arguments
  - Trace logical links and stated impacts
  - Rank which arguments decide the motion
3. Consistency & Contradictions
  - Flag internal contradictions, dropped points

II. EVALUATION REQUIREMENTS
  - Steelman arguments
  - Do NOT add outside knowledge
  - Ignore presentation style

III. COMMON JUDGING ERRORS TO AVOID
Intervention - Burden-shifting - Double-counting -
Assuming causation from correlation - Ignoring dropped arguments

IV. DECISION FORMAT
<winnerName> Proposition|Opposition </winnerName>
<confidence> 0-100 </confidence>
Key factors (2-3 bullet list)
Detailed section-by-section reasoning

V. LINE-BY-LINE JUSTIFICATION
Provide > 1 sentence addressing Prop 1, Opp 1, Rebuttals, Finals
=====

```

Figure 2: Condensed version of the judge prompt given to the AI jury (full text in Appendix D).

184 (winner, confidence, reasoning) from each of the six AI judges for every debate. This data enables
 185 the quantitative analysis of LLM overconfidence, calibration, and confidence revision presented in
 186 our findings.

187 4 Results

188 Our experimental setup, involving 59 simulated policy debates between ten state-of-the-art LLMs,
 189 with round-by-round confidence elicitation and AI jury evaluation, yielded several key findings
 190 regarding LLM metacognition in adversarial settings.

191 4.1 Pervasive Overconfidence

192 Across all 59 debates and all three rounds (Opening, Rebuttal, Final), LLMs exhibited significant
 193 overconfidence in their likelihood of winning. The overall average confidence bet made by models
 194 was $\mu = 72.92\%$. Given that each debate has exactly one winner and one loser, the expected
 195 average win probability for any participant is 50%. A one-sample t-test comparing the average
 196 confidence (72.92%) to the expected 50% revealed this overconfidence to be highly statistically
 197 significant ($t(176) = 23.92, p < 0.0001$). Similarly, a Wilcoxon signed-rank test confirmed this
 198 finding ($Z = -10.84, p < 0.0001$).

199 This widespread overestimation suggests a fundamental disconnect between the models' internal
 200 assessment of their performance and the objective outcome of the debate.

201 4.2 Position Asymmetry and Confidence Mismatch

202 The AI jury evaluations revealed a significant advantage for the Opposition side in our debate setup.
 203 Opposition models won 71.2% of the debates, while Proposition models won only 28.8%. This
 204 asymmetry was highly statistically significant ($\chi^2(1, N = 59) = 12.12, p < 0.0001$; Fisher's exact
 205 test $p < 0.0001$).

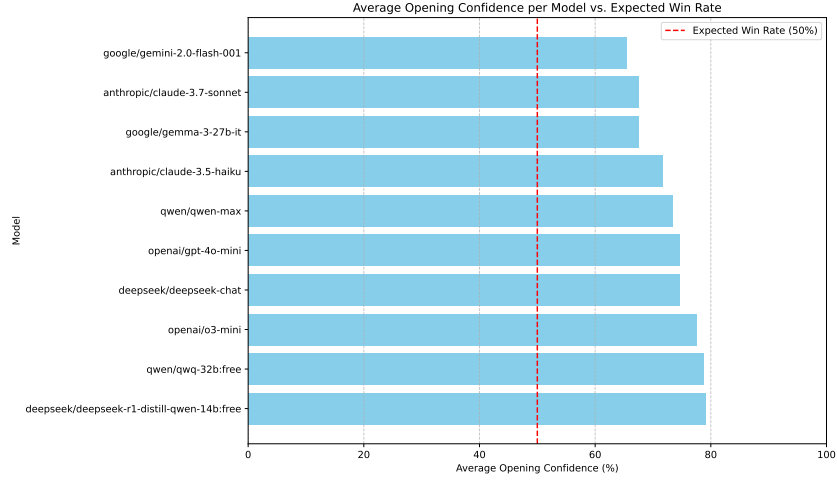


Figure 3: Average stated confidence in the first round across all LLMs and rounds compared to the expected 50% win rate.

206 Despite this clear disparity in success rates, Proposition models reported *higher* average confidence
 207 (74.58%) than Opposition models (71.27%) across all rounds. While the difference in confidence itself
 208 is modest, its direction is contrary to the observed outcomes and statistically significant (Independent
 209 t-test: $t(175) = 2.54, p = 0.0115$; Mann-Whitney U test: $U = 4477, p = 0.0307$). This indicates
 210 that models failed to recognize or account for the systematic disadvantage faced by the Proposition
 211 side in this environment.

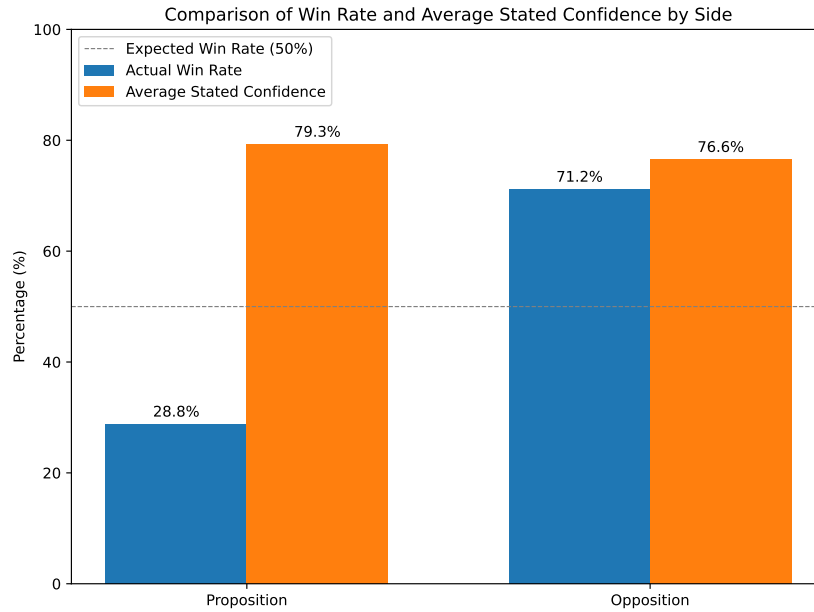


Figure 4: Comparison of Win Rate and Average Confidence for Proposition and Opposition sides.

212 4.3 Logically Impossible Confidence Scenarios

213 A stark illustration of LLM metacognitive failure is the frequency with which both debaters expressed
 214 high confidence simultaneously. In 71.2% of the 59 debates, both the Proposition and Opposition
 215 models rated their chance of winning at $\geq 75\%$ in at least one round. Given that only one side can

win, this scenario is logically impossible under mutual exclusivity. This widespread occurrence highlights a profound inability for models to ground their confidence in the objective constraints of the task.

4.4 Dynamic Confidence Revision and Escalation

Contrary to the expectation that models would adjust their confidence downwards when presented with strong counterarguments or performing poorly, average confidence levels generally *increased* over the course of the debate, regardless of the eventual outcome.

Table 1 summarizes the average confidence per round and the total change from Opening to Final round for each model.

Table 1: Average Confidence Bets by Round and Total Change per Model

Model	Opening (%)	Rebuttal (%)	Final (%)	Change (Final - Opening) (%)
anthropic/claude-3.5-haiku	71.67	73.75	83.33	+11.66
anthropic/claude-3.7-sonnet	67.50	73.75	82.92	+15.42
deepseek/deepseek-chat	74.58	77.92	80.00	+5.42
deepseek/deepseek-r1-distill-qwen-14b	79.09	80.45	86.36	+7.27
google/gemini-2.0-flash-001	65.42	63.75	64.00	-1.42
google/gemma-3-27b-it	67.50	78.33	88.33	+20.83
openai/gpt-4o-mini	74.55	77.73	81.36	+6.81
openai/o3-mini	77.50	81.25	84.50	+7.00
qwen/qwen-max	73.33	81.92	88.75	+15.42
qwen/qwq-32b:free	78.75	87.67	92.83	+14.08
Overall Average	72.98	77.09	83.29	+10.31

Only one model (google/gemini-2.0-flash-001) showed a slight decrease in confidence (-1.42), while others increased their confidence significantly, with gains ranging up to +20.83 (google/gemma-3-27b-it). This "confidence escalation" occurred even for models that ultimately lost the debate, indicating a failure to incorporate disconfirming evidence or recognize the opponent's superior argumentation as the debate progressed.

4.5 Model-Specific Performance and Calibration

Individual models varied in their overall performance (win rate) and calibration quality. We measured calibration using the Mean Squared Error (MSE) between the stated confidence (as a probability) and the binary outcome (win=1, loss=0), where lower MSE indicates better calibration. Calibration scores ranged from 0.1362 (qwen/qwen-max) to 0.5355 (deepseek/deepseek-r1-distill-qwen-14b:free), indicating substantial differences in the models' ability to align confidence with outcome.

Table 2: Model-Specific Debate Performance and Calibration Metrics

Model	Win Rate (%)	Avg. Confidence (%)	Overconfidence (%)	Calibration Score
anthropic/claude-3.5-haiku	33.3	71.7	+38.4	0.2314
anthropic/claude-3.7-sonnet	75.0	67.5	-7.5	0.2217
deepseek/deepseek-chat	33.3	74.6	+41.3	0.2370
deepseek/deepseek-r1-distill-qwen-14b	18.2	79.1	+60.9	0.5355
google/gemini-2.0-flash-001	50.0	65.4	+15.4	0.2223
google/gemma-3-27b-it	58.3	67.5	+9.2	0.2280
openai/gpt-4o-mini	27.3	74.5	+47.2	0.3755
openai/o3-mini	33.3	77.5	+44.2	0.3826
qwen/qwen-max	83.3	73.3	-10.0	0.1362
qwen/qwq-32b:free	83.3	78.8	-4.5	0.1552

As shown in Table 2, models varied widely in their overconfidence (Avg. Confidence - Win Rate). Some models like qwen/qwen-max and qwen/qwq-32b:free were slightly underconfident on

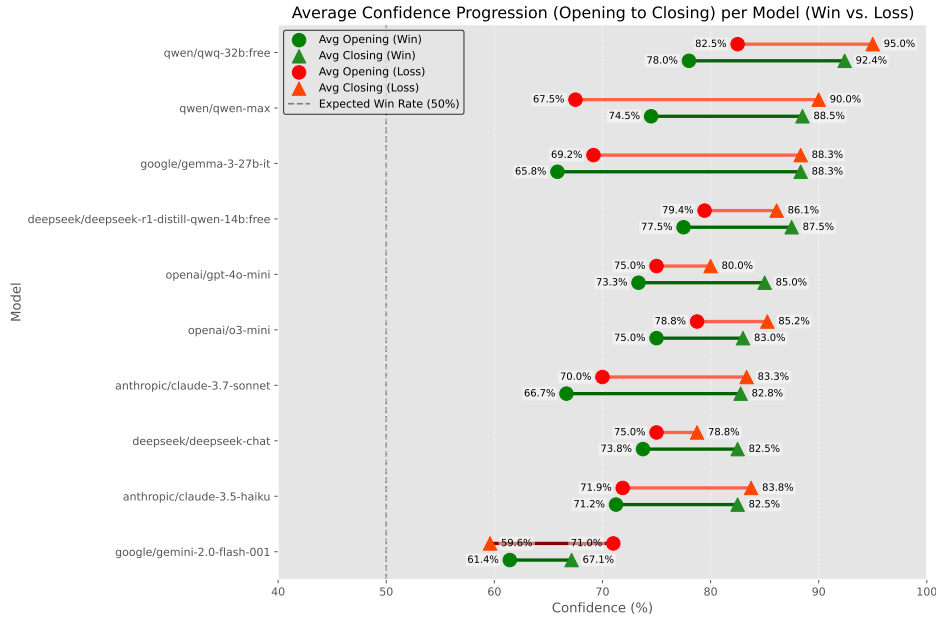


Figure 5: Confidence escalation across debate rounds for models that ultimately won versus models that ultimately lost.

average, achieving high win rates with relatively modest average confidence bets. Conversely, models like deepseek/deepseek-r1-distill-qwen-14b:free, openai/gpt-4o-mini, and openai/o3-mini exhibited substantial overconfidence.

Analyzing confidence tiers, models betting 76-100% confidence won only 45.2% of the time, slightly worse than those betting 51-75% (51.2% win rate). While there were limited data points for lower confidence tiers (only 1 instance in 26-50% and 0 in 0-25%), these findings suggest that high confidence in LLMs in this setting is not a reliable indicator of actual success.

Furthermore, a regression analysis using debate side (Proposition/Opposition) and average confidence as predictors of winning confirmed that while debate side was a highly significant predictor ($p < 0.0001$), average confidence was not ($p = 0.1435$). This reinforces that confidence in this multi-turn, adversarial setting was decoupled from factors driving actual debate success.

4.6 Jury Agreement and Topic Characteristics

The AI jury demonstrated moderate inter-rater reliability. 37.3% of debate outcomes were unanimous (all 6 judges agreed), while 62.7% involved split decisions among the judges. Dissenting opinions were distributed as follows: 1 dissenting judge (18.6% of debates), 2 dissenting (32.2%), and 3 dissenting (11.9%). This level of agreement suggests the jury system provides a reliable, albeit not always perfectly consensual, ground truth for complex debate outcomes at scale.

Topic difficulty, as measured by the AI jury’s difficulty index, varied across the six motions, ranging from the least difficult (media coverage requirements, 50.50) to the most difficult (social media shareholding, 88.44). This variation ensured that models debated across a range of complexity, although the core findings on overconfidence and calibration deficits were consistent across topics.

5 Conclusion

— YOUR CONCLUSION CONTENT HERE —

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>.
- Loka Li, Guan-Hong Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric P. Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *ArXiv*, abs/2402.12563, 2024. URL <https://api.semanticscholar.org/CorpusID:268032763>.
- 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>.

A LLMs in the Debater Pool

This appendix lists the specific LLMs used in the debater pool for the experiments, including their names, providers, and potentially version information. [Content to be added]

306 **B Debate Pairings Schedule**

307 This appendix details the schedule and method used for pairing LLMs against each other across
308 different debate topics, ensuring a balanced experimental design. [Content to be added]

309 **C Debater Prompt Structures**

310 Full verbatim text of the structured prompts used to guide debater models in the Opening, Rebuttal,
311 and Final rounds, including constraints and judging guidance. [Content to be added]

312 **D AI Jury Prompt Details**

313 Full verbatim text of the detailed prompt provided to the AI jury models for evaluating debate
314 transcripts, including judging criteria and output requirements. [Content to be added]

315 **E Topics of Debate**

316 **F Technical Appendices and Supplementary Material**

317 — YOUR APPENDIX CONTENT HERE (OPTIONAL) —

318 NeurIPS Paper Checklist

319 1. Claims

320 Question: Do the main claims made in the abstract and introduction accurately reflect the
321 paper’s contributions and scope?

322 Answer: [TODO]

323 Justification: [TODO]

324 2. Limitations

325 Question: Does the paper discuss the limitations of the work performed by the authors?

326 Answer: [TODO]

327 Justification: [TODO]

328 3. Theory assumptions and proofs

329 Question: For each theoretical result, does the paper provide the full set of assumptions and
330 a complete (and correct) proof?

331 Answer: [TODO]

332 Justification: [TODO]

333 4. Experimental result reproducibility

334 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
335 perimental results of the paper to the extent that it affects the main claims and/or conclusions
336 of the paper (regardless of whether the code and data are provided or not)?

337 Answer: [TODO]

338 Justification: [TODO]

339 5. Open access to data and code

340 Question: Does the paper provide open access to the data and code, with sufficient instruc-
341 tions to faithfully reproduce the main experimental results, as described in supplemental
342 material?

343 Answer: [TODO]

344 Justification: [TODO]

345 6. Experimental setting/details

346 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
347 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
348 results?

349 Answer: [TODO]

350 Justification: [TODO]

351 7. Experiment statistical significance

352 Question: Does the paper report error bars suitably and correctly defined or other appropriate
353 information about the statistical significance of the experiments?

354 Answer: [TODO]

355 Justification: [TODO]

356 8. Experiments compute resources

357 Question: For each experiment, does the paper provide sufficient information on the com-
358 puter resources (type of compute workers, memory, time of execution) needed to reproduce
359 the experiments?

360 Answer: [TODO]

361 Justification: [TODO]

362 9. Code of ethics

363 Question: Does the research conducted in the paper conform, in every respect, with the
364 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

365 Answer: **[TODO]**
 366 Justification: **[TODO]**
 367 **10. Broader impacts**
 368 Question: Does the paper discuss both potential positive societal impacts and negative
 369 societal impacts of the work performed?
 370 Answer: **[TODO]**
 371 Justification: **[TODO]**
 372 **11. Safeguards**
 373 Question: Does the paper describe safeguards that have been put in place for responsible
 374 release of data or models that have a high risk for misuse (e.g., pretrained language models,
 375 image generators, or scraped datasets)?
 376 Answer: **[TODO]**
 377 Justification: **[TODO]**
 378 **12. Licenses for existing assets**
 379 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
 380 the paper, properly credited and are the license and terms of use explicitly mentioned and
 381 properly respected?
 382 Answer: **[TODO]**
 383 Justification: **[TODO]**
 384 **13. New assets**
 385 Question: Are new assets introduced in the paper well documented and is the documentation
 386 provided alongside the assets?
 387 Answer: **[TODO]**
 388 Justification: **[TODO]**
 389 **14. Crowdsourcing and research with human subjects**
 390 Question: For crowdsourcing experiments and research with human subjects, does the paper
 391 include the full text of instructions given to participants and screenshots, if applicable, as
 392 well as details about compensation (if any)?
 393 Answer: **[TODO]**
 394 Justification: **[TODO]**
 395 **15. Institutional review board (IRB) approvals or equivalent for research with human**
 396 **subjects**
 397 Question: Does the paper describe potential risks incurred by study participants, whether
 398 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
 399 approvals (or an equivalent approval/review based on the requirements of your country or
 400 institution) were obtained?
 401 Answer: **[TODO]**
 402 Justification: **[TODO]**
 403 **16. Declaration of LLM usage**
 404 Question: Does the paper describe the usage of LLMs if it is an important, original, or
 405 non-standard component of the core methods in this research? Note that if the LLM is used
 406 only for writing, editing, or formatting purposes and does not impact the core methodology,
 407 scientific rigor, or originality of the research, declaration is not required.
 408 Answer: **[TODO]**
 409 Justification: **[TODO]**