
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. Debate provides a natural framework to stress-test these metacognitive abilities because it requires participants to respond to direct challenges, adapt to new information, and continually reassess the relative strength of competing positions—particularly when their arguments are directly contradicted or new evidence emerges. In adversarial settings, where one side must ultimately prevail, a rational agent should recognize when its position has been weakened and adjust its confidence accordingly. This is especially true when debaters have equal capabilities, as neither should maintain an unreasonable expectation of advantage.

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. The debate setup ensures both sides have equal access to information and equal opportunity to present their case. 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. Four key findings emerge: First, LLMs are systematically overconfident: average confidence is 72.92%, despite a 50% expected win rate. Second, this overconfidence paradoxically increases when models are more likely to lose—Proposition debaters won only 28.8% of debates yet expressed higher average confidence than Opposition models (74.58% vs. 71.27%). Third, instead of converging toward 50% as counter-evidence accumulates, average stated win probability climbs from 69% (opening) to 78% (closing). This "confidence escalation" occurs even in losing models that should recognize their deteriorating position. Fourth, overconfidence persists even though all models know they face opponents of equal capability, with no inherent advantage. In 71.2% of debates, both debaters report high confidence ($\geq 75\%$)—a logically incoherent outcome.

These findings raise serious concerns about deploying LLMs in roles requiring accurate self-assessment or real-time adaptation to new evidence and arguments. 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. This 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. 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

89 tests whether models can *dynamically monitor* their epistemic standing in a debate—arguably a more
90 socially and cognitively demanding task.

91 **Debate as Evaluation and Oversight.** Debate has been proposed as a mechanism for AI alignment,
92 where two agents argue and a human judge evaluates which side is more truthful or helpful [Irving
93 et al., 2018]. More recently, Brown-Cohen et al. [2023] propose “doubly-efficient debate,” showing
94 that honest agents can win even when outmatched in computation, if the debate structure is well-
95 designed. While prior work focuses on using debate to elicit truthful outputs or train models, we
96 reverse the lens: we use debate as a testbed for evaluating *epistemic self-monitoring*. Our results
97 suggest that current LLMs, even when incentivized and prompted to reflect, struggle to track whether
98 they are being outargued.

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

106 **Human Overconfidence Baselines.** This section will present literature on human overconfidence
107 in reasoning tasks and debates. We will discuss established findings on how humans often exhibit
108 similar overconfidence patterns and relate this to our LLM findings. Key references for human
109 calibration baselines will be introduced.

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

116 3 Methodology

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

124 3.1 Debate Simulation Environment

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

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

133 3.2 Structured Debate Framework

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

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

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

144 3.3 Core Prompt Structures & Constraints

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

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

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

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

156 3.4 Dynamic Confidence Elicitation

157 After generating the content for *each* of their three speeches (including the concurrent opening),
158 models were required to provide a private “confidence bet”.

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

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

167 3.5 Evaluation Methodology: The AI Jury

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

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

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

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

```

===== 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.

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

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

```
===== 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).

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

200 3.6 Data Collection

201 The final dataset comprises the full transcripts of 59 debates, the round-by-round confidence bets
 202 (amount and private thoughts) from both debaters in each debate, and the detailed structured verdicts
 203 (winner, confidence, reasoning) from each of the six AI judges for every debate. This data enables
 204 the quantitative analysis of LLM overconfidence, calibration, and confidence revision presented in
 205 our findings.

206 **[TODO - FLESH OUT STATS TESTS]** We conducted statistical hypothesis tests for each key
 207 finding, including t-tests for overconfidence claims, chi-square tests for position asymmetry, and

208 regression analysis for confidence escalation patterns. Additionally, we evaluated which LLMs made
 209 the most accurate predictions of debate outcomes.

210 4 Results

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

214 4.1 Pervasive Overconfidence and Logical Impossibility

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

222 This widespread overestimation suggests a fundamental disconnect between the models’ internal
 223 assessment of their performance and the objective outcome of the debate.

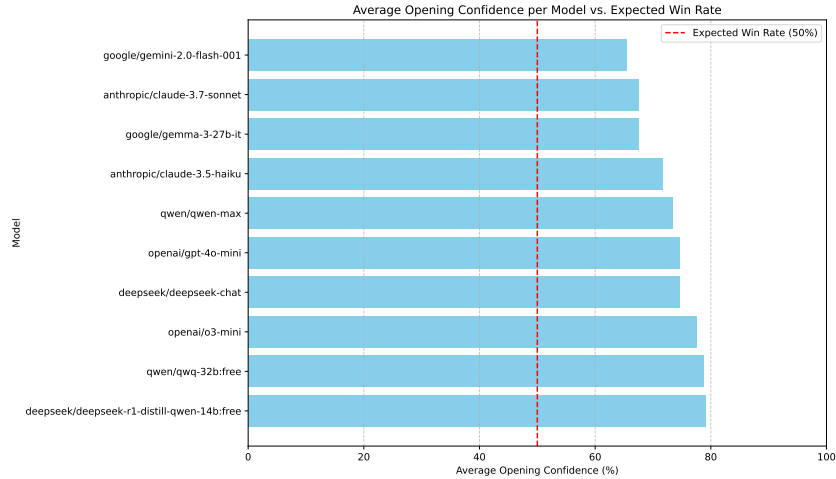


Figure 3: Average stated confidence in the first round across all LLMs and rounds compared to the expected 50% win rate.

224 4.2 Position Asymmetry and Confidence Mismatch

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

229 Despite this clear disparity in success rates, Proposition models reported *higher* average confidence
 230 (74.58%) than Opposition models (71.27%) across all rounds. While the difference in confidence itself
 231 is modest, its direction is contrary to the observed outcomes and statistically significant (Independent
 232 t-test: $t(175) = 2.54, p = 0.0115$; Mann-Whitney U test: $U = 4477, p = 0.0307$). This indicates
 233 that models failed to recognize or account for the systematic disadvantage faced by the Proposition
 234 side in this environment.

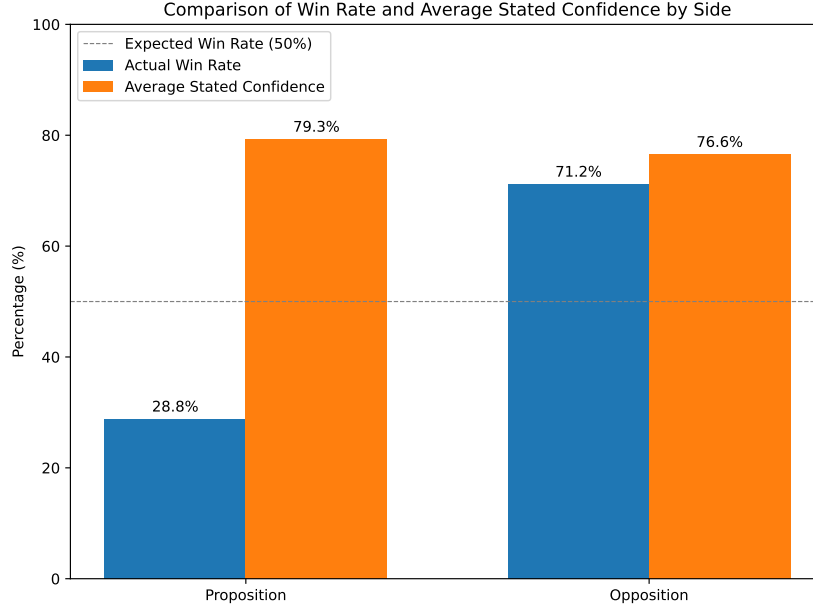


Figure 4: Comparison of Win Rate and Average Confidence for Proposition and Opposition sides.

4.3 Logically Impossible Confidence Scenarios

A stark illustration of LLM metacognitive failure is the frequency with which both debaters expressed high confidence simultaneously. In 71.2% of the 59 debates, both the Proposition and Opposition 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.

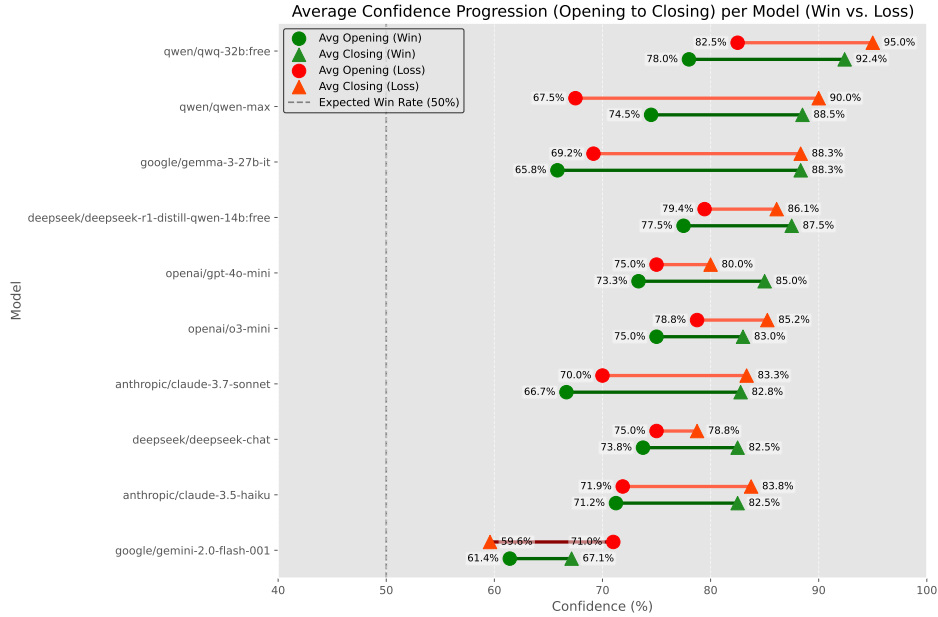


Figure 5: Confidence escalation across debate rounds for models that ultimately won versus models that ultimately lost.

[STATISTICAL VERIFICATION, TBA] We will expand this section with statistical verification of the confidence escalation pattern, including analysis of confidence trends for winning versus losing models. The data will demonstrate how this pattern contradicts rational Bayesian updating.

4.5 Persistence Against Identical Models

This subsection will present results from our ablation study on identical model debates. Despite debating against the exact same model architecture, LLMs maintain significant overconfidence levels. Table [X] shows confidence levels for each model when debating against itself. Average confidence across rounds remains well above the theoretically expected 50%, with most models increasing their confidence from opening to closing rounds. This demonstrates that overconfidence persists even when models should recognize that their opponent has identical capabilities with no inherent advantage.

4.6 Strategic Confidence in Public Settings

When models knew their confidence assessments would be visible to opponents, we observed different patterns compared to private betting scenarios. This section will analyze how models adapt their stated confidence when it may strategically influence opponents. Table [Y] presents comparative confidence levels between public and private conditions. The results have implications for chain-of-thought faithfulness and reveal potential strategic bluffing behavior that differs from genuine metacognitive assessment.

4.7 Model Performance, Calibration, and Evaluation Reliability

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

274 scores ranged from 0.1362 (qwen/qwen-max) to 0.5355 (deepseek/deepseek-r1-distill-qwen-14b:free),
 275 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

276 As shown in Table 2, models varied widely in their overconfidence (Avg. Confidence - Win Rate).
 277 Some models like qwen/qwen-max and qwen/qwq-32b:free were slightly underconfident on
 278 average, achieving high win rates with relatively modest average confidence bets. Conversely,
 279 models like deepseek/deepseek-r1-distill-qwen-14b:free, openai/gpt-4o-mini, and
 280 openai/o3-mini exhibited substantial overconfidence.

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

285 Furthermore, a regression analysis using debate side (Proposition/Opposition) and average confidence
 286 as predictors of winning confirmed that while debate side was a highly significant predictor ($p <$
 287 0.0001), average confidence was not ($p = 0.1435$). This reinforces that confidence in this multi-turn,
 288 adversarial setting was decoupled from factors driving actual debate success.

289 4.8 Jury Agreement and Topic Characteristics

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

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

299 **[LLM PREDICTION ACCURACY ANALYSIS, TBA]** This section will include an analysis of
 300 LLM prediction accuracy, examining which models were most effective at estimating their true
 301 likelihood of success, and whether any models showed superior metacognitive capabilities.

302 5 Discussion

303 5.1 Metacognitive Limitations and Possible Explanations

304 Our findings reveal significant limitations in LLMs’ metacognitive abilities, specifically their capacity
 305 to accurately assess their argumentative position and revise confidence in adversarial contexts. Several
 306 explanations may account for these observed patterns:

307 First, post-training for human preferences may inadvertently reinforce overconfidence. Models
 308 trained via RLHF are often rewarded for confident, assertive responses that match human preferences,
 309 potentially at the expense of epistemic calibration.

310 Second, training datasets predominantly feature successful task completion rather than explicit
311 failures or uncertainty. This bias may limit models’ ability to recognize and represent losing positions
312 accurately.

313 Third, the observed confidence patterns may reflect more general human biases toward expressing
314 confidence around 70%, with 7/10 serving as a common attractor state in human confidence judgments.
315 LLMs may be mimicking this human tendency rather than performing proper Bayesian updating.

316 5.2 Implications for AI Safety and Deployment

317 The confidence escalation phenomenon identified in this study has significant implications for AI
318 safety and responsible deployment. In high-stakes domains like legal analysis, medical diagnosis,
319 or research, overconfident systems may fail to recognize when they are wrong or when additional
320 evidence should cause belief revision.

321 The persistence of overconfidence even in controlled experimental conditions suggests this is a
322 fundamental limitation rather than a context-specific artifact. This has particular relevance for
323 multi-agent systems, where models must negotiate, debate, and potentially admit error to achieve
324 optimal outcomes. If models maintain high confidence despite opposition, they may persist in flawed
325 reasoning paths or fail to incorporate crucial counterevidence.

326 5.3 Potential Mitigations and Guardrails

327 Our ablation study testing explicit 50% win probability instructions shows [placeholder for results].
328 This suggests that direct prompting approaches may help mitigate but not eliminate confidence biases.

329 Other potential mitigation strategies include:

- 330 • Developing dedicated calibration training objectives
- 331 • Implementing confidence verification systems through external validation
- 332 • Creating debate frameworks that explicitly penalize overconfidence or reward accurate
333 calibration
- 334 • Designing multi-step reasoning processes that force models to consider opposing viewpoints
335 before finalizing confidence assessments

336 5.4 Future Research Directions

337 Future work should explore several promising directions:

- 338 • Investigating whether human-LLM hybrid teams exhibit better calibration than either humans
339 or LLMs alone
- 340 • Developing specialized training approaches specifically targeting confidence calibration in
341 adversarial contexts
- 342 • Exploring the relationship between model scale, training methods, and confidence calibration
- 343 • Testing whether emergent abilities in frontier models include improved metacognitive
344 assessments
- 345 • Designing debates where confidence is directly connected to resource allocation or other
346 consequential decisions

347 6 Conclusion

348 — YOUR CONCLUSION CONTENT HERE —

349 References

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

- 352 Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller,
353 Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark,
354 and Deep Ganguli. Which economic tasks are performed with ai? evidence from millions of claude
355 conversations, 2025. URL <https://arxiv.org/abs/2503.04761>.
- 356 Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint*
357 *arXiv:1805.00899*, 2018. URL <https://arxiv.org/abs/1805.00899>.
- 358 Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas
359 Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly)
360 know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2207.05221)
361 [2207.05221](https://arxiv.org/abs/2207.05221).
- 362 Loka Li, Guan-Hong Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric P. Xing, and Kun
363 Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language
364 models. *ArXiv*, abs/2402.12563, 2024. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:268032763)
365 [268032763](https://api.semanticscholar.org/CorpusID:268032763).
- 366 Colin Rivera, Xinyi Ye, Yonsei Kim, and Wenpeng Li. Linguistic assertiveness affects factuality
367 ratings and model behavior in qa systems. In *Findings of the Association for Computational*
368 *Linguistics (ACL)*, 2023. URL <https://arxiv.org/abs/2305.04745>.
- 369 Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their
370 knowledge of language. *arXiv preprint arXiv:2503.07513*, 2025. URL [https://arxiv.org/](https://arxiv.org/abs/2503.07513)
371 [abs/2503.07513](https://arxiv.org/abs/2503.07513).
- 372 Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea
373 Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated
374 confidence scores from language models fine-tuned with human feedback. In *Proceedings of the*
375 *2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL
376 <https://arxiv.org/abs/2305.14975>.
- 377 Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms
378 express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings*
379 *of the 2024 International Conference on Learning Representations (ICLR)*, 2024. URL [https:](https://arxiv.org/abs/2306.13063)
380 [/arxiv.org/abs/2306.13063](https://arxiv.org/abs/2306.13063).
- 381 Rongwu Xu, Brian S. Lin, Han Qiu, et al. The earth is flat because...: Investigating llms’ belief
382 towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.06717*, 2023. URL
383 <https://arxiv.org/abs/2312.06717>.
- 384 Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei
385 Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments,
386 2025. URL <https://arxiv.org/abs/2504.03160>.
- 387 Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of
388 uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference*
389 *on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL [https://arxiv.](https://arxiv.org/abs/2302.13439)
390 [org/abs/2302.13439](https://arxiv.org/abs/2302.13439).

391 **A LLMs in the Debater Pool**

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

394 **B Debate Pairings Schedule**

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

397 **C Debater Prompt Structures**

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

400 **D AI Jury Prompt Details**

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

403 **E Topics of Debate**

404 **F Technical Appendices and Supplementary Material**

405 — YOUR APPENDIX CONTENT HERE (OPTIONAL) —

NeurIPS Paper Checklist

1. Claims

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

Answer: [TODO]

Justification: [TODO]

2. Limitations

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

Answer: [TODO]

Justification: [TODO]

3. Theory assumptions and proofs

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

Answer: [TODO]

Justification: [TODO]

4. Experimental result reproducibility

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

Answer: [TODO]

Justification: [TODO]

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [TODO]

Justification: [TODO]

6. Experimental setting/details

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

Answer: [TODO]

Justification: [TODO]

7. Experiment statistical significance

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

Answer: [TODO]

Justification: [TODO]

8. Experiments compute resources

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

Answer: [TODO]

Justification: [TODO]

9. Code of ethics

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

453 Answer: **[TODO]**
 454 Justification: **[TODO]**

455 **10. Broader impacts**
 456 Question: Does the paper discuss both potential positive societal impacts and negative
 457 societal impacts of the work performed?
 458 Answer: **[TODO]**
 459 Justification: **[TODO]**

460 **11. Safeguards**
 461 Question: Does the paper describe safeguards that have been put in place for responsible
 462 release of data or models that have a high risk for misuse (e.g., pretrained language models,
 463 image generators, or scraped datasets)?
 464 Answer: **[TODO]**
 465 Justification: **[TODO]**

466 **12. Licenses for existing assets**
 467 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
 468 the paper, properly credited and are the license and terms of use explicitly mentioned and
 469 properly respected?
 470 Answer: **[TODO]**
 471 Justification: **[TODO]**

472 **13. New assets**
 473 Question: Are new assets introduced in the paper well documented and is the documentation
 474 provided alongside the assets?
 475 Answer: **[TODO]**
 476 Justification: **[TODO]**

477 **14. Crowdsourcing and research with human subjects**
 478 Question: For crowdsourcing experiments and research with human subjects, does the paper
 479 include the full text of instructions given to participants and screenshots, if applicable, as
 480 well as details about compensation (if any)?
 481 Answer: **[TODO]**
 482 Justification: **[TODO]**

483 **15. Institutional review board (IRB) approvals or equivalent for research with human**
 484 **subjects**
 485 Question: Does the paper describe potential risks incurred by study participants, whether
 486 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
 487 approvals (or an equivalent approval/review based on the requirements of your country or
 488 institution) were obtained?
 489 Answer: **[TODO]**
 490 Justification: **[TODO]**

491 **16. Declaration of LLM usage**
 492 Question: Does the paper describe the usage of LLMs if it is an important, original, or
 493 non-standard component of the core methods in this research? Note that if the LLM is used
 494 only for writing, editing, or formatting purposes and does not impact the core methodology,
 495 scientific rigor, or originality of the research, declaration is not required.
 496 Answer: **[TODO]**
 497 Justification: **[TODO]**