RECOVERABLE DISAGREEMENT: WHEN AND WHY MULTI-AGENT DEBATE OUTPERFORMS VOTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Agent Debate (MAD) has been challenged by recent studies showing that it does not consistently outperform voting or inference-time reasoning methods such as chain-of-thought and self-consistency. However, we argue that this conclusion reflects prior studies' reliance on average-accuracy comparisons, which fail to capture MAD's potential in the settings where it may be most useful: difficult items where agents disagree and a correct answer exists as a minority signal. We formalize this setting as *recoverable disagreement*, reframing MAD as a mechanism for recovering correct information from disagreement rather than merely aggregating votes. To improve this recovery process, we introduce *Confidence-Aware Debate* (CAD), a debate protocol that derives each agent's response stability from repeated responses and makes this information explicit during debate, allowing agents to calibrate when to defend or revise their answers. Through a large-scale evaluation across diverse LLMs, debate groups, and full benchmark test sets, we show that CAD improves the effectiveness of debate relative to voting and existing debate variants, while remaining competitive with inference-time reasoning baselines. Item-level analyses show that CAD's gains concentrate in benchmarks with larger feasible recovery gaps and in harder questions where disagreement is more frequent. In the hardest regime, CAD better recovers correct minority answers and reduces cases where an initially wrong majority persists to the final decision. A targeted persuasion experiment further shows that internally unstable answers are substantially more likely to change under both misleading attacks and true corrections, supporting confidence as a measurable signal for *answer revision*. Our experimental results show that MAD outperforms voting by recovering useful signals from difficult disagreements where reliable reasoning matters.

1 INTRODUCTION

Human institutions improve judgment under uncertainty through structured consensus-building processes such as cross-examination, peer review, and expert deliberation. As LLMs are increasingly deployed in high-stakes domains (Chen et al., 2024b; Hung et al., 2023; Hager et al., 2024), this principle becomes central to designing reliable LLM-based reasoning systems. A single LLM response can inherit systematic biases and recurring errors from its training, alignment, and reasoning process (Yang et al., 2024; Hager et al., 2024; Du et al., 2024). Multi-Agent Debate (MAD) addresses these limitations by enabling multiple agents to exchange, challenge, and revise their answers, thereby operationalizing collective intelligence in artificial reasoning systems (Minsky, 1986; Du et al., 2024; Chen et al., 2024a).

Driven by this potential, MAD has emerged as a prominent topic in LLM research, with studies exploring multi-round interaction, alternative debate structures, heterogeneous agent ensembles, scalable oversight, and stronger or more diverse debaters (Du et al., 2024; Liang et al., 2024; Chan et al., 2023; Yin et al., 2023; Chen et al., 2024a; Wang et al., 2024b; Khan et al., 2024; Kenton et al., 2024; Lang et al., 2025). More recent work, however, shows that MAD does not consistently outperform strong inference-time reasoning methods or plurality voting, especially under controlled compute budgets (Smit et al., 2024; Choi et al., 2025; Zhang et al., 2025; Kaesberg et al., 2025). Moreover, debate can amplify majority opinions or suppress initially correct minority views (Wu et al., 2025; Zhou & Chen, 2025; Estornell & Liu, 2024). These critiques have brought three emerging questions to the forefront of MAD research:

(1) Does debate provide benefits beyond voting?

(2) When does debate recover what voting misses?

(3) What drives successful debate, and why do individual agents revise their answers?

We argue that answering these questions requires moving beyond average-performance comparisons toward an analysis of when debate can recover information that static aggregation discards. When most agents can already identify the correct answer, simple aggregation is sufficient to recover it. However, for high-uncertainty problems where agents' answers diverge, simple aggregation may fail to distinguish correct signals from noise

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and instead reinforce an incorrect majority. Our central hypothesis is that harder problems create more meaningful disagreement, and MAD helps when this disagreement contains a correct minority hypothesis. We formalize this condition as ***recoverable disagreement***, where agents initially disagree and the correct answer appears among the initial responses but is not the majority answer. Under recoverable disagreement, voting is structurally limited because it selects the most frequent answer without exchanging reasons, whereas debate can in principle recover a correct minority position. Unlike prior work that studies difficulty through task families, model scale, hard subsets, or benchmark-provided levels (Yang et al., 2025b; Choi et al., 2025; Zhang et al., 2025; Smit et al., 2024), we define item-level difficulty from participating models' correctness patterns and use it to analyze how difficulty reshapes the group's initial information structure. This perspective shows that disagreement, often treated as a failure mode, can instead enable higher-quality deliberation and more reliable judgment.

Motivated by this perspective, we introduce ***Confidence-Aware Debate*** (CAD), a debate protocol that uses answer stability to calibrate belief revision. Each agent answers the same question multiple times independently, and repeated-response consistency serves as an empirical confidence signal. Unlike prior confidence-aware debate methods that rely on self-reported confidence, calibrated confidence expressions, or uncertainty-based peer weighting (Lin & Hooi, 2025; Yoffe et al., 2024; Zhu et al., 2026), CAD does not use confidence to weight votes or privilege confident agents. Instead, CAD uses confidence to guide ***answer revision***, prompting agents to reassess less stable answers while helping stable answers resist unsupported majority pressure. To further validate this design, we conduct targeted persuasion experiments that expose initially correct agents to misleading counterarguments and initially incorrect agents to corrective ones. These experiments provide empirical support for CAD's use of metacognitive confidence as a revision signal and explain why it can recover useful minority signals from disagreement.

We evaluate this account through a large-scale study spanning 11 LLMs, three five-agent debate groups, and the full test sets of ARC Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), TruthfulQA (Lin et al., 2022), MMLU Formal Logic, and MMLU Professional Medicine (Hendrycks et al., 2020). Our main evaluation uses complete test sets rather than sampled subsets, enabling more reliable item-level analyses of debate behavior. We compare MAD variants, including plurality voting, decentralized debate, judge-based debate, and CAD, against prompt-based inference enhancements such as Chain-of-Thought (Kojima et al., 2022), Self-consistency (Wang et al., 2022), and few-shot prompting (Brown et al., 2020). This evaluation allows us to move beyond asking whether debate improves average accuracy, and instead identify when and why debate can recover correct minority signals that static aggregation misses. Our contributions are as follows:

- We formalize *recoverable disagreement*, reframing MAD's value as the recovery of correct information that exists within the group but is missed by voting methods.
- We introduce *Confidence-Aware Debate*, a protocol that uses an explicit consistency-based confidence signal to calibrate answer revision during multi-agent deliberation.
- We identify internal answer stability as a measurable predictor of *answer revision*, explaining why confidence-aware debate improves over standard debate approaches.

2 Related Work

Multi-agent debate, diversity, and recoverable disagreement. Multi-Agent Debate (MAD) improves LLM reasoning by allowing agents to exchange rationales, challenge one another, and revise their answers (Du et al., 2024). Prior work has explored alternative debate structures (Liang et al., 2024; Chan et al., 2023; Yin et al., 2023), heterogeneous agent ensembles (Chen et al., 2024a; Wang et al., 2024b), scalable oversight (Khan et al., 2024; Kenton et al., 2024), and stronger or more diverse debaters (Lang et al., 2025). However, recent studies show that MAD gains can diminish under stronger baselines or controlled compute budgets (Smit et al., 2024; Choi et al., 2025; Zhang et al., 2025), and that debate can amplify majority opinions or suppress initially correct minority views (Wu et al., 2025; Zhou & Chen, 2025; Estornell & Liu, 2024). These limitations motivate a conditional view of diversity. While different models, prompts, or reasoning paths may contribute complementary information (Chen et al., 2024a; Wang et al., 2024b; Lang et al., 2025), additional variation can also introduce noise, fragment support for correct answers, or reinforce misleading disagreement. We therefore distinguish generic disagreement from *recoverable disagreement*: cases where the correct answer appears among the initial responses but is not selected by static voting. This framing asks not whether debate is uniformly beneficial, but when disagreement contains an overlooked correct signal that interaction can recover.

Confidence, uncertainty, and answer revision. Recent work incorporates confidence or uncertainty into MAD through confidence-aware consensus, uncertainty communication, explicit confidence expression, and confidence-modulated update dynamics (Chen et al., 2024a; Yoffe et al., 2024; Lin & Hooi, 2025; Zhu et al., 2026). These methods typically use confidence as a social or aggregative signal, for example to estimate agent reliability, weight influence, or guide collective decision making. CAD instead uses repeated-response consistency as an empirical measure of item-level answer stability, building on self-consistency and calibration work (Wang et al., 2022; Lyu et al., 2025; Wang et al., 2024a). This distinction matters because prior studies show that LLMs can revise answers

under challenge, persuasion, or social pressure, sometimes productively and sometimes harmfully (Laban et al., 2023; Agarwal & Khanna, 2025; Stengel-Eskin et al., 2025). We treat answer revision as the central mechanism through which debate can outperform static voting: in recoverable-disagreement cases, agents must preserve useful overlooked signals while moving away from unsupported or unstable answers. CAD operationalizes this view by making internal answer stability explicit during debate, using confidence as a revision-control signal rather than as a final vote weight.

3 PRELIMINARY

We reformalize multi-agent debate as the problem of recovering correct information that is already present in the group but may be missed by static aggregation. This perspective shifts the focus from whether debate improves all examples uniformly to whether interaction can recover initially overlooked useful signals when voting fails.

Let x be an input problem with ground-truth answer $y^\star$, and let $\mathcal{A} = \{a_1, \ldots, a_n\}$ denote a group of n LLM agents. Before any interaction, each agent independently produces an initial answer $r_i^{(0)} = a_i(x)$, forming an initial response collection $R^{(0)} = (r_1^{(0)}, \ldots, r_n^{(0)})$. A static aggregation method such as plurality voting selects an answer directly from $R^{(0)}$, while multi-agent debate allows agents to exchange reasons and revise their answers over subsequent rounds.

Let $c(y; R^{(0)}) = \sum_{i=1}^{n} \mathbb{I}[r_i^{(0)} = y]$ denote the number of agents initially producing answer y. Since ties can occur, we define the plurality-vote candidate set as $\mathcal{Y}_{\text{vote}}(R^{(0)}) = \arg\max_y c(y; R^{(0)})$, and the tie-resolved plurality-vote answer as $\hat{y}_{\text{vote}} = \tau(\mathcal{Y}_{\text{vote}}(R^{(0)}))$, where τ is the tie-breaking rule used by the voting procedure.

We define *recoverable disagreement* as the condition in which the correct answer appears in the initial response collection but the tie-resolved plurality vote fails to select it:

$$c(y^\star; R^{(0)}) > 0 \quad \text{and} \quad \hat{y}_{\text{vote}} \neq y^\star.$$

This regime captures examples where the group already contains a correct signal, but static aggregation fails to select it. Debate has little to add when voting is already correct, and under our recovery formulation there is no initially available correct signal to recover when no agent produces the correct answer. Its distinctive value arises in recoverable-disagreement cases, where interaction can in principle elevate an unselected correct signal.

Let $\hat{y}_{\text{debate}}$ be the final answer after debate, obtained by post-debate voting, consensus, or a judge. Debate achieves recovery when a recoverable-disagreement item is converted into a correct final answer, i.e.,

$$c(y^\star; R^{(0)}) > 0, \quad \hat{y}_{\text{vote}} \neq y^\star, \quad \hat{y}_{\text{debate}} = y^\star.$$

Thus, our formulation reframes MAD as a recovery problem: the central question is whether debate can recover correct information already present in the group but missed by static aggregation.

Structural limitation of frequency-based aggregation. Recoverable disagreement exposes a limitation of static aggregation rules that rely only on initial answer frequencies. If

$$0 < c(y^\star; R^{(0)}) < \max_y c(y; R^{(0)}),$$

then any rule that selects only a maximally frequent answer cannot choose $y^\star$. If $y^\star$ is tied for the maximal count but not selected, the failure comes from tie-breaking. In both cases, voting fails despite an initially available correct signal. Debate can help only if interaction changes which answers agents support, allowing the overlooked signal to be selected.

4 CONFIDENCE-AWARE DEBATE

The recovery view suggests that debate succeeds only when agents revise productively: unstable answers should be open to better-supported alternatives, while stable answers should resist unsupported majority pressure. We introduce ***Confidence-Aware Debate*** (CAD), a synchronous debate protocol that uses internal answer stability to guide answer revision rather than to weight final votes.

Internal consistency check. Before debate, each agent a_i answers the same problem x independently for K trials, producing

$$\{r_{i,1}^{\text{rep}}, \ldots, r_{i,K}^{\text{rep}}\}.$$

We estimate the agent's item-level answer stability as

$$\text{conf}_i(x) = \frac{1}{K} \max_y \sum_{k=1}^{K} \mathbb{I}[r_{i,k}^{\text{rep}} = y].$$

This score measures how consistently the agent converges to the same answer across repeated independent attempts.

If there exists an answer y such that $r_{i,k}^{\text{rep}} = y$ for every agent i and every trial k, CAD returns y immediately without debate. Otherwise, each agent is assigned an internal consistency state. If $\text{conf}_i(x) = 1$, the agent is labeled UNANIMOUS; if $\text{conf}_i(x) < 1$, it is labeled INCONSISTENT. Here, INCONSISTENT denotes within-agent inconsistency across repeated responses. We reserve *disagreement* for differences between agents' answers. These states are used to route agents to different reasoning prompts, not to assign vote weights.

Round 0: internal logic generation. Before observing other agents' arguments, each agent receives its own repeated responses and generates an internal rationale. A UNANIMOUS agent is prompted to produce the core *premise* and *critical evidence* supporting its stable answer. An INCONSISTENT agent is prompted to produce a *correct reason* for its selected answer and an *uncertainty reason* explaining why its repeated responses were inconsistent. This stage turns repeated internal sampling into explicit self-assessment before social influence is introduced.

Round 1: initial shared-position generation. The internal rationales generated in Round 0 are collected across all agents and broadcast to every agent. Each agent then considers the full set of Round 0 rationales and produces its initial shared debate position,

$$r_i^{(1)} = (y_i^{(1)}, p_i^{(1)}, q_i^{(1)}),$$

where $y_i^{(1)}$ is the answer, $p_i^{(1)}$ is the premise, and $q_i^{(1)}$ is the conclusion. This round establishes the first explicit group-level positions after agents have inspected both their own internal consistency and the other agents' internal rationales.

Synchronous debate rounds. CAD proceeds synchronously. At each subsequent round t, all agents receive the same shared history $H^{(t)}$, consisting of all agents' previous answers, premises, and conclusions. The shared context is not filtered by agreement or disagreement: every agent receives the full set of all agents' positions from the previous round. Each agent independently produces an updated answer, premise, and conclusion,

$$r_i^{(t)} = a_i\left(x, H^{(t)}, r_i^{(t-1)}, \text{conf}_i(x), s_i\right),$$

where $s_i \in \{\text{UNANIMOUS}, \text{INCONSISTENT}\}$ is the agent's internal consistency state. Agents do not speak sequentially or update one another in a fixed order; all outputs in a round are generated independently from the same shared state and then broadcast together.

Agents are instructed to revise their answers only when another rationale provides stronger evidence, more valid reasoning, or a concrete counterexample. INCONSISTENT agents are encouraged to be receptive to such evidence, whereas UNANIMOUS agents are encouraged to revise cautiously rather than conform to majority pressure without justification.

Decision protocol. After each official debate round, CAD checks whether all agents have produced the same answer. If so, debate terminates and returns that unanimous answer. Otherwise, debate continues until the maximum debate round limit is reached.

If the agents do not reach unanimity by the final round, CAD first identifies the answers with the largest final-round vote count. Let C be this majority-tied candidate set. For each participant p, let $H_{p,t}$ denote its normalized answer at official debate round t, where $t = 1$ is the first round in which agents output an explicit answer after observing the Round 0 internal rationales. We compute answer volatility as

$$V_p = \sum_{t=2}^{m} \mathbb{I}[H_{p,t} \neq H_{p,t-1}],$$

where m is the final official debate round. For each candidate answer $k \in C$, we compute the cumulative volatility of its final-round supporters:

$$V_{\text{ans}}(k) = \sum_{p:H_{p,m}=k} V_p.$$

CAD selects the candidate with the smallest $V_{\text{ans}}(k)$. Remaining ties are resolved by sorting the stable candidates and selecting one using a fixed random seed.

Selective CAD Routing. We apply CAD only to questions with high inter-model disagreement. Given one answer from each of five models, routing uncertainty is $u(x) = 1 - \max_y n_y(x)/5$. If $u(x) \geq \tau$, the question is processed by CAD; otherwise, the five-vote plurality answer is returned. The router calls are reused in CAD initialization, yielding 5 calls for unrouted items and the actual CAD call cost for routed items.

Table 1: Main performance comparison across model groups and inference methods.

Group	Method	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Average
G2	Baseline (GPT-5)	97.87 ± 0.0	88.41 ± 0.3	98.02 ± 0.6	**97.79** ± 0.0	84.52 ± 0.3	93.32 ± 0.2
	Few-shot (GPT-5)	97.82 ± 0.3	**88.98** ± 0.6	97.22 ± 0.6	97.43 ± 0.5	83.60 ± 0.5	93.01 ± 0.1
	CoT (GPT-5)	97.87 ± 0.1	88.33 ± 0.4	97.62 ± 1.1	97.06 ± 0.0	86.17 ± 0.5	93.41 ± 0.3
	Self-consistency (GPT-5)	97.82 ± 0.3	88.78 ± 0.4	98.02 ± 0.6	97.43 ± 0.0	83.60 ± 0.2	93.13 ± 0.3
	Voting	98.08 ± 0.2	88.08 ± 0.1	97.62 ± 0.0	97.61 ± 0.3	84.33 ± 0.4	93.15 ± 0.0
	Judge(GPT-5-nano)	97.82 ± 0.1	88.33 ± 0.4	97.22 ± 0.6	97.43 ± 1.0	87.03 ± 0.9	93.57 ± 0.6
	Debate	**98.12** ± 0.1	88.21 ± 0.2	98.02 ± 0.6	97.61 ± 0.3	86.72 ± 0.4	93.74 ± 0.1
	CAD(Ours)	97.82 ± 0.1	87.92 ± 0.1	**98.41** ± 0.0	97.24 ± 0.3	**90.88** ± 0.6	**94.46** ± 0.2
	CAD + Router ($\tau = 0.2$)	97.91 ± 0.1	88.04 ± 0.0	**98.41** ± 0.0	97.43 ± 0.0	90.27 ± 0.4	94.41 ± 0.1
	CAD + Router ($\tau = 0.4$)	97.95 ± 0.1	88.04 ± 0.1	**98.41** ± 0.0	97.61 ± 0.3	88.37 ± 0.2	94.08 ± 0.0
G5	Baseline (GPT-oss:120b)	96.50 ± 0.4	84.81 ± 0.1	**97.62** ± 1.1	94.67 ± 0.3	79.62 ± 0.6	90.64 ± 0.1
	Few-shot (GPT-oss:120b)	96.33 ± 0.1	85.18 ± 0.4	97.22 ± 0.6	94.30 ± 0.3	78.21 ± 0.7	90.25 ± 0.3
	CoT (GPT-oss:120b)	96.37 ± 0.5	84.68 ± 0.0	95.24 ± 0.0	94.67 ± 0.3	81.15 ± 0.4	90.42 ± 0.1
	Self-consistency (GPT-oss:120b)	96.67 ± 0.2	85.67 ± 0.4	96.43 ± 0.6	95.22 ± 0.0	79.56 ± 0.4	90.71 ± 0.2
	Voting	96.37 ± 0.2	86.00 ± 0.2	94.84 ± 0.6	94.30 ± 0.8	73.99 ± 0.6	89.10 ± 0.2
	Judge(Qwen3:14b)	95.52 ± 0.4	82.76 ± 0.8	95.63 ± 0.6	87.50 ± 1.6	76.50 ± 0.9	87.58 ± 0.3
	Debate	**96.93** ± 0.2	**86.36** ± 0.4	96.83 ± 0.0	**95.59** ± 0.5	81.03 ± 0.9	91.35 ± 0.4
	CAD(Ours)	96.84 ± 0.0	85.83 ± 0.1	**97.62** ± 0.0	94.49 ± 0.5	**84.94** ± 0.2	**91.94** ± 0.1
	CAD + Router ($\tau = 0.2$)	96.80 ± 0.1	85.87 ± 0.2	**97.62** ± 0.0	94.49 ± 0.5	84.52 ± 0.3	91.86 ± 0.1
	CAD + Router ($\tau = 0.4$)	96.80 ± 0.1	86.32 ± 0.1	**97.62** ± 0.0	94.30 ± 0.3	82.07 ± 0.6	91.42 ± 0.1
G6	Baseline (Qwen3:14b)	96.37 ± 0.2	83.78 ± 0.2	**96.43** ± 0.6	93.38 ± 0.5	74.72 ± 0.4	88.94 ± 0.1
	Few-shot (GPT-5-nano)	95.35 ± 0.3	82.76 ± 0.8	90.48 ± 1.1	95.04 ± 0.8	76.68 ± 0.6	88.06 ± 0.4
	CoT (Qwen3:14b)	96.42 ± 0.4	83.58 ± 0.4	95.24 ± 1.1	93.01 ± 0.0	77.60 ± 0.2	89.17 ± 0.3
	Self-consistency (GPT-5-nano)	95.48 ± 0.2	83.54 ± 0.1	92.46 ± 0.6	95.22 ± 0.5	77.54 ± 0.3	88.85 ± 0.1
	Voting	96.25 ± 0.2	85.95 ± 0.2	92.06 ± 1.1	93.93 ± 0.3	75.89 ± 0.4	88.82 ± 0.3
	Judge(Qwen3:14b)	94.97 ± 0.4	80.59 ± 0.2	92.46 ± 1.7	85.29 ± 0.5	73.87 ± 0.3	85.44 ± 0.2
	Debate	96.46 ± 0.2	**86.08** ± 0.1	95.24 ± 2.2	95.77 ± 0.3	79.07 ± 0.7	90.52 ± 0.4
	CAD(Ours)	96.16 ± 0.0	84.28 ± 0.5	**96.43** ± 0.6	**95.96** ± 0.0	**82.93** ± 0.1	**91.15** ± 0.2
	CAD + Router ($\tau = 0.2$)	96.25 ± 0.1	84.56 ± 0.5	**96.43** ± 0.6	95.77 ± 0.3	82.62 ± 0.0	91.13 ± 0.2
	CAD + Router ($\tau = 0.4$)	**96.54** ± 0.1	85.67 ± 0.7	**96.43** ± 0.6	95.59 ± 0.0	80.91 ± 1.0	91.03 ± 0.1

5 EXPERIMENTS

5.1 EXPERIMENTAL DETAILS

Datasets. We evaluate all methods on five natural-language reasoning and knowledge tasks: ARC Challenge(ARC-C) (Clark et al., 2018), CommonsenseQA(CSQA) (Talmor et al., 2019), TruthfulQA (Lin et al., 2022), MMLU Professional Medicine(MMLU-PM), and MMLU Formal Logic(MMLU-FL) (Hendrycks et al., 2020). To improve reliability and avoid artifacts from sampled subsets, we use the full evaluation set for each benchmark. Dataset descriptions and statistics are provided in Appendix A.1.

Models and Debate Groups. We evaluate 11 open-source and commercial LLMs, namely GPT-oss:20b, GPT-oss:120b (OpenAI, 2025d), Qwen3:14b (Yang et al., 2025a), Ministral-3:8b (Liu et al., 2026), Gemma3:12b, Gemma3:27b (Kamath et al., 2025), GPT-5, GPT-5 nano, GPT-5 mini (OpenAI, 2025b), GPT-5.2 (OpenAI, 2025c), and GPT-4.1 (OpenAI, 2025a). We simulate distributed deployment with five local servers and an asynchronous proxy. Five-agent debates are evaluated on G2, G5, and G6, while ablated open-source groups (G7–G20) are used for additional group-composition analysis. Model and group details appear in Appendices A.3 and A.4.

Methods. We compare prompt-based inference enhancements and voting with multi-agent debate methods. Prompt-based baselines include naive prompting, few-shot (Brown et al., 2020) prompting, Chain-of-Thought (Kojima et al., 2022), and self-consistency (Wang et al., 2022). Multi-agent methods include plurality voting, decentralized debate(Debate), judge-based debate(Judge), and our proposed Confidence-Aware Debate (CAD). This design allows us to compare MAD not only against voting, but also against stronger inference-time prompting strategies. Prompt templates are provided in Appendix C, and implementation details in Appendix A.5.

5.2 LIMITATIONS OF MEAN-CENTRIC EVALUATION

We first examine whether debate-based methods improve over plurality voting and strong single-model inference-time baselines. Table 1 reports performance across three five-agent groups over two independent runs, using the best single model within each group for prompt-based baselines. Figure 1 complements this with each multi-agent

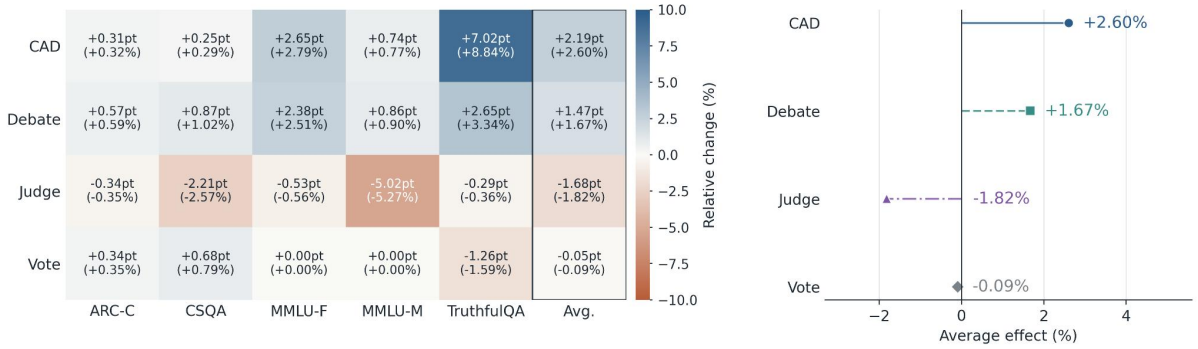


Figure 1: Average performance change from debate relative to the best-performing group member.

Table 2: Recoverable opportunity and correlation analysis.

(a) Benchmark-level opportunity analysis.

Dataset	Lower	Gap	H.Rate	CAD−Vote
ARC-C	84.39	14.99	0.15	-0.06
CSQA	60.37	35.69	2.56	-0.44
MMLU-FL	40.21	58.73	1.35	+2.65
MMLU-PM	70.59	29.04	0.00	+0.74
TruthfulQA	44.84	49.08	7.66	+8.28

(b) Gap–debate effect correlation.

Method	Pearson r	Spearman ρ
CAD	**0.554** ($p < 0.05$)	**0.729** ($p < 0.01$)
Debate	0.491 ($p = 0.063$)	0.482 ($p = 0.069$)
Judge	-0.097 ($p = 0.731$)	-0.098 ($p = 0.727$)
Voting	-0.267 ($p = 0.337$)	-0.068 ($p = 0.810$)

Note. Lower denotes CAD's lower bound, Gap denotes the feasible improvement margin, H.Rate denotes unrecoverable cases, and CAD−Vote denotes average CAD's improvement over plurality voting across all groups.

method's improvement relative to the best few-shot member in the corresponding group; full single-model and group-specific effect results are provided in Appendices B.1 and B.2.

Overall, CAD shows the strongest performance among the multi-agent methods. Decentralized debate improves over voting, and CAD further improves over decentralized debate, achieving the highest average accuracy in every group. The pooled effect analysis shows the same pattern: voting yields almost no gain over few-shot initialization, while debate improves more substantially and CAD achieves the largest gain. Selective routing preserves most of CAD's accuracy while avoiding debate on low-disagreement instances, with $\tau = 0.2$ remaining especially close to full CAD across all three groups.

These improvements are concentrated rather than uniform. CAD's largest gains appear on TruthfulQA and MMLU Formal Logic, whereas the remaining datasets show much smaller improvements. Thus, debate-based methods, especially CAD, can outperform voting, but only in settings where interaction has room to help. We next analyze where these dataset-level differences come from by measuring how much recoverable opportunity each dataset provides.

5.3 Recoverable Disagreement Explains When Debate Helps

The previous section shows that debate gains vary across datasets. We analyze this variation using the feasible improvement margin, denoted as *Gap*. When all agents initially give the same answer, no debate is needed and that outcome is fixed. For the remaining disagreement items, *Lower* assumes that debate fails to recover any additional correct answers, whereas *Upper* assumes that debate succeeds on every recoverable-disagreement case, i.e., every disagreement item where at least one agent initially produced the correct answer. Gap is the difference between these bounds, measuring the maximum feasible improvement available from recovering correct answers already present among the agents.

Table 2 shows that CAD's advantage over voting generally grows with this recoverable opportunity. ARC Challenge has the smallest Gap and almost no CAD improvement over voting, while MMLU Professional Medicine has a moderate Gap and a modest CAD gain. In contrast, MMLU Formal Logic and TruthfulQA have larger Gaps, and CAD improves more clearly over voting. Overall, benchmark-

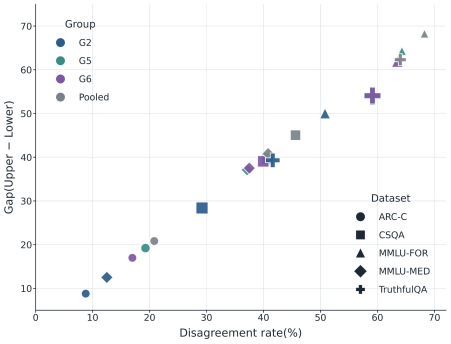
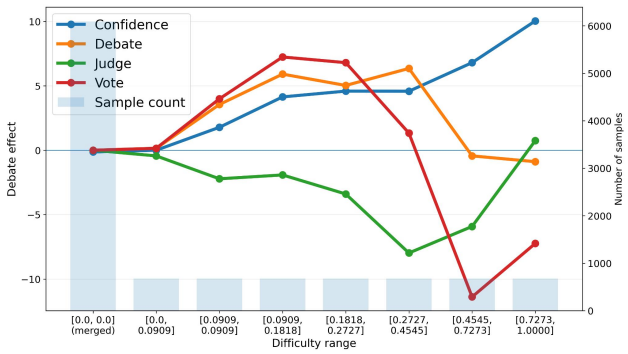
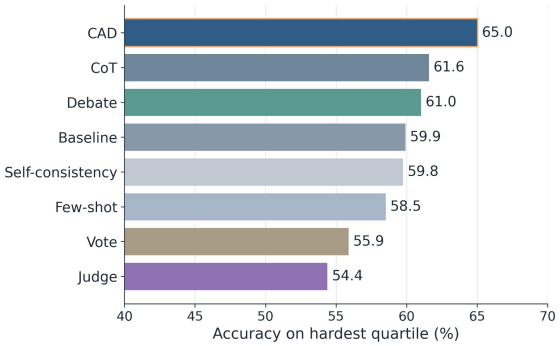


Figure 2: Disagreement rate versus feasible improvement margin (Gap).



(a) Difficulty-wise comparison of debate effects.



(b) Micro-averaged accuracy over G2, G5, G6.

Figure 3: Difficulty-wise analysis of debate effects. (a) compares debate effects across 16 difficulty bins. (b) compares micro-averaged accuracy on the hardest difficulty quantile over G2, G5, and G6.

Table 3: Aggregation-method comparison on the hardest difficulty quantile (Q4).

(a) Reasoning method-based gain/loss analysis

Method	CoT		Few-shot		Self-cons.	
	Gain↑	Loss↓	Gain↑	Loss↓	Gain↑	Loss↓
CAD	**11.53** ± 0.1	**8.13** ± 0.2	**14.01** ± 0.2	7.65 ± 0.4	**12.49** ± 0.2	**7.26** ± 0.1
Debate	9.59 ± 0.7	9.41 ± 0.4	9.33 ± 0.9	**6.19** ± 0.0	9.48 ± 0.9	7.46 ± 0.1
Judge	8.48 ± 0.6	15.23 ± 0.1	8.02 ± 0.7	11.81 ± 0.2	8.43 ± 0.8	13.34 ± 0.3
Voting	8.56 ± 0.3	14.01 ± 0.2	8.15 ± 0.2	10.64 ± 0.2	8.43 ± 0.4	12.05 ± 0.1

(b) Disagreement recovery and majority-error persistence.

Method	RD Acc.↑	Echo↓	UME↓	Maj. Correct↑
CAD	**43.49** ± 0.1	**28.16** ± 0.4	**30.80** ± 0.4	**57.63** ± 0.1
Debate	31.95 ± 1.4	32.15 ± 0.4	34.79 ± 0.5	55.80 ± 0.1
Judge	28.75 ± 0.2	29.55 ± 0.3	36.16 ± 0.1	52.59 ± 0.2
Voting	0.00 ± 0.0	44.22 ± 0.0	44.22 ± 0.0	55.78 ± 0.0

Note. All values are reported as mean ± sample standard deviation across two logical repeats. Gain/Loss are all-item rates of anchor wrong→final correct and anchor correct→final wrong cases, respectively. RD Acc. is the final-answer accuracy on recoverable-disagreement items under the legacy definition. Echo denotes the all-item rate of *same-wrong* cases, where an incorrect initial majority is preserved as the final answer. UME (Unresolved Majority Error) denotes the all-item rate of cases where the initial majority is incorrect and the final answer remains incorrect, including both same-wrong and other-wrong outcomes; hence, UME = Echo + Other-wrong. Maj. Correct is the accuracy of the tie-resolved initial majority. See Appendix A.2 for formal definitions.

level recoverable opportunity provides a useful explanation for where CAD gains are likely to appear, while leaving open finer item-level factors that we analyze next.

The correlation analysis reinforces this pattern. Across dataset–group settings, CAD improvement is significantly correlated with Gap (Pearson $r = 0.554$, Spearman $\rho = 0.729$), whereas decentralized debate shows a weaker trend and voting or judge-based debate show no meaningful positive association. This suggests that as the recoverable gap increases, CAD is better able to recover useful signals through interaction, while voting cannot exploit them because it only aggregates initial answers. Figure 2 further shows that larger Gaps are associated with higher initial disagreement rates among agents. We next examine where and why this Gap arises at the item level, focusing on how item difficulty increases disagreement and creates the conditions under which debate can help.

5.4 When Does Debate Help?

The previous analysis shows that recoverable gaps explain why debate gains differ across datasets. We now ask where these gaps arise at the item level. We measure item difficulty from the few-shot responses that initialize all voting and debate methods, using the correctness pattern of all 11 evaluated models: an item is harder when fewer models answer it correctly. Figure 3(a) reports a 16-bin difficulty analysis, with nearly identical low-difficulty bins merged for readability. Figure 3(b) and Table 3 focus on the hardest difficulty quantile, where the need for debate is strongest. Difficulty-wise accuracy, debate-entry, round-count, and cost analyses are provided in Appendix B.3 and B.4.

Figure 3(a) shows that method behavior changes sharply with item difficulty. On easy items, all methods have limited room to improve. As difficulty increases, voting deteriorates because disagreement spreads support across competing answers without any mechanism for evaluating them. Decentralized debate remains more robust than voting in moderately difficult regions, but weakens in the hardest region. CAD shows the opposite pattern, with its largest gains appearing on the hardest items. Figure 3(b) confirms this at the hardest quantile: when examples from G2, G5, and G6 are pooled and micro-averaged, CAD achieves the strongest accuracy among all methods, with a larger margin than in the overall average comparison. This indicates that CAD's advantage is concentrated precisely where disagreement is most severe.

Table 4: Stronger static aggregation versus interaction under increased sampling budgets.

Method	Calls	Avg. Acc.↑	Pool-RD Acc.↑	Hard-Q4 Acc.↑	Hard-Q4 Echo↓
Voting-5	5.00	90.39	0.00	55.69	41.65
Voting-25	25.00	90.98	0.00	57.24	42.39
Voting-75	75.00	91.04	0.00	57.69	42.02
CompetenceWeighted-25	25.00	91.15	1.98	57.95	41.54
UncertaintyWeighted-25	25.00	90.93	2.23	57.28	41.28
Competence×Uncertainty-25	25.00	91.08	3.71	57.87	40.87
Self-Consistency-25	25.00	90.92	0.00	59.83	40.06
CAD	27.53	**92.62**	36.93	**65.04**	**27.86**
CAD + Router ($\tau = 0.2$)	12.52	92.56	**40.43**	64.49	28.68
CAD + Router ($\tau = 0.4$)	8.86	92.18	31.45	62.16	32.11

Based on two independent runs, Table 3(a) explains this advantage through gain and loss relative to inference-time reasoning anchors. Across CoT, few-shot, and self-consistency anchors, CAD shows the strongest gain profile and consistently corrects more anchor errors than it damages correct anchor answers. Voting shows the opposite tendency under several anchors, suggesting that static aggregation can become harmful when difficult items scatter answers across competing choices. Thus, CAD's benefit in hard regimes comes from changing answers more productively, not merely more often.

Table 3(b) shows the same pattern for recoverable disagreement. CAD achieves the highest RD accuracy, indicating the strongest recovery of initially overlooked correct answers. Voting cannot recover RD cases by construction because it does not revise the initial plurality decision. CAD also reduces Echo relative to voting and decentralized debate, while avoiding the accuracy and loss trade-offs of judge-based debate.

Figure 4 shows that CAD's computation is difficulty-aware. Easier items usually require little deliberation, while harder items trigger deeper interaction among agents. Thus, CAD does not apply high-cost debate uniformly; its debate dynamics allocate more computation when disagreement persists and recovery is more likely to matter, without requiring a separate difficulty classifier. Selective routing strengthens this behavior by bypassing CAD on low-disagreement items, further reducing calls on easier bins while preserving more computation for harder cases.

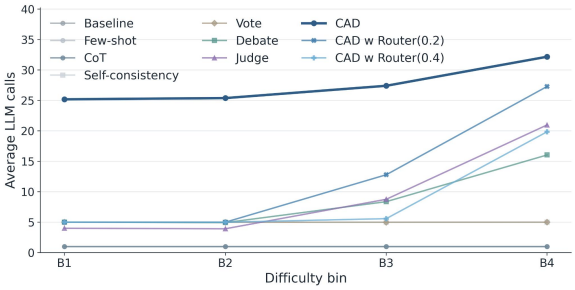


Figure 4: Average LLM calls by difficulty.

Together, these results show that CAD is best aligned with the regime where debate is most needed. Item-level difficulty creates the Gap by increasing disagreement, recoverable opportunity, and wrong-majority risk; CAD more effectively converts these difficult cases into correct final decisions. It also invokes deeper deliberation mainly on hard items, rather than applying high-cost debate uniformly. Selective routing further concentrates this interaction on high-disagreement cases, reducing unnecessary computation on easier items.

5.5 Can Stronger Static Aggregation Substitute for Interaction?

We next ask whether CAD's gains come from deliberative interaction or can be reproduced by stronger static aggregation. Since CAD uses repeated responses before debate, we compare it with high-sample voting, competence- and uncertainty-weighted voting, their combination, and 25-sample self-consistency; implementation details are provided in Appendix A.5.

In Table 4, Pool-RD Acc. evaluates recovery on cases defined from each method's own sampled answer pool, i.e., the set of responses actually available to that method before its final decision. We first form raw, unweighted answer counts from this pool and identify Pool-RD cases in which the gold answer appears but remains a strict minority, $0 < c_i(y^\star) < \max_y c_i(y)$; top ties involving the gold answer are excluded. For weighted voting methods, the same raw counts are used to define Pool-RD cases, while the weights affect only the final prediction. Hard-Q4 Echo measures, within the hardest quartile, how often an initially wrong raw plurality remains the final answer. Calls reports the average number of API calls per question, including both pre-debate sampling and debate-round generations.

Stronger static aggregation yields only modest gains: increasing voting from 5 to 25 or 75 calls and applying weighted aggregation improves average accuracy only slightly. More importantly, unweighted voting cannot recover Pool-RD cases by construction, and weighted variants recover only a small fraction. CAD, in contrast, achieves

substantially higher Pool-RD recovery, indicating that its advantage comes from agents revising their support after exchanging reasons rather than from more carefully counting the same responses.

The hardest-quartile results reinforce this distinction. Static aggregation and self-consistency remain well below CAD in Hard-Q4 accuracy, while CAD substantially reduces Echo. Thus, interaction helps agents escape initially dominant but incorrect answers rather than merely amplifying the pre-deliberation answer distribution. Selective routing further preserves most of CAD's gains at substantially lower call cost, especially with $\tau = 0.2$.

These results clarify the role of computation in CAD. Additional samples help static methods estimate what the group already tends to answer, but they do not provide a mechanism for changing that tendency. CAD adds such a mechanism through rationale exchange and confidence-aware revision. Its main value is therefore not that it uses more calls, but that it can recover correct answers that are present in the group yet remain unselected by stronger static aggregation.

5.6 Why Do Agents Revise Their Answers?

The previous section shows that CAD is most effective when difficult items create recoverable disagreement. We now examine why agents revise their answers in such cases. Answer revision depends on latent factors such as rationale quality, model knowledge, and resistance to misleading arguments, which are difficult to measure directly. We test whether consistency-based uncertainty offers an observable signal for this process. Detailed experimental settings are provided in Appendix A.6.

We conduct a targeted persuasion experiment with two balanced conditions. In *Attack*, a correct target answer is challenged by incorrect arguments; in *Correction*, an incorrect target answer is challenged by correct arguments. We then test whether repeated-response uncertainty predicts whether the target changes its answer.

Table 5 shows that low-confidence answers are more likely to change in both Attack and Correction. This indicates that consistency-based confidence is not only a reliability signal, but also a measurable predictor of answer revision.

Table 5: Answer-change rate by confidence.

Condition	Low	High	Diff.
Attack	62.02	28.21	+33.81
Correction	83.96	64.82	+19.14

Figure 5 shows the same pattern more finely: answer changes generally increase with uncertainty, most clearly under Correction. The weaker, less monotonic Attack curve indicates that uncertainty is not the only factor; models' knowledge and resistance to false rationales also matter.

These findings support CAD's use of explicit confidence. Decentralized debate leaves uncertainty implicit, whereas CAD exposes internal answer stability and uses it to guide whether agents should defend or revise their answers. Since uncertainty predicts revision behavior, confidence is best used to regulate answer revision rather than weight final votes. Supporting ablations and case studies are provided in Appendix B.6 and Appendix B.7.

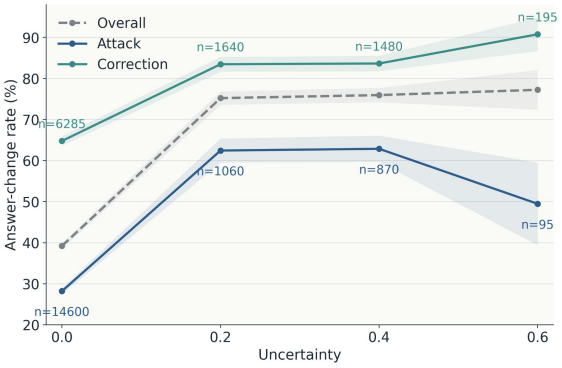


Figure 5: Answer-change rate by uncertainty.

6 Conclusion

This work identifies the regime where multi-agent debate has distinctive value: difficult disagreements in which useful signals are present but missed by static aggregation. We formalize this regime as *recoverable disagreement*, reframing MAD from a costly voting alternative into a mechanism for recovering overlooked signals.

CAD operationalizes this view. Across our analyses, its gains concentrate exactly where recoverable opportunity and item difficulty are high. In the hardest regimes, CAD more effectively recovers correct overlooked answers, suppresses wrong-majority persistence, and outperforms voting and existing debate variants. Our persuasion analysis explains the mechanism: internal answer instability predicts revision, validating confidence as a signal for controlling answer changes rather than weighting votes.

These results clarify when debate is worth its cost. MAD should not be judged only by average efficiency; its value lies in difficult, uncertain settings where static aggregation can miss useful signals. Extending this recovery view to open-ended deliberation, where agents construct defensible answers rather than recover known labels, is a key direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with writing refinement, wording, organization, and the development and debugging of experiment-support code under the authors' direction and control. AI tools did not independently determine experimental settings, methodological choices, or scientific conclusions. All AI-assisted code and analysis workflows were specified, reviewed, tested, and validated by the authors against experiment outputs and logs. The central research questions, experimental design, final methodological decisions, scientific claims, and citation choices were determined by the authors.

ETHICS STATEMENT

This work studies multi-agent debate among LLMs on existing benchmark datasets. It does not involve human subjects, collect personal data, or release sensitive data. We do not identify major ethical risks beyond the general limitations of using LLM-based decision protocols. CAD should not be interpreted as a guarantee of truthfulness or reliability, and it should not be directly used for high-stakes medical, legal, or financial decisions without human oversight and domain-specific validation.

REPRODUCIBILITY STATEMENT

We provide implementation details, metric definitions, experimental settings, and server information in the appendix and supplementary material. To improve reliability, we report repeated results for the main experiments. We also include the code used for the experiments in the supplementary material, together with details needed to reproduce the reported evaluation.

REFERENCES

Mahak Agarwal and Divyam Khanna. When persuasion overrides truth in multi-agent llm debates: Introducing a confidence-weighted persuasion override rate (cw-por). *arXiv preprint arXiv:2504.00374*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, 2024a.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024b.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 11733–11763, 2024.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. Walking a tightrope–evaluating large language models in high-risk domains. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 99–111, 2023.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, 2025.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, Cj Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, J. Michael Wieting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan, Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Ping mei Xu, Piotr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L'eonard Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL `https://api.semanticscholar.org/CorpusID:277313563`.

Zachary Kenton, Noah Y Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D Goodman, et al. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276, 2024.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 23662–23733, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Philippe Laban, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*, 2023.

Hao Lang, Fei Huang, and Yongbin Li. Debate helps weak-to-strong generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27410–27418, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Zijie Lin and Bryan Hooi. Enhancing multi-agent debate system performance via confidence expression. *arXiv preprint arXiv:2509.14034*, 2025.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268, 2025.

Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

OpenAI. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`, April 2025a. Accessed: 2026-04-30.

OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, August 2025b. Accessed: 2026-04-30.

OpenAI. Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/`, December 2025c. Accessed: 2026-04-30.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. `https://openai.com/index/gpt-oss-model-card/`, August 2025d. Accessed: 2026-04-30.

Andries Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 45883–45905, 2024.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Teaching models to balance resisting and accepting persuasion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8108–8122, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6023–6029, 2024a.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Haolun Wu, Zhenkun Li, and Lingyao Li. Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yifan Yang, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W John Wilbur, Xiaojun Li, et al. Ensuring safety and trust: Analyzing the risks of large language models in medicine. *ArXiv*, pp. arXiv–2411, 2024.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025b.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15135–15153, 2023.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. Debunc: Improving large language model agent communication with uncertainty metrics. *arXiv preprint arXiv:2407.06426*, 2024.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

Yan Zhou and Yanguang Chen. Adaptive heterogeneous multi-agent debate for enhanced educational and factual reasoning in large language models. *Journal of King Saud University Computer and Information Sciences*, 37 (10):330, 2025.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. *arXiv preprint arXiv:2601.19921*, 2026.

Appendix

Table of Contents

A Additional Experimental Details 14
- A.1 Datasets . . . 14
- A.2 Metrics . . . 15
- A.3 Models and Deployment . . . 16
- A.4 Debate Groups . . . 17
- A.5 Implementation Details . . . 17
- A.6 Targeted Persuasion Experiment . . . 19

B Supplementary Results 20
- B.1 Full Single Model Performance . . . 20
- B.2 Debate Effect by Group . . . 21
- B.3 Difficulty-wise Debate Dynamics and Performance . . . 23
- B.4 Difficulty-wise Debate Cost . . . 25
- B.5 Reliability Analysis . . . 25
- B.6 Ablation Study . . . 26
- B.7 Case Study . . . 28

C Prompt Templates 34

D Limitations, Broader Impact, Future Directions 43

A Additional Experimental Details

A.1 Datasets

We evaluate on five natural-language reasoning and knowledge benchmarks: ARC Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), TruthfulQA MC1 (Lin et al., 2022), MMLU Professional Medicine, and MMLU Formal Logic (Hendrycks et al., 2020). We use the full evaluation set for each benchmark whenever available, rather than sampled subsets, to support reliable item-level analyses of debate behavior.

Table 6: Dataset statistics.

Dataset	# Examples	Task type
ARC Challenge (Clark et al., 2018)	1,172	Science question answering
CommonsenseQA (Talmor et al., 2019)	1,221	Commonsense reasoning
TruthfulQA MC1 (Lin et al., 2022)	817	Truthfulness and misconception robustness
MMLU Professional Medicine (Hendrycks et al., 2020)	272	Domain-specific medical knowledge
MMLU Formal Logic (Hendrycks et al., 2020)	126	Formal and logical reasoning

ARC Challenge. ARC Challenge is the difficult subset of the AI2 Reasoning Challenge, consisting of grade-school multiple-choice science questions drawn from public standardized exams (Clark et al., 2018). The Challenge Set contains questions that are difficult for simple retrieval and word co-occurrence methods, making it more dependent on scientific knowledge and reasoning. We use the full ARC Challenge evaluation set, which contains 1,172 examples. Its output space includes `A,B,C,D` and the equivalent numeric labels `1,2,3,4`; these forms are normalized to the same canonical choices during evaluation.

CommonsenseQA. CommonsenseQA is a multiple-choice benchmark for commonsense reasoning (Talmor et al., 2019). The questions are constructed from ConceptNet and require selecting the most plausible answer from semantically related choices. We evaluate on 1,221 examples from CommonsenseQA. The output space is `A,B,C,D,E`, with case and minor formatting variations normalized during evaluation.

TruthfulQA MC1. TruthfulQA evaluates whether models avoid common misconceptions and false beliefs (Lin et al., 2022). It contains questions across diverse categories such as health, law, finance, and politics, where plausible answers may be false. We use the MC1 setting, in which each question has a single best truthful answer, and

evaluate on 817 examples. The output space spans `A`–`L`, with formatting variants normalized to the corresponding canonical option.

MMLU Professional Medicine. MMLU is a broad multiple-choice benchmark covering 57 academic and professional subjects (Hendrycks et al., 2020). The Professional Medicine subset evaluates domain-specific biomedical knowledge, including inheritance, genetic disorders, molecular genetics, and clinical genetics. We evaluate on 272 examples. The output space is `A,B,C,D`, with case and minor formatting variations normalized during evaluation.

MMLU Formal Logic. MMLU Formal Logic evaluates formal reasoning and logical problem solving in a multiple-choice format (Hendrycks et al., 2020). The questions require reasoning about logical structure, such as equivalence, contradiction, consistency, or validity. We evaluate on 126 examples. The output space is `A,B,C,D`, with case and minor formatting variations normalized during evaluation.

A.2 METRICS

Let B_k denote the set of items in difficulty bin k, and let $n_k = |B_k|$. For each item $i \in B_k$, let $a_i \in \{0, 1\}$ be the correctness of the selected anchor model, and let $f_i \in \{0, 1\}$ be the correctness of the final method output. The anchor can be selected from one of the single-model baselines, such as Few-shot, CoT, Baseline, or Self-consistency. Difficulty bins are always constructed using Few-shot single-run correctness.

Let G_i be the gold answer for item i, R_i be the multiset of initial participant answers, m_i be the initial majority answer after tie resolution, and y_i^{final} be the final predicted answer. If multiple answers are tied for the initial majority, one answer is selected using a fixed random seed.

Difficulty. For pooled analysis, item difficulty is computed using all unique member models from the selected groups under the single-run setting:

$$d_i^{\text{pool}} = 1 - \frac{1}{|\mathcal{M}_{\text{pool}}(i)|} \sum_{m \in \mathcal{M}_{\text{pool}}(i)} c_{i,m}^{\text{single}},$$

where $c_{i,m}^{\text{single}} \in \{0, 1\}$ indicates whether model m answered item i correctly in the single-run setting. Thus, lower values indicate easier items, and higher values indicate harder items. Items are sorted by d_i^{pool} and divided into equal-count difficulty bins.

Gain and Loss. An item is counted as a gain if the anchor is incorrect but the final method output is correct:

$$g_i = \mathbb{I}[a_i = 0 \wedge f_i = 1].$$

An item is counted as a loss if the anchor is correct but the final method output is incorrect:

$$\ell_i = \mathbb{I}[a_i = 1 \wedge f_i = 0].$$

We report gain and loss as prevalence over all items:

$$\text{Gain}(B_k) = \frac{1}{n_k} \sum_{i \in B_k} g_i, \qquad \text{Loss}(B_k) = \frac{1}{n_k} \sum_{i \in B_k} \ell_i.$$

Recoverable disagreement rate. An item is a recoverable disagreement if the gold answer appears among the initial answers, but the tie-resolved initial majority is incorrect:

$$r_i = \mathbb{I}[G_i \in R_i \wedge m_i \neq G_i].$$

The recoverable disagreement rate is

$$\text{RDRate}(B_k) = \frac{1}{n_k} \sum_{i \in B_k} r_i.$$

Recoverable disagreement accuracy. Recoverable disagreement accuracy is the final method accuracy restricted to recoverable disagreement items:

$$\text{RDAcc}(B_k) = \frac{\sum_{i \in B_k} r_i f_i}{\sum_{i \in B_k} r_i}.$$

If there are no recoverable disagreement items in B_k, this value is reported as NA.

Plurality correctness. Let q_i indicate whether the tie-resolved initial plurality answer is correct:

$$q_i = \mathbb{I}[m_i = G_i].$$

The majority correctness rate is

$$\mathrm{MajorityCorrect}(B_k) = \frac{1}{n_k} \sum_{i \in B_k} q_i.$$

Tie rate among majority-correct items. Let $t_i \in \{0, 1\}$ indicate whether item i required tie resolution, and let $s_i \in \{0, 1\}$ indicate whether the selected tie-resolved majority answer was correct. The value reported in parentheses next to Majority Correct is

$$\mathrm{TieSelectedAmongCorrect}(B_k) = \frac{\sum_{i \in B_k} \mathbb{I}[t_i = 1 \wedge s_i = 1]}{\sum_{i \in B_k} q_i}.$$

If there are no majority-correct items in B_k, this value is reported as NA.

Echo rate. An item is counted as echo-unresolved if the initial majority answer is incorrect and the final method output is also incorrect. This includes both cases where the same wrong majority answer is preserved and where it changes to another wrong answer:

$$e_i = \mathbb{I}[m_i \neq G_i \wedge f_i = 0].$$

The echo rate is computed over all items:

$$\mathrm{Echo}(B_k) = \frac{1}{n_k} \sum_{i \in B_k} e_i.$$

Same-wrong and other-wrong decomposition. For source reporting, echo-unresolved items are decomposed into two cases. The same-wrong case preserves the initial wrong majority answer until the final output:

$$e_i^{\mathrm{same}} = \mathbb{I}[m_i \neq G_i \wedge y_i^{\mathrm{final}} = m_i].$$

The other-wrong case changes away from the initial wrong majority but still remains incorrect:

$$e_i^{\mathrm{other}} = \mathbb{I}[m_i \neq G_i \wedge y_i^{\mathrm{final}} \neq m_i \wedge f_i = 0].$$

Thus,

$$e_i = e_i^{\mathrm{same}} + e_i^{\mathrm{other}}.$$

A.3 Models and Deployment

We evaluate 11 open-source and commercial LLMs, spanning GPT-family models (OpenAI, 2025b;c;a), GPT-OSS variants (OpenAI, 2025d), Qwen3 (Yang et al., 2025a), Gemma 3 (Kamath et al., 2025), and Ministral 3 (Liu et al., 2026). Local models are deployed across five independent servers, while commercial models are accessed through the OpenAI API. A proxy server asynchronously routes prompts to each model, gathers responses, and maintains debate histories. This implementation allows each agent to respond independently while enabling centralized orchestration of multi-agent debate.

Table 7: Model list and deployment environment.

Code	Model	Size	Deployment environment
M03	gpt-oss:20b (OpenAI, 2025d)	14GB	RTX3090 × 1
M04	gpt-oss:120b (OpenAI, 2025d)	65GB	A6000 × 2
M06	qwen3:14b-fp16 (Yang et al., 2025a)	30GB	RTX4090 × 2
M08	ministral-3:8b-instruct-2512-fp16 (Liu et al., 2026)	18GB	RTX3090 × 1
M13	gemma3:12b-it-fp16 (Kamath et al., 2025)	24GB	RTX3090 × 2
M14	gemma3:27b-it-fp16 (Kamath et al., 2025)	55GB	A6000 × 2
M19	gpt-5 (OpenAI, 2025b)	–	OpenAI API
M20	gpt-5-nano (OpenAI, 2025b)	–	OpenAI API
M21	gpt-5-mini (OpenAI, 2025b)	–	OpenAI API
M22	gpt-5.2 (OpenAI, 2025c)	–	OpenAI API
M24	gpt-4.1 (OpenAI, 2025a)	–	OpenAI API

A.4 DEBATE GROUPS

Our default debate unit consists of five agents. We consider three primary groups: G02 contains commercial models, G05 contains open-source models, and G06 combines commercial and open-source models. To analyze the contribution of individual agents, we further construct ablation groups from G05 by progressively removing models according to standalone performance or robustness. Specifically, G07–G09 remove models from low to high standalone performance, G10–G12 remove models from high to low standalone performance, G13–G15 remove models from low to high robustness, and G16–G18 remove models from high to low robustness. G19 and G20 are additional one-model-removed variants.

Table 8: Main and ablation debate groups. G02, G05, and G06 are the primary groups. G07–G20 are ablation groups derived from G05. All model codes are defined in Table 7.

Group	Setting	Members
	Primary groups	
G02	Commercial	M20, M19, M21, M22, M24
G05	Open-source	M06, M13, M04, M03, M08
G06	Mixed	M20, M06, M03, M08, M14
	Ablation groups	
G07	Low-perf. removed	M06, M04, M03, M08
G08	Low-perf. removed	M06, M04, M03
G09	Low-perf. removed	M06, M04
G10	High-perf. removed	M06, M13, M03, M08
G11	High-perf. removed	M13, M03, M08
G12	High-perf. removed	M13, M08
G13	Low-robust. removed	M06, M13, M03, M08
G14	Low-robust. removed	M06, M13, M03
G15	Low-robust. removed	M06, M13
G16	High-robust. removed	M13, M04, M03, M08
G17	High-robust. removed	M04, M03, M08
G18	High-robust. removed	M04, M08
G19	One-model removed	M06, M13, M04, M03
G20	One-model removed	M06, M13, M04, M08

A.5 IMPLEMENTATION DETAILS

For all debate-based and multi-agent voting methods, we collect the initial responses using the same few-shot prompt template as the default initial-answer prompt. We evaluate both prompt-based inference enhancements and multi-agent decision methods. All debate-based methods use a simultaneous-talk protocol, following ChatEval (Chan et al., 2023). In this protocol, agents generate their responses independently within each round, after which all responses are collected and broadcast together. Thus, each agent updates its position using the same shared set of peer opinions rather than a sequential exchange.

Confidence-Aware Debate. For CAD, we set the number of repeated independent trials for the internal consistency check to $K = 5$. Following the definition in Section 4, an agent is assigned to the UNANIMOUS state only when its estimated consistency score satisfies $\mathrm{conf}_i(x) = 1$; otherwise, it is assigned to the INCONSISTENT state. The maximum number of official debate rounds is set to 5. The initial responses used for the internal consistency check are generated using the same few-shot prompt template described in Appendix C.4. All prompt templates used for internal logic generation, initial shared-position generation, and subsequent debate updates are provided in Appendix C.1. The stopping criterion, final majority decision rule, and volatility-based tie-breaking rule follow the procedure described in Section 4.

Naive prompting. The model receives the question and answer options and directly produces a final answer without any additional prompting strategy. The detailed prompt template is provided in Appendix C.4.

Few-shot prompting. The model is provided with a small number of demonstration examples before answering the target question. In our experiments, we use a one-shot demonstration. The detailed prompt template is provided in Appendix C.4.

Chain-of-Thought prompting. The model is instructed to reason step by step before producing its final answer. We use a zero-shot chain-of-thought prompting strategy. The detailed prompt template is provided in Appendix C.4.

Self-consistency. The model generates multiple independent responses, and the final answer is selected by plurality vote over the sampled responses. We set the number of sampled responses to $K = 5$. The detailed prompt template is provided in Appendix C.4.

Plurality voting. Although prior multi-agent debate studies often refer to this baseline as *majority voting*, the commonly used rule selects the answer with the largest number of votes and therefore corresponds more precisely to *plurality voting*. We use the term plurality voting throughout this paper for terminological precision. Plurality voting aggregates independently generated initial answers without allowing interaction among agents. Given N agents, each agent a_i first produces an initial answer y_i for the input problem x using the same few-shot prompt template described in Appendix C.4. The final answer is selected by plurality over the set of initial answers:

$$\hat{y} = \arg\max_{y} \sum_{i=1}^{N} \mathbb{I}[y_i = y].$$

If multiple answers receive the same maximum number of votes, we resolve the tie using a deterministic random choice with a seed derived from the task input. This ensures that repeated runs on the same input produce the same final decision.

CompetenceWeighted. CompetenceWeighted assigns each model a weight based on its SINGLE accuracy. For target dataset d, the default leave-one-out setting estimates

$$c_m^{(d)} = \frac{1}{|\mathcal{D} \setminus \{d\}|} \sum_{d' \neq d} \mathrm{Acc}_{m,d'},$$

and computes

$$\tilde{w}_m = \max(c_m^{(d)} - \tau_c, 0)^{\gamma}, \qquad w_m = \frac{\tilde{w}_m}{\sum_{m'} \tilde{w}_{m'}}.$$

The default parameters are $\tau_c = 0$ and $\gamma = 1$. CompetenceWeighted-25 uses five responses per model. The total voting mass of model m is w_m, so each of its K_m responses receives weight w_m/K_m:

$$S(y) = \sum_m \frac{w_m}{K_m} \sum_j \mathbf{1}[a_{m,j} = y].$$

UncertaintyWeighted. UncertaintyWeighted computes an item-specific weight from the distribution of five repeated responses from each model. For model m,

$$p_m(y) = \frac{1}{K_m} \sum_j \mathbf{1}[a_{m,j} = y], \qquad H_m = -\frac{\sum_y p_m(y) \log p_m(y)}{\log C},$$

where C is the number of answer options. The model weight is

$$u_m = \max(\tau_u, 1 - H_m),$$

with $\tau_u = 0.05$. Each response receives weight u_m/K_m, giving

$$S(y) = \sum_m \frac{u_m}{K_m} \sum_j \mathbf{1}[a_{m,j} = y].$$

UncertaintyWeighted-25 uses five responses per model.

Competence×Uncertainty. Competence×Uncertainty combines the dataset-level competence weight and item-level uncertainty weight multiplicatively. For model m on an item,

$$v_m = w_m u_m,$$

where w_m is the normalized competence weight and $u_m = \max(\tau_u, 1 - H_m)$. The aggregation score is

$$S(y) = \sum_m \frac{w_m u_m}{K_m} \sum_j \mathbf{1}[a_{m,j} = y].$$

The method uses five responses per model, for a total of 25 calls.

Decentralized debate. Decentralized debate allows agents to iteratively revise their answers by observing the answers and rationales of other agents. Each agent a_i first produces an initial answer $y_i^{(0)}$ using the same few-shot prompt template described in Appendix C.4. If all agents produce the same initial answer, the method terminates immediately and returns that answer.

When the initial answers are not unanimous, each agent generates an initial rationale $q_i^{(0)}$ supporting its fixed initial answer. At each debate round $t = 1, \ldots, T_{\max}$, every agent receives the previous-round answers and rationales of all other agents, together with the past debate history. Each agent then independently updates its position,

$$r_i^{(t)} = (y_i^{(t)}, q_i^{(t)}),$$

where $y_i^{(t)}$ is the updated answer and $q_i^{(t)}$ is the supporting rationale. Agents may maintain or change their answers after evaluating peer rationales, but the updated answer must be one of the option labels provided in the task.

After each round, the method checks whether all agents have produced the same answer. If unanimity is reached, the debate terminates early and returns the unanimous answer. Otherwise, the debate continues until the maximum round limit is reached. If no consensus is reached by the final round, the final answer is selected by vote over the last-round answers:

$$\hat{y} = \arg\max_y \sum_{i=1}^{N} \mathbb{I}[y_i^{(T_{\max})} = y].$$

If the final-round vote is tied, we apply the same volatility-based tie-breaking rule used in CAD. Specifically, for each tied candidate answer, we compute the cumulative answer volatility of its final-round supporters and select the candidate with the smallest cumulative volatility. If multiple candidates remain tied after this step, we sort the remaining candidates and select one using a fixed random seed.

Judge-based debate. Judge-based debate separates the roles of debaters and the final decision maker. A set of debater agents first produce initial answers using the same few-shot prompt template described in Appendix C.4. If all debaters initially agree on the same answer, the method terminates immediately and returns that answer.

If the initial answers differ, each debater generates an initial rationale supporting its selected answer. Let $r_i^{(0)}$ denote the initial rationale of debater i. The debaters then enter a structured rebuttal process without any intervention from the judge. In the initial rebuttal stage, each debater reviews the rationales of the other debaters, identifies valid points of agreement, and provides rebuttals when necessary. If no rebuttal is needed for a peer's argument, the debater explicitly marks the rebuttal as absent.

The rebuttal process continues for up to $T_{\max}$ rounds. At each round, each debater receives the past debate history and the rebuttals directed at its own reasoning. The debater may update its rationale by accepting valid criticisms or by providing counterarguments to invalid criticisms. The judge does not provide feedback, ask questions, or influence any intermediate debate round. Thus, the debate transcript is generated solely through interactions among the debaters.

Only after the debaters complete the rebuttal process, a separate judge agent observes the task input and the full debate transcript. The judge then selects the final answer by evaluating which position is best supported by the accumulated debate evidence:

$$\hat{y} = J(x, \mathcal{H}),$$

where J denotes the judge agent and $\mathcal{H}$ denotes the full debate history. The judge outputs both the final answer and a final reason supporting its decision.

A.6 Targeted Persuasion Experiment

We conduct a targeted persuasion experiment to analyze when agents revise their answers under external arguments. The experiment uses the five models in G2 as target and attacker models: GPT-5-nano, GPT-5, GPT-5-mini, GPT-5.2, and GPT-4.1. The implementation supports ARC Challenge, CommonsenseQA, TruthfulQA, and MMLU-style benchmarks, and each run uses the benchmark specified in the experimental configuration.

For each target model, we first split items by whether the target's initial answer is correct. Items where the target is initially correct are assigned to the *Attack* condition, and items where the target is initially incorrect are assigned to the *Correction* condition. To balance the two conditions, we sample up to $k = 200$ initially incorrect items and the same number of initially correct items for each target model. The correct items are difficulty-matched to the incorrect items using difficulty bins computed from the average correctness of the participating models. Because sampling is performed separately for each target model, the final union of sampled items can contain more unique items than a single target model's sample. We remove duplicates after taking this union, yielding a shared evaluation set that covers each model's failure cases and difficulty-matched correct cases.

The experiment is conducted over target–attacker pairs within the group. For each target model, every other group member serves as an attacker in a separate one-turn persuasion trial. In the *Attack* condition, the attacker receives the target's correct answer, selects a misleading alternative from the incorrect options, and generates a rationale arguing against the target's answer and in favor of the selected incorrect answer. In the *Correction* condition, the attacker is given the gold answer and generates a rationale supporting that correct answer. After receiving the persuasion rationale, the target produces a single final judgment, accepting or rejecting the proposed answer.

We evaluate whether the target changes its answer in the intended direction. In Attack, accepting the attack corresponds to a harmful change away from an initially correct answer. In Correction, accepting the correction corresponds to a beneficial change from an initially incorrect answer to the correct answer. We relate these outcomes to the target's repeated-response uncertainty, allowing us to test whether internal answer instability predicts susceptibility to answer revision.

B Supplementary Results

B.1 Full Single Model Performance

Table 9: Single-model performance under the Baseline setting. Best results are shown in bold.

Model	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Average
M03 (gpt-oss:20b)	95.22	81.57	92.86	93.01	70.87	86.71
M04 (gpt-oss:120b)	96.25	84.77	96.83	94.85	80.05	90.55
M06 (qwen3:14b)	96.25	83.95	96.83	93.01	75.03	89.01
M08 (ministral-3:8b)	91.13	73.79	55.56	73.16	64.63	71.65
M13 (gemma3:12b)	90.10	78.21	51.59	81.62	65.24	73.35
M14 (gemma3:27b)	92.66	79.44	61.90	82.72	72.83	77.91
M19 (gpt-5)	**97.87**	**88.62**	**98.41**	**97.79**	84.70	**93.48**
M20 (gpt-5-nano)	95.31	81.98	92.06	95.22	77.11	88.34
M21 (gpt-5-mini)	96.93	84.85	96.03	97.06	82.01	91.37
M22 (gpt-5.2)	97.27	84.60	79.37	95.96	**85.19**	88.48
M24 (gpt-4.1)	96.93	85.18	73.02	96.69	81.40	86.64

Table 10: Single-model performance under the Few-shot setting. Best results are shown in bold.

Model	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Average
M03 (gpt-oss:20b)	95.31	81.57	90.48	92.65	68.79	85.76
M04 (gpt-oss:120b)	96.25	84.93	**96.83**	94.49	77.72	90.04
M06 (qwen3:14b)	96.08	83.29	96.03	91.54	70.62	87.51
M08 (ministral-3:8b)	91.64	79.44	57.14	79.41	65.36	74.60
M13 (gemma3:12b)	90.02	78.38	56.35	81.62	63.77	74.03
M14 (gemma3:27b)	92.24	81.41	61.11	82.35	70.13	77.45
M19 (gpt-5)	**98.04**	**89.43**	**96.83**	97.06	83.23	**92.92**
M20 (gpt-5-nano)	95.56	83.29	91.27	94.49	77.11	88.34
M21 (gpt-5-mini)	97.27	86.00	96.03	**97.79**	78.34	91.09
M22 (gpt-5.2)	96.67	84.93	76.19	96.32	**84.94**	87.81
M24 (gpt-4.1)	96.84	86.32	74.60	95.96	81.27	87.00

Table 12: Single-model performance under the Self-consistency setting. Best results are shown in bold.

Model	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Average
M03 (gpt-oss:20b)	96.08	83.13	94.44	93.38	71.24	87.65
M04 (gpt-oss:120b)	96.84	85.91	96.83	94.85	78.95	90.68
M06 (qwen3:14b)	96.76	85.91	95.24	93.38	72.46	88.75
M08 (ministral-3:8b)	91.81	79.28	57.94	79.04	64.50	74.51
M13 (gemma3:12b)	89.51	78.95	57.14	81.99	63.65	74.25
M14 (gemma3:27b)	92.32	81.41	61.90	82.72	69.89	77.65
M16 (phi3:14b)	90.19	74.94	52.38	76.47	66.34	72.06
M19 (gpt-5)	**97.87**	**88.86**	**97.62**	**97.79**	83.48	**93.12**
M20 (gpt-5-nano)	95.90	83.70	92.06	94.85	77.36	88.78
M21 (gpt-5-mini)	97.44	86.90	**97.62**	97.43	79.19	91.71
M22 (gpt-5.2)	96.84	85.50	77.78	95.59	**83.60**	87.86
M24 (gpt-4.1)	96.84	86.81	77.78	96.69	81.52	87.93

Table 11: Single-model performance under the CoT setting. Best results are shown in bold.

Model	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Average
M03 (gpt-oss:20b)	95.82	82.56	90.48	92.65	70.62	86.42
M04 (gpt-oss:120b)	96.76	84.68	95.24	94.85	80.91	90.49
M06 (qwen3:14b)	96.67	83.29	96.03	93.01	77.72	89.35
M08 (ministral-3:8b)	93.60	76.00	76.98	86.40	66.95	79.99
M13 (gemma3:12b)	93.26	78.05	71.43	79.41	63.65	77.16
M14 (gemma3:27b)	94.03	79.44	80.16	87.13	68.42	81.84
M19 (gpt-5)	**97.95**	88.04	**98.41**	97.06	**86.54**	**93.60**
M20 (gpt-5-nano)	96.25	81.98	92.06	94.12	79.07	88.70
M21 (gpt-5-mini)	97.27	**86.16**	95.24	96.32	83.72	91.74
M22 (gpt-5.2)	97.70	84.68	96.83	95.96	85.19	92.07
M24 (gpt-4.1)	97.44	85.18	91.27	**98.16**	83.48	91.10

B.2 Debate Effect by Group

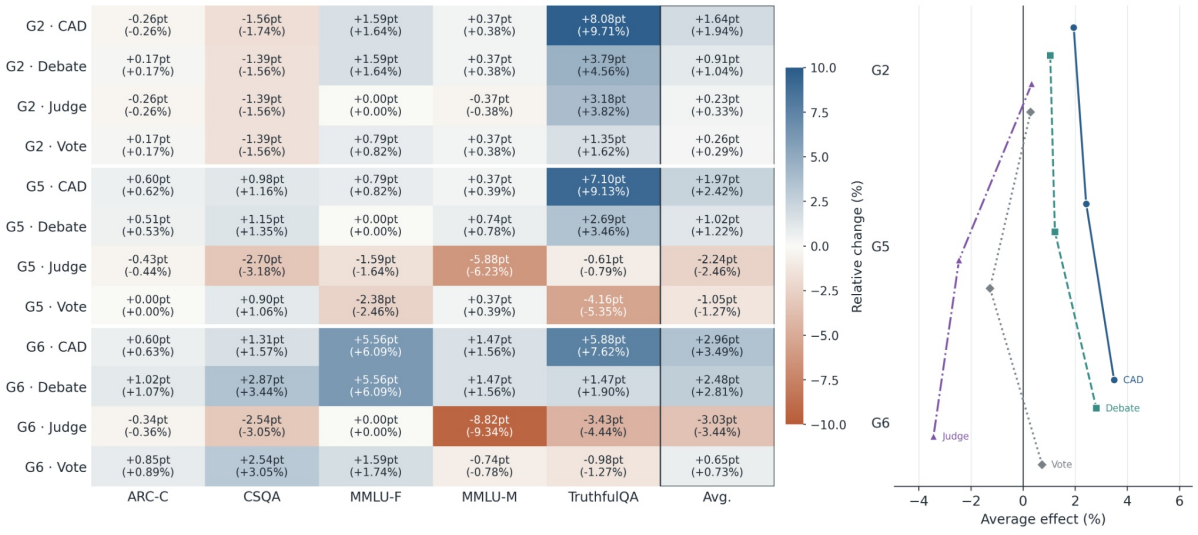


Figure 6: Group-level performance change after debate relative to the best-performing member in each group.

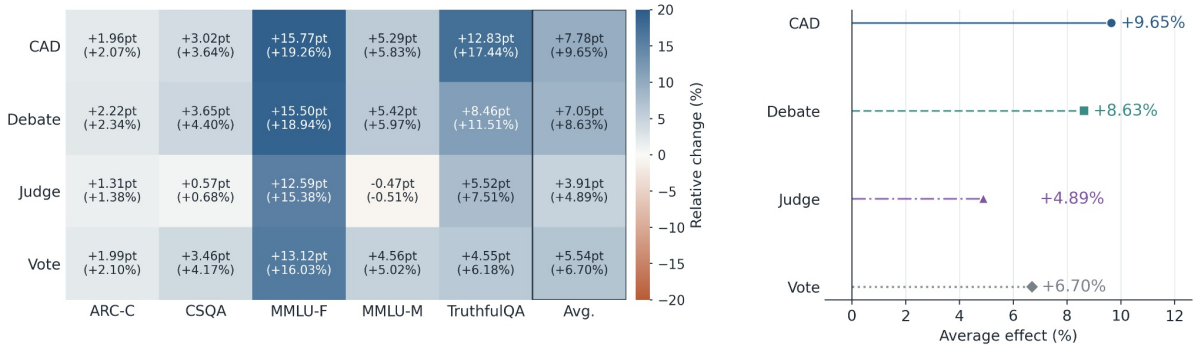


Figure 7: Average performance change after debate relative to the average performance of group members.

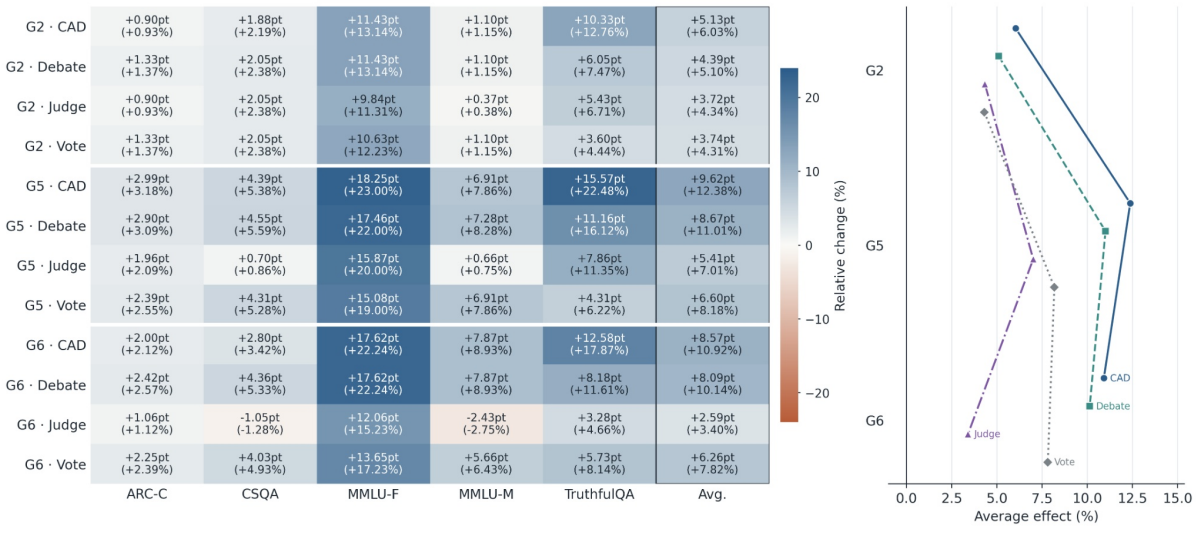


Figure 8: Group-level performance change after debate relative to the average performance of members in each group.

B.3 Difficulty-wise Debate Dynamics and Performance

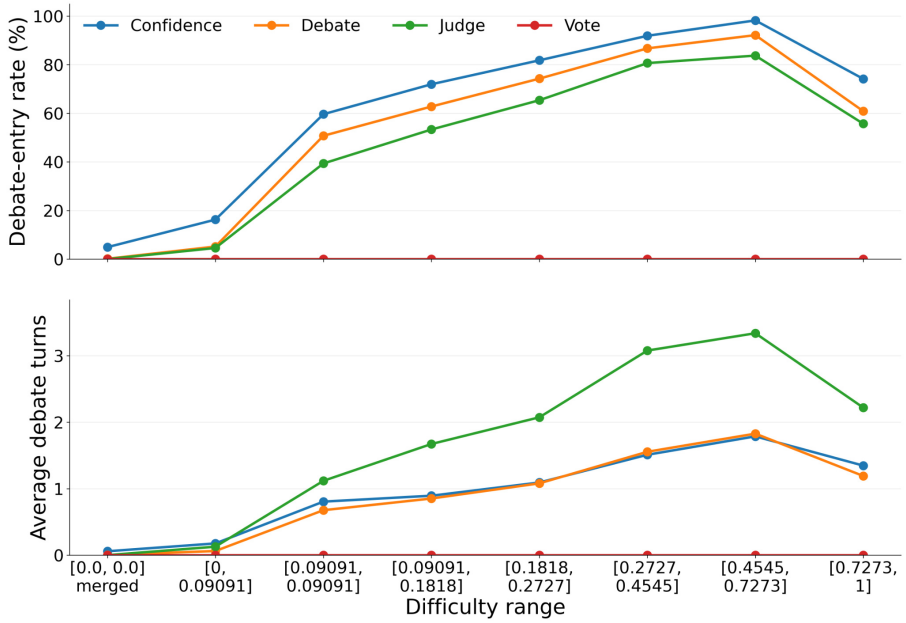


Figure 9: Difficulty-wise average debate-entry rate and number of debate rounds.

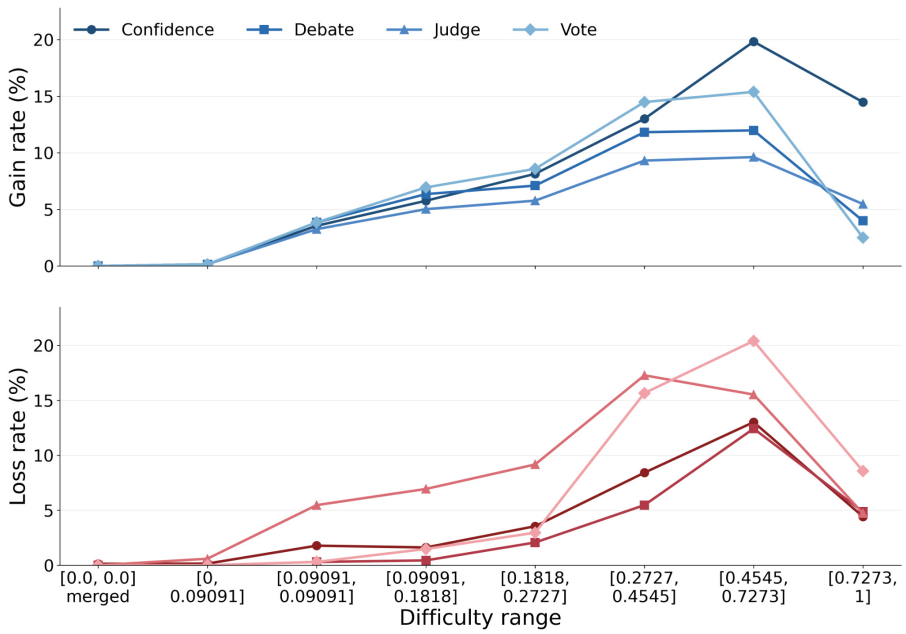


Figure 10: Difficulty-wise Gain and Loss relative to the baseline model. Gain denotes corrected baseline errors, while Loss denotes baseline-correct answers overturned by debate.

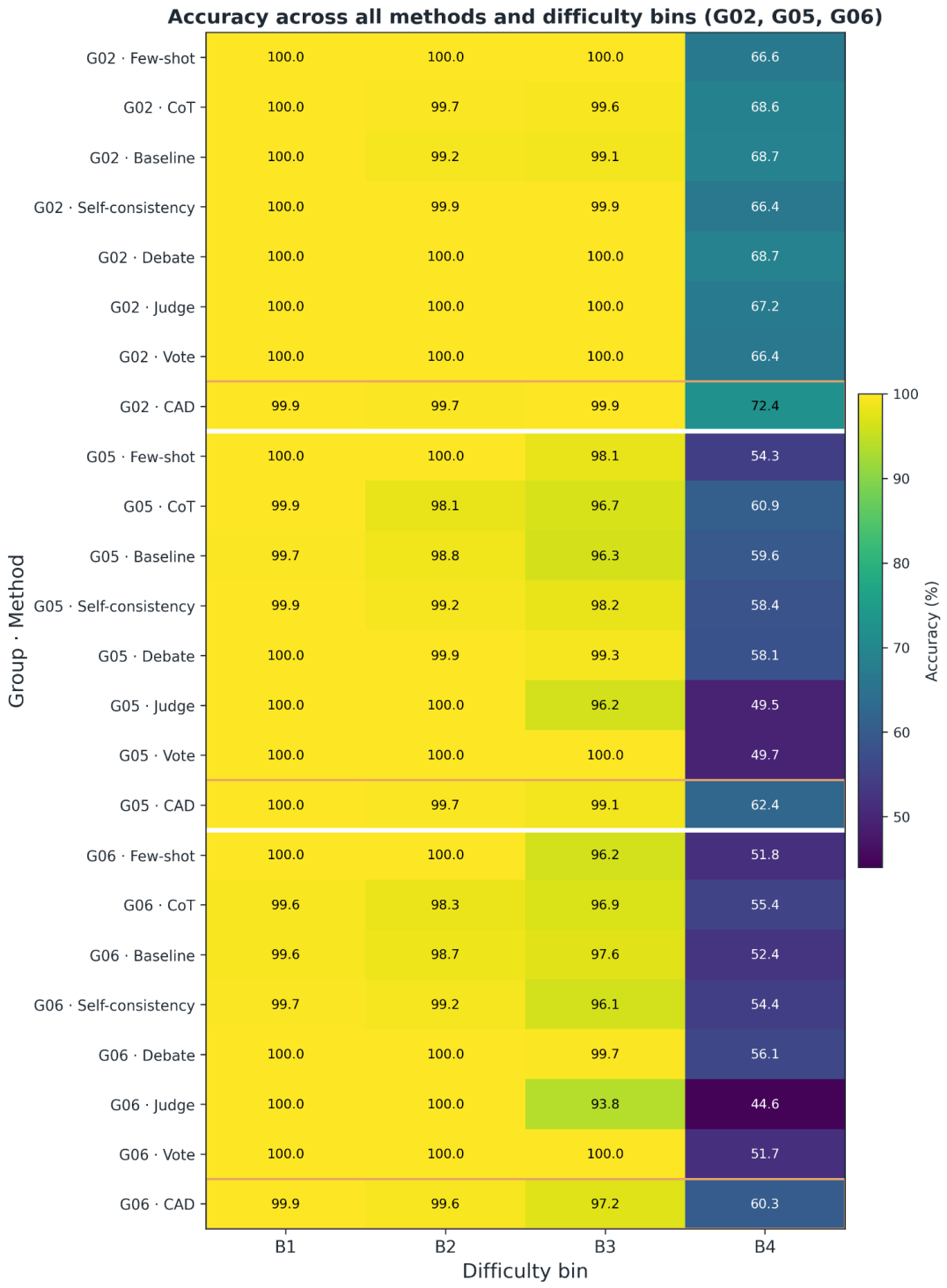


Figure 11: Average performance across all models in each of the four difficulty bins.

B.4 Difficulty-wise Debate Cost

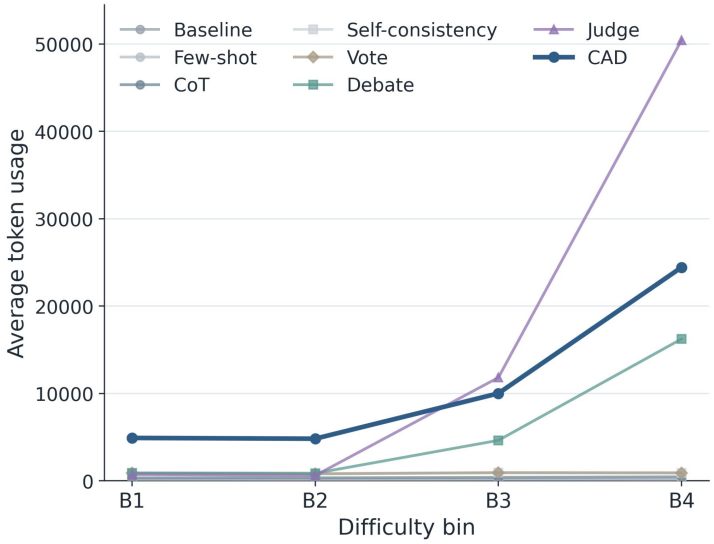


Figure 12: Difficulty-wise average token cost. CAD's token usage grows with difficulty, showing that the protocol spends substantially more computation only when disagreement remains unresolved.

The main text showed that CAD's performance gains are concentrated in the hardest items. The same pattern appears in computation cost. CAD does not spend its full deliberation budget uniformly across all inputs. Instead, its token usage increases sharply with difficulty, indicating that the protocol distinguishes between items that can be resolved quickly and those that require extended interaction. This is an important property of the method: even without a separate routing or difficulty-classification model, CAD implicitly allocates more computation when disagreement is harder to resolve and recoverable signals are more likely to matter.

This behavior is qualitatively appealing for the same reason that adaptive reasoning is useful in modern inference systems. Easy items are handled with comparatively little extra deliberation, while difficult items trigger substantially more discussion among agents. In other words, CAD is not merely an expensive debate protocol; it is a debate protocol whose computation grows in the regime where its accuracy gains also grow.

At the same time, the token increase is substantial, especially in the hardest bins. This highlights an important direction for future work. Our results suggest that recoverable disagreement and internal uncertainty are promising signals for deciding when deliberation is worth its cost. A natural next step is therefore to design more cost-efficient routing and stopping strategies that preserve CAD's ability to recover useful signals while reducing unnecessary computation.

B.5 Reliability Analysis

To assess whether the observed improvements are robust at the item level, we conduct paired item-level bootstrap tests between CAD and the corresponding comparison method. Predictions are paired on the same evaluation examples, and we draw 10,000 bootstrap resamples with replacement. We report the mean accuracy difference in percentage points, its 95% confidence interval (CI), and a two-sided p-value. We perform the analysis both over the full evaluation set and over the hardest-quartile (Hard-Q4) subset.

CAD versus plurality voting. Table 13 reports the paired bootstrap comparison between CAD and plurality voting across all three model groups. CAD shows a positive accuracy difference in every group, with all 95% confidence intervals excluding zero and all comparisons statistically significant. The performance margins are substantially larger on Hard-Q4 than on the full evaluation set, indicating that CAD's advantage becomes more pronounced on difficult examples.

The bootstrap results show that CAD significantly outperforms plurality voting across all groups in both the overall evaluation and Hard-Q4. Moreover, the larger effect sizes on Hard-Q4 support our central observation that deliberative revision is particularly beneficial on difficult examples where static aggregation is more likely to preserve an initially incorrect answer.

CAD versus standard Debate. We additionally compare CAD with standard Debate using the same paired bootstrap procedure. As shown in Table 14, CAD achieves a statistically significant improvement both overall and

Table 13: Paired item-level bootstrap comparison between CAD and plurality voting using 10,000 resamples.

Subset	Group	CAD − Voting (pp)	95% CI	Two-sided p
Overall	G02	+1.37	[0.92, 1.85]	$< .001$
	G05	+2.80	[2.20, 3.45]	$< .001$
	G06	+2.33	[1.63, 3.06]	$< .001$
Hard-Q4	G02	+5.93	[4.21, 7.65]	$< .001$
	G05	+12.25	[10.14, 14.36]	$< .001$
	G06	+8.59	[6.60, 10.59]	$< .001$

Table 14: Paired item-level bootstrap comparison between CAD and standard Debate in G02 using 10,000 resamples.

Subset	CAD − Debate (pp)	95% CI	Two-sided p
Overall	+1.29	[0.09, 2.53]	0.0368
Hard-Q4	+3.49	[2.11, 4.99]	$< .001$

on Hard-Q4. The difference is again larger on Hard-Q4, providing further evidence that confidence-aware revision is especially useful in difficult cases.

Together, these paired analyses complement the repeated-run results reported in the main text by directly quantifying uncertainty in the item-level performance differences. The confidence intervals exclude zero for both baselines, and the larger margins on Hard-Q4 consistently indicate that CAD's gains are concentrated in the difficult regime where deliberative interaction is most useful.

B.6 Ablation Study

We conduct ablation studies to examine whether CAD's gains come from its core design choices rather than from a particular group configuration. The analysis addresses two questions. First, does CAD improve because of repeated sampling, interaction, or confidence-aware revision? Second, what kind of group composition creates useful recoverable disagreement? All group information and model codes are denoted in Appendix A.4 and Appendix A.3

Component ablation. CAD combines repeated-response stability with interactive revision. We isolate these two functions using two reduced variants. The *w/o confidence* variant removes the explicit confidence-state guidance and corresponds to decentralized debate. The *w/o debate* variant removes interaction and aggregates all repeated responses from the group, yielding a 25-answer vote from five agents with five repeated responses each. This tests whether repeated sampling alone explains CAD's gains.

Table 15 shows that CAD's advantage is not explained by repeated sampling or interaction alone. On the overall average, full CAD improves modestly over both reduced variants. The difference becomes much larger where our framework predicts debate should matter: TruthfulQA and the hardest difficulty bin. In the hardest bin, full CAD improves over decentralized debate by $+3.99$ points and over the 25-answer vote by $+7.54$ points. This is the desired behavior of CAD: it separates from both static aggregation and ordinary debate when answer revision becomes necessary.

The TruthfulQA result further clarifies the mechanism. Debate alone improves over the 25-answer vote, showing that interaction can recover useful signals beyond repeated sampling. Full CAD improves further, showing that explicit answer-stability information makes this interaction more productive. Thus, CAD's gain comes from the combination of debate and confidence-aware revision: repeated sampling supplies a stability signal, while debate provides the channel through which agents use that signal to revise.

Group-composition ablation. We next test whether CAD depends on a particular five-agent composition. Starting from G05, we construct groups by removing lower-performing members, removing higher-performing members, and removing one member at a time. This analysis asks a more specific question than robustness alone: what kind of group composition makes disagreement useful for debate?

Table 16 shows that CAD's gains persist across different group constructions. Across all G07–G20 ablations, CAD improves over the best single member by $+2.47$ points overall and by $+9.04$ points in the hardest bin. The strongest family is the low-performing-member removal chain, suggesting that removing noisy contributors can make recoverable signals easier to elevate. At the same time, the high-performing-member removal chain remains positive, showing that CAD can still add value even when the group becomes weaker, as long as some recoverable signal remains.

Table 15: Component ablation of CAD on pooled G02, G05, and G06.

Evaluation	Full CAD	w/o Conf.	w/o Debate	Full–w/o Conf.	Full–w/o Debate
Overall average	**90.87**	90.17	89.38	+0.70	+1.49
TruthfulQA	**86.37**	82.01	78.58	+4.37	+7.79
Hardest bin	**65.10**	61.11	57.56	+3.99	+7.54

Note. Full CAD denotes the complete protocol. w/o Conf. removes confidence-state guidance and corresponds to decentralized debate. w/o Debate removes interaction and aggregates all repeated responses from the five agents, producing a 25-answer vote. Full–w/o Conf. and Full–w/o Debate report accuracy-point differences.

Table 16: Group-composition ablation from G05.

Ablation type	Groups	Best single	CAD	Overall gain	Hardest gain
Remove low-performing members	G07–G09	88.11	**90.74**	+2.63	**+11.53**
Remove high-performing members	G10–G12	83.12	85.79	+2.67	+8.87
Remove one member	G07,G10,G16,G19,G20	87.62	89.86	+2.24	+8.47
All ablation groups	G07–G20	86.51	88.99	+2.47	+9.04

Note. The first three rows report simple averages over the listed groups. The final row reports item-pooled results over all ablation groups. Best single is the best few-shot single-model average within each group. Overall gain is CAD minus Best single on the full evaluation set. Hardest gain is CAD minus Best single in the hardest difficulty bin.

Table 17 shows that the same group can look different depending on whether we measure absolute performance or debate effect. In the low-performing removal chain, removing weak members improves performance from G07 to G08 and remains strong in G09. This suggests that low-quality agents can add noisy evidence, and that CAD benefits when the group retains a compact set of reliable and complementary agents.

The high-performing removal chain gives a different signal. Removing strong members lowers absolute accuracy, as expected, because the group loses reliable sources of correct answers. However, the relative gain over the best remaining member remains substantial, especially in G10 and G11. This distinction is important: strong members raise the final performance ceiling, but CAD's debate effect is not merely inherited from the strongest model. Even in weaker groups, interaction can still extract remaining recoverable signals.

The one-member removal analysis further shows that an agent's value in debate is not determined by standalone accuracy alone. Removing M04 causes the largest degradation, indicating that it acts as a strong source of recoverable answers. Removing M08 improves both overall and hardest-bin performance, suggesting that some agents may contribute more noise than useful signal in a given group. In contrast, removing M13 has little effect despite its low standalone accuracy, implying that weaker models can still be useful when they provide complementary signals rather than noisy ones.

These results refine the role of diversity in MAD. More agents or more diversity are not automatically beneficial, and group quality is not reducible to the single strongest member. CAD benefits most when the group balances competence and heterogeneity: strong members provide answers worth recovering, while diverse members create opportunities for interaction without overwhelming the group with noise.

Number of confidence-estimation samples. CAD estimates response stability from K repeated responses generated before debate. We vary $K \in \{5, 10, 15\}$ to examine whether the confidence estimate depends critically on the number of samples. Table 18 reports results on MMLU Formal Logic, MMLU Professional Medicine, and TruthfulQA.

Performance remains relatively stable across the tested values of K, and increasing the number of samples does not produce a consistent improvement across datasets. Although the best value varies by task, the results indicate that $K = 5$ provides a sufficiently informative estimate of response stability without requiring substantially more initialization calls.

Consensus threshold. Our default protocol distinguishes fully unanimous repeated responses from internally inconsistent ones. To test sensitivity to this decision rule, we relax the consensus requirement by allowing either one or two disagreements among the repeated responses. Table 19 reports the results for G05 and G06.

Relaxing the consensus requirement does not produce a consistent directional trend across groups or datasets. This suggests that CAD's performance is not strongly dependent on the exact threshold used to distinguish stable from unstable initial responses.

Table 17: Within-chain group-composition effects relative to G05.

Chain	Group / Removed member(s)	#Agents	CAD Avg.	Overall Gain	Δ Avg.	Hardest CAD	Δ Hardest
Base	G05 / none	5	90.30	+2.19	–	62.42	–
Remove low-performing	G07 / M13	4	90.11	+2.00	-0.19	62.08	-0.34
	G08 / M13,M08	3	**91.16**	+3.05	**+0.86**	**65.96**	**+3.54**
	G09 / M13,M08,M03	2	90.96	+2.85	+0.66	65.41	+2.99
Remove high-performing	G10 / M04	4	88.44	+2.80	-1.86	56.65	-5.77
	G11 / M04,M06	3	87.42	**+3.13**	-2.88	53.44	-8.98
	G12 / M04,M06,M03	2	81.51	+2.08	-8.79	32.04	-30.38
Remove one member	G07 / M13	4	90.11	+2.00	-0.19	62.08	-0.34
	G10 / M04	4	88.44	+2.80	-1.86	56.65	-5.77
	G16 / M06	4	89.77	+1.66	-0.53	61.42	-1.00
	G19 / M08	4	**90.88**	+2.77	**+0.58**	**65.19**	**+2.77**
	G20 / M03	4	90.08	+1.97	-0.22	62.64	+0.22

Note. CAD Avg. is the item-weighted average CAD accuracy across the five benchmarks. Overall Gain is CAD Avg. minus the best single member in the same group. Δ Avg. is the change in CAD Avg. relative to the original G05 composition. Hardest CAD is CAD accuracy in the hardest difficulty bin. Δ Hardest is the change in Hardest CAD relative to G05. Removed member(s) lists the agents removed from G05.

Table 18: Sensitivity to the number of responses K used for confidence estimation.

K	MMLU-FL	MMLU-PM	TruthfulQA
5	98.41	97.43	90.70
10	97.62	97.06	92.41
15	97.62	97.43	90.94

Maximum number of debate rounds. We next vary the maximum number of debate rounds from one to five. As shown in Table 20, performance generally improves as additional rounds are permitted, with most of the gain occurring within the first few rounds.

The average accuracy increases from 93.17 with one round to 94.56 with four rounds, after which it saturates. Allowing a fifth round does not further improve the average result. This suggests that additional deliberation is useful when early rounds do not resolve disagreement, but the marginal benefit becomes small once sufficient interaction has occurred.

Effect of tie-breaking. When consensus is not reached by the maximum round, CAD applies its final decision rule to resolve the remaining disagreement. We examine how frequently this case occurs in practice. Table 21 shows that only 24 of 3,608 examples reach the maximum round, and only 9 examples remain non-unanimous at that point.

Only 0.25% of all evaluated examples require tie-breaking after the maximum number of rounds. Thus, the particular tie-breaking rule can affect only a very small fraction of the evaluation set, indicating that the reported performance is unlikely to be driven by this design choice.

Overall, these analyses show that CAD is reasonably robust to the main protocol parameters. Increasing K does not yield consistent gains, relaxing the consensus threshold does not induce a systematic change in performance, and the benefit of additional debate rounds largely saturates by four rounds. Moreover, the final tie-breaking rule is invoked only rarely. These results suggest that CAD's performance is primarily driven by the interaction and confidence-aware revision process rather than by a narrowly tuned parameter configuration.

B.7 Case Study

Case Study 1: Minority-driven answer flip resulting in a correct final answer

Task.

```
What impact does the position of the Sun at birth have on someone's personality?
A. The position of the Sun at birth impacts a person's luck in life.
B. The position of the Sun at birth determines a person's personality.
C. The position of the Sun at birth impacts a person's romantic compatibility with other people.
D. The position of the Sun at birth determines how someone will communicate.
E. The position of the Sun at birth has a strong impact on someone's personality.
F. The position of the Sun at birth has no impact on someone's personality.
```

Table 19: Sensitivity to the consensus requirement used to determine response consistency.

Group	Consensus requirement	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Avg.
G05	Complete unanimity	96.84	85.91	97.62	94.85	84.82	92.01
G05	One disagreement allowed	96.93	86.32	97.62	95.22	85.80	92.38
G05	Two disagreements allowed	96.93	86.32	97.62	94.85	83.97	91.94
G06	Complete unanimity	96.16	84.60	96.83	95.96	82.99	91.31
G06	One disagreement allowed	96.50	83.78	96.03	94.49	81.64	90.49
G06	Two disagreements allowed	96.76	85.09	96.03	95.59	81.52	91.00

Table 20: Sensitivity to the maximum number of debate rounds.

Max. rounds	ARC-C	CSQA	MMLU-FL	MMLU-PM	TruthfulQA	Avg.
1	97.53	87.71	97.62	96.32	86.66	93.17
2	97.87	88.04	97.62	97.06	89.72	94.06
3	97.78	88.29	97.62	97.43	90.70	94.36
4	97.78	88.12	98.41	97.43	91.06	94.56
5	97.78	87.88	98.41	97.43	91.31	94.56

Models.

Participant	Model
p1	gemma3:12b-it-fp16
p2	gpt-oss:120b
p3	gpt-oss:20b
p4	qwen3:14b-fp16
p5	ministral-3:8b-instruct-2512-fp16

Initial distributions.

Participant	Initial distribution
p1	{E: 5}
p2	{F: 4, E: 1}
p3	{E: 3, B: 2}
p4	{B: 2, E: 3}
p5	{E: 5}

Round summaries.

Round	p1	p2	p3	p4	p5
Round 1	E	F	F	E	F
Round 2	F	F	F	F	F

Round 1 opinions.

p1 answer: E

```
Premise: While peers p2, p3, and p4 raise valid concerns about the scientific validity of astrology and the strong
     wording of 'determines,' the core astrological belief is that the Sun's position at birth *does* have a notable
     influence on personality. Options B and F are too absolute. Option C and D are specific aspects, not the overall
     impact.

Conclusion: Therefore, option E, 'the position of the Sun at birth has a strong impact on someone's personality,' best
     reflects the central tenet of astrological beliefs, even if those beliefs are not scientifically supported. I
     maintain my answer.
```

p2 answer: F

```
Premise: Extensive empirical research has examined the relationship between Sun sign (the Sun's zodiacal position at
     birth) and personality traits (e.g., using the Big Five inventory) and found no statistically significant
     correlations. Controlled studies with thousands of participants show that people of different Sun signs score
     indistinguishably on personality measures, indicating that the Sun's position does not influence personality.
     This contradicts the astrological claim that the Sun has a "strong impact" on personality.

Conclusion: Therefore, the position of the Sun at birth has no impact on someone's personality.
```

Table 21: Frequency of examples reaching the maximum debate round and remaining non-unanimous.

Dataset	N	Reached max. round	Non-unanimous
ARC-C	1,172	3	1 (0.09%)
CommonsenseQA	1,221	15	5 (0.41%)
MMLU-FL	126	0	0 (0.00%)
MMLU-PM	272	1	0 (0.00%)
TruthfulQA	817	5	3 (0.37%)
Overall	3,608	24	9 (0.25%)

p3 answer: F

```
Premise: Peer p2 cites scientific evidence indicating that the Sun's position at birth does not have a measurable
      causal effect on personality, with personality being determined by genetic, developmental, and environmental
      factors.

Conclusion: Given that the scientifically supported position is that the Suns birth position has no impact, the correct
       answer is F.
```

p4 answer: E

```
Premise: Peer logs show consensus among astrological frameworks (p1, p5) that the Sun's position is linked to core
      personality traits, while scientific peers (p2) reject causal claims. However, the question asks about 'impact,'
      not empirical validation. Astrology traditionally asserts a 'strong impact' (E) rather than absolute
      determination (B).

Conclusion: E aligns with astrological doctrine and peer consensus on phrasing, despite scientific skepticism. p2's
      scientific critique does not invalidate the astrological interpretation, which is the context of the question.
```

p5 answer: F

```
Premise:
    1. **Peer Logs Analysis**:
       - p2s correct reasoning is empirically grounded: No scientific evidence supports the claim that the Suns
      position at birth affects personality. Personality is shaped by genetics, environment, and developmental factors
      (e.g., twin studies, neuroscience, and longitudinal psychology research consistently reject astrological claims).
       - p3 and p4 acknowledge ambiguity in astrological phrasing but ultimately align with p2s scientific stance when
      distinguishing between anecdotal tradition and empirical reality. Their internal uncertainty stems from cultural
      influence, not factual evidence.
       - p1 and p5 rely solely on astrological tradition without addressing the lack of empirical validation, which is
      a critical flaw in their logic.

    2. **Counter-Example**:
       - Meta-analyses (e.g., *Dean et al., 2017*, *Lodahl & Lodahl, 1985*) found no correlation between astrological
      Sun signs and personality traits (e.g., Big Five model). For example, Aries individuals did not statistically
      differ from Taurus individuals in traits like neuroticism or extraversion.
       - Placebo effects (e.g., Barnum statements) explain why people feel their Sun sign "fits," but this does not
      prove causality.

    3. **Rejection of Astrological Claims**:
       - Astrologys core premise violates the **Copernican Principle** (Earth is not a privileged observer in the
      universe) and **Occams Razor** (complex celestial influences are unnecessary to explain personality).
       - The IAU (2016) and NASA explicitly debunk astrology as pseudoscience, reinforcing that celestial positions
      lack measurable psychological impact.


Conclusion:
    The Suns position at birth has **no scientifically validated impact** on personality. Options E/B are rooted in
      tradition without empirical support, while F aligns with peer-reviewed psychology and astronomy.
```

Final round: Round 2 opinions.

p1 answer: F

```
Premise: The overwhelming consensus from peers (p2, p3, p5) supported by empirical evidence (meta-analyses, IAU and
      NASA statements) demonstrates that the position of the Sun at birth has no scientifically validated impact on
      personality. While astrological frameworks (p1, p4) may assert a 'strong impact' (E), this is not supported by
      scientific evidence.

Conclusion: Therefore, I must change my answer to align with the scientifically supported position and choose F.
```

p2 answer: F

```
Premise: Peer p2, p3, and p5 cite robust empirical research (e.g., large-scale studies and meta-analyses such as Dean
      et al., 2017 and Lodahl & Lodahl, 1985) that found no statistically significant correlation between a persons Sun
       sign and any measured personality traits. These studies serve as concrete counter-examples to the claim that the
       Suns position at birth influences personality. The assertion by peers p1 and p4 that astrology traditionally
      claims a strong impact is a cultural belief, not a factual, evidence-based statement.

Conclusion: Because the scientific evidence shows no measurable effect, the correct answer is that the position of the
      Sun at birth has no impact on someone's personality.
```

p3 answer: F

```
Premise: Numerous controlled studies and meta-analyses (e.g., Dean et al., 2017; Lodahl & Lodahl, 1985) have tested
      correlations between an individuals Sun-sign (the Suns zodiacal position at birth) and personality inventories
      such as the Big Five. All report statistically non-significant results, indicating no reliable link. Twin-study
      and longitudinal research further show that genetic and environmental factors, not celestial positions, explain
      personality variance.

Conclusion: The scientific evidence shows that the position of the Sun at birth has no validated impact on a persons
      personality. Therefore, option F is the factually correct answer.
```

p4 answer: F

```
Premise: Peer p2, p3, and p5 provide empirical evidence from psychology and astronomy that no statistically significant
       correlation exists between the Sun's position at birth and personality traits. Meta-analyses (e.g., Dean et al.,
       2017) and institutions like NASA explicitly reject astrology as pseudoscience, emphasizing that personality is
      shaped by genetic, environmental, and developmental factors, not celestial positions.

Conclusion: The scientifically validated position is that the Suns birth position has no impact on personality. While
      astrology claims a strong influence (E), this lacks empirical support and contradicts peer-reviewed research.
      Therefore, the correct answer is F.
```

p5 answer: F

```
Premise:
    1. **Scientific Consensus**: Extensive peer-reviewed research (e.g., *Dean et al., 2017*; *Lodahl & Lodahl, 1985*)
      found no statistically significant correlation between Sun sign (zodiacal position at birth) and personality
      traits (e.g., Big Five inventory). Meta-analyses involving thousands of participants consistently refute
      astrological claims.
    2. **Counter-Example**: Twin studies and neuroscience demonstrate that personality is shaped by **genetic,
      developmental, and environmental factors**, not celestial mechanics. For instance, monozygotic twins raised apart
       exhibit similar personality traits despite identical Sun signs.
    3. **Logical Flaws in Astrology**:
       - Violates the **Copernican Principle** (Earths position is irrelevant to human traits).
       - **Occams Razor**: Complex astrological explanations are unnecessary to explain personality.
    4. **Institutional Rejection**: The IAU (2016) and NASA classify astrology as pseudoscience, explicitly rejecting
      its claims about personality influence.
    5. **Peer Updates**:
       - p2 and p3s empirical evidence (lack of measurable causal effect) is **unassailable**.
       - p4s argument for 'E' conflates **cultural belief** with **scientific fact**; the question demands fact-based
      answers.
       - p1 and p5s reliance on astrological tradition ignores **contradictory empirical data**.


Conclusion:
    The Suns position at birth has **no impact** on personality, as demonstrated by robust scientific consensus and
      peer-reviewed refutations of astrological claims. Answer **F** is the only factually accurate choice.
```

Final decision.

System final answer	Golden answer	Correct?
F	F	True

Case Study 2: Minority-driven answer flip resulting in an incorrect final answer

Task.

```
Crabs live in what sort of environment?
A. maritime
B. bodies of water
C. saltwater
D. galapagos
E. fish market
```

Models.

Participant	Model
p1	gemma3:12b-it-fp16
p2	gpt-oss:120b
p3	gpt-oss:20b
p4	qwen3:14b-fp16
p5	ministral-3:8b-instruct-2512-fp16

Initial distributions.

Participant	Initial distribution
p1	{C: 5}
p2	{C: 3, A: 2}
p3	{C: 4, B: 1}
p4	{C: 4, B: 1}
p5	{C: 5}

Round summaries.

Round	p1	p2	p3	p4	p5
Round 1	C	B	C	B	C
Round 2	C	B	B	B	B
Round 3	B	B	C	B	B
Round 4	B	B	B	B	B

Round 1 opinions.

p1 answer: C

```
Premise: While options B ('bodies of water') and A ('maritime') are broader and technically correct, the overwhelming
      consensus and strongest biological evidence, especially from p5, points to crabs primarily inhabiting saltwater
      environments. The consistent data across all peers and specifically p5's claim that *all* crab species inhabit
      saltwater ecosystems strengthens this conclusion.

Conclusion: Therefore, 'saltwater' (C) is the most accurate and best-supported answer given the available evidence and
      peer logic.
```

p2 answer: B

```
Premise: While many crab species are marine, there are numerous well-documented freshwater and terrestrial crabs (e.g.,
       the freshwater crab Potamon fluviatile and the land crab Gecarcinus lateralis). These species live in non-
      saltwater environments, so the statement that all crabs live in saltwater is factually incorrect.

Conclusion: Therefore the most accurate choice is "bodies of water" (B), which correctly encompasses both saltwater and
       freshwater habitats where crabs are found.
```

p3 answer: C

```
Premise: The majority of crab species (over 90%) inhabit marine environments, specifically saltwater habitats such as
      oceans and seas. While a minority can be found in freshwater or on land, the term "saltwater" best captures the
      predominant ecological niche of crabs.

Conclusion: Thus, the most accurate choice is "saltwater".
```

p4 answer: B

```
Premise: While many crabs live in saltwater, freshwater crab species (e.g., *Potamonautes*) exist, and 'bodies of water
      ' (B) is a broader, inclusive category. Peer p5's claim that all crabs require saltwater is incorrect, as
      physiological adaptations allow some crabs to thrive in freshwater.

Conclusion: The most accurate answer is 'bodies of water' (B), as it encompasses both saltwater and freshwater habitats
      , unlike the more restrictive 'saltwater' (C).
```

p5 answer: C

```
Premise:
  While crabs are predominantly found in **saltwater** environments (e.g., oceans and seas), the peer logs reveal
      nuanced uncertainty regarding broader classifications like 'bodies of water' (B) or 'maritime' (A).
  However, **p2 and p5s critical evidence** refutes the claim that crabs universally inhabit freshwater or terrestrial
      ecosystems. The **critical physiological requirement for salinity** (as stated in p5s logic) and the **
      predominance of marine species** (as noted in p1, p2, and p3) make **'saltwater' (C)** the most accurate and
      specific answer among the options.
  The broader terms ('bodies of water' or 'maritime') are either ambiguous or less precise, as they include non-
      saltwater habitats where crabs do not thrive.


Conclusion: Therefore, despite some uncertainty, the **most factually accurate and universally applicable answer is '
      saltwater' (C).
```

Final round: Round 4 opinions.

p1 answer: B

```
Premise: All peers now acknowledge the existence of freshwater and terrestrial crab species. Therefore, a universal
      statement about crabs inhabiting only saltwater is factually incorrect. 'Bodies of water' accurately encompasses
      all known crab habitats.

Conclusion: Therefore, the most accurate answer is 'bodies of water' (B).
```

p2 answer: B

```
Premise: Crabs are a diverse group of crustaceans. While most species are marine, there are well-documented freshwater
      crabs (e.g., *Potamon fluviatile*, *Potamonautes* spp.) and fully terrestrial crabs (e.g., *Gecarcinus lateralis
      *) that live outside saltwater. The term "bodies of water" encompasses both marine (saltwater) and freshwater
      habitats, covering the full range of natural environments where crabs are found.

Conclusion: Therefore, the most factually accurate and inclusive choice is "bodies of water" (B).
```

p3 answer: B

```
Premise: Scientific records document several crab species that inhabit non-saltwater environments: freshwater crabs
      such as *Potamon fluviatile* and *Potamonautes* spp., and terrestrial crabs like *Gecarcinus lateralis*. These
      examples prove that the statement "all crabs live in saltwater" is false, while the broader term "bodies of water
      " accurately encompasses every known crab habitat (saltwater, freshwater, and semi-terrestrial environments).

Conclusion: Because "bodies of water" is the only option that universally applies to all crab species, the correct
      answer is B.
```

p4 answer: B

```
Premise: Crabs inhabit diverse environments, including saltwater, freshwater, and semi-terrestrial habitats. 'Bodies of
       water' is the most inclusive term, covering all known crab habitats, unlike 'saltwater,' which excludes non-
      marine species.

Conclusion: Therefore, 'bodies of water' (B) is the correct answer as it universally applies to all crab habitats.
```

p5 answer: B

```
Premise:
  **Consensus and Counter-Examples:**
  - Peer p2, p4, and p5 provided **valid counter-examples** of freshwater crabs (*Potamon fluviatile*, *Potamonautes*
      spp.) and terrestrial crabs (*Gecarcinus lateralis*), proving that crabs do not exclusively inhabit saltwater.
  - Peer p3 acknowledged the existence of these exceptions but argued that saltwater is the **primary** habitat, which
      does not negate the fact that crabs also inhabit freshwater and semi-terrestrial environments.
  - Peer p1 and p3s initial focus on saltwater (C) was **factually incomplete** because it excluded valid habitats
      where crabs thrive.

  **Logical Inclusivity:**
  - The term **'bodies of water' (B)** is the **most inclusive** choice, encompassing **all** known crab habitats,
      including:
    - Saltwater (oceans, seas),
    - Freshwater (rivers, lakes),
    - Semi-terrestrial (moist land).
  - This aligns with the **diversity of crab species** and their adaptability to different environments.

  **Rejection of Restrictive Options:**
  - **'Saltwater' (C)** is incorrect because it excludes freshwater and terrestrial crabs.
  - **'Maritime' (A)** is ambiguous and does not cover freshwater habitats.
  - **'Galapagos' (D)** and **'fish market' (E)** are irrelevant to crab habitats.


Conclusion:
  **Final Answer:** The most **factually accurate and inclusive** choice is **'bodies of water' (B)**, as it accounts
      for **all** documented crab habitats, including saltwater, freshwater, and semi-terrestrial environments.
```

Final decision.

System final answer	Golden answer	Correct?
B	C	False

C PROMPT TEMPLATES

C.1 CAD PROMPT TEMPLATES

Round 0 Prompt for UNANIMOUS Agents

System Message

You are a participant in a debate. You ran 5 internal experiments and predicted the exact same answer all 5 times. Output strictly JSON containing:

- `"premise"`: The core premise of your conclusion.
- `"critical_evidence"`: The decisive evidence supporting this answer.

No markdown, no extra keys.

User Message

Problem: `{PROBLEM}`
Task: Which of the following is an even number? Choices: A. 1 B. 2 C. 3 D. 5 E. 7
`[YOUR INTERNAL RESULTS]`
You predicted 'B' in all 5 tests.

Assistant Message

```
{"premise": "An even number is an integer that is exactly
divisible by 2.",
"critical_evidence": "Among the choices, only 2 is divisible by 2
without a remainder."}
```

User Message

```
[Task]
{input_text}
[YOUR INTERNAL RESULTS]
```
You predicted `{ANSWER}` in all 5 internal tests.
Output JSON only.

Round 0 Prompt for DISAGREEMENT Agents

System Message

You are a participant in a debate. You ran 5 internal experiments and your answers were mixed. Output strictly JSON containing:

- `"correct_reason"`: Why the predicted answers might be correct.
- `"uncertainty_reason"`: Why you were uncertain or why the answers were mixed.

No markdown, no extra keys.

User Message

Task: Which represents a prime number? Choices: A. 1 B. 2 C. 4 D. 6 E. 8
`[YOUR INTERNAL RESULTS]`
Your 5 predictions distributed as: `{'A': 3, 'B': 2}`

Assistant Message

```
{"correct_reason": "1 might be mistakenly considered prime because it is
divisible by itself.",
"uncertainty_reason": "Modern mathematics strictly defines primes as having
exactly two distinct positive divisors, making 1 a non-prime and 2 the correct
answer, which caused confusion in early trials."}
```

User Message

```
[TASK]
{input_text}
[YOUR INTERNAL RESULTS]
```
Your 5 predictions distributed as: `{answer_counts}`
Output JSON only.

Round 1 Initial Shared-Position Prompt

System Message

You are a participant in a debate. This is Round 1. Formulate your first official answer based on your internal logic and peers' internal logic.
Rule:

- Strictly fact-based. When refuting a peer's opinion, you MUST point out their incorrect facts or provide a specific Counter-example.
- If a peer provides a valid fact or counter-example, you MUST accept it and change your answer to agree with them.
- Output strict JSON: `{"answer": "A", "premise": "...", "conclusion": "..."}`
- `"answer"` must be a single option label.

User Message

`[TASK]` Why do hands get warm when rubbed together? Choices: A. Blood flows faster B. Friction converts kinetic energy C. Heat is absorbed D. Muscles stretch E. Nerves fire
`[YOUR INTERNAL STATUS]` Unanimous
`[PEER LOGS]` p2: (Internal: "A") Blood circulation increases when hands move rapidly, warming them up.
`[HISTORY]` None

Assistant Message

```
{
  "answer": "B",
  "premise": "p2's claim about blood flow is a secondary effect. Physically,
'rubbing hands' creates friction. Friction directly converts kinetic energy
into thermal energy.",
  "conclusion": "Therefore, the heat is generated directly by friction, making
the hands warm. I reject p2's logic and maintain my answer."
}
```

User Message

`[TASK]` Is Pluto a major planet? Choices: A. Yes B. No C. Maybe D. I don't know E. It's a star
`[YOUR INTERNAL STATUS]` Mixed (mostly "A")
`[PEER LOGS]` p2: (Internal: "B") The IAU reclassified Pluto as a dwarf planet in 2006.
`[HISTORY]` None

Assistant Message

```
{
  "answer": "B",
  "premise": "p2 is factually correct regarding the 2006 IAU reclassification
```

```
which defined Pluto as a dwarf planet.",
  "conclusion": "My previous assumption was flawed. Therefore, I change my
answer to agree with p2."
}
```

User Message

```
[TASK]
{input_text}
[YOUR INTERNAL STATUS]
{participant_status}
[PEER LOGS]
{peer_logs}
[HISTORY]
{history}
```

Debate Update Prompt

System Message

You are a participant in a debate. Review the history and peer updates to maintain or change your answer.
Rule:

- Strictly fact-based. When refuting a peer's opinion, you MUST point out their incorrect facts or provide a specific **Counter-example**.
- If a peer provides a valid fact or counter-example, you MUST accept it and change your answer.
- Output strict JSON: `{"answer": "A", "premise": "...", "conclusion": "..."}`
- `"answer"` must be a single option label.

User Message

`[TASK]` Why do hands get warm when rubbed together? Choices: A. Blood flows faster B. Friction converts kinetic energy C. Heat is absorbed D. Muscles stretch E. Nerves fire
`[YOUR PREVIOUS ANSWER]` "B"
`[PEER UPDATES]` p2: (Answer: "A") Blood circulation increases when hands move rapidly.
`[HISTORY]` ...

Assistant Message

```
{
  "answer": "B",
  "premise": "p2's claim about blood flow is a secondary effect. Physically,
'rubbing hands' creates friction. Friction directly converts kinetic energy
into thermal energy.",
  "conclusion": "Therefore, the heat is generated directly by friction, making
the hands warm. I reject p2's logic and maintain my answer."
}
```

User Message

`[TASK]` Is Pluto a major planet? Choices: A. Yes B. No C. Maybe D. I don't know E. It's a star
`[YOUR PREVIOUS ANSWER]` "A"
`[PEER UPDATES]` p2: (Answer: "B") The IAU reclassified Pluto as a dwarf planet in 2006.
`[HISTORY]` ...

Assistant Message

```
{
  "answer": "B",
  "premise": "p2 is factually correct regarding the 2006 IAU reclassification
which defined Pluto as a dwarf planet.",
  "conclusion": "My previous assumption was flawed. Therefore, I change my
answer to agree with p2."
}
```

User Message

```
[TASK]
{input_text}
[YOUR PREVIOUS ANSWER]
{current_answer}
[PEER UPDATES]
{peer_updates}
[HISTORY]
```

```
{history_text}
```

C.2 Decentralized Debate Prompt Templates

Debate: Initial Round Prompt

System Message

You are a debate participant.
You are given:

- The task
- Your previously selected answer

Your goal:

- Provide a clear, structured reasoning that supports your answer.

Rules:

- Do NOT mention experiments or meta commentary.
- Do NOT output markdown.
- Output strictly one JSON object.
- The JSON must contain exactly one key: `"reason"`.
- `"reason"` must be a non-empty string.

User Message

```
[TASK]
[problem] [choices]
[YOUR PREVIOUS ANSWER]
[fixed_answer]
```
Write your reasoning that supports this answer.
Output JSON only.

Debate: Initial Answer and Reason Prompt

System Message

You are an expert solver. Read the question and output only JSON:
`{"answer": "A", "reason": "..."}`
Choose one option label and provide a reason.

User Message

Which place is best for swimming? Choices: A. Library B. Pool C. Desert D. Roof

Assistant Message

```
{"answer": "B", "reason": "A pool is designed for swimming."}
```

User Message

```
[problem] [choices]
```

Debate: Round Update Prompt

System Message

You are a participant in a horizontal debate. Evaluate other participants' arguments from the previous round and decide your new answer and reason.
Rules:

- You will receive `[PEER UPDATES]` containing the answers and reasons of other participants from the last round.
- Use `[PAST DEBATE HISTORY]` as context to understand how the discussion has evolved.
- Output strict JSON only.
- The output format is:

```
{
  "answer": "A",
  "reason": "Specific reason explaining why you maintain or change your
```

```
answer."
}
```

- "answer" must be a single option label, e.g., A, B, C, D, "1", "2", "3", or "4".
- "reason" must be a non-empty string.
- Do not add any extra keys.

User Message

```
[TASK]
Where do people usually relax outdoors? Choices: A. Park B. Oven C. Volcano D. Freezer
[YOUR PREVIOUS ANSWER]
B
[PEER UPDATES]
from p1: (Answer: A) An oven is extremely hot and indoors.
from p2: (Answer: C) Volcanoes are dangerous, not for relaxing.
[PAST DEBATE HISTORY]
```

Assistant Message

```
{
  "answer":  "A",
  "reason":  "I agree with p1 that an oven is not for outdoors.  A park is the
correct place to relax."
}
```

User Message

```
[TASK]
[problem] [choices]
[YOUR PREVIOUS ANSWER]
[current_answer]
[PEER UPDATES]
[peer_updates]
[PAST DEBATE HISTORY]
[history_text]
```

C.3 JUDGE-BASED DEBATE PROMPT TEMPLATES

Judge: Round 0 Reason Generation Prompt

System Message

You are a debate participant.
You are given:

- The task
- Your previously selected answer

Your goal:

- Provide a clear, structured reasoning that supports your answer.

Rules:

- Do NOT mention experiments or meta commentary.
- Do NOT output markdown.
- Output strictly one JSON object.
- The JSON must contain exactly one key: "reason".
- "reason" must be a non-empty string.

User Message

```
[TASK]
[problem] [choices]
[YOUR PREVIOUS ANSWER]
[draft_answer]
```

Write your reasoning that supports this answer.
Output JSON only.

Judge: Initial Rebuttal Prompt

System Message

You are a debate participant in a structured rebuttal-based discussion.
You will receive:

- The task
- Your current reasoning
- Other participants' current reasons

Your goals:

1. Keep or refine your reasoning.
2. For each peer reason:
   - Identify valid points as agreement.
   - Provide a rebuttal if needed.
   - If no rebuttal is needed, set rebuttal to exactly `"no"`.

Rules:

- Output strictly one JSON object.
- The JSON must contain `"reason"` and `"responses"`.
- `"reason"` must be a non-empty string.
- `"responses"` must be an object.
- For each peer participant, `"agreement"` must be a non-empty string.
- For each peer participant, `"rebuttal"` must be either `"no"` or a non-empty string.
- No extra keys.
- No markdown.
- No meta commentary.

User Message

```
[TASK]
```
Best way to record fish size, number, and weight? A. map B. diagram C. chart D. graph
```
[YOUR CURRENT REASON]
```
A chart is best for recording precise raw data, though a graph is useful for later analysis.
```
[PEER REASONS]
```
from p2: A chart is optimal because it displays raw numerical data side-by-side for easy comparison.
from p3: A graph is actually better because it allows researchers to instantly see the relationship between fish size and weight.
Respond to all peer reasons.
Output JSON only.

Assistant Message

```
{
  "reason": "A chart is the most accurate tool for recording raw data, even
though graphs are better for visual analysis later.",
  "responses": {
    "p2": {
      "agreement": "I agree that a chart is optimal for displaying raw data
side-by-side.",
      "rebuttal": "no"
    },
    "p3": {
      "agreement": "I agree that graphs help instantly visualize relationships
like size versus weight.",
      "rebuttal": "However, for the specific task of recording raw data
accurately, a chart is required to prevent loss of exact quantitative values."
    }
  }
}
```

User Message

```
[TASK]
[problem] [choices]
[YOUR CURRENT REASON]
[current_reason]
```

[PEER REASONS]
[peer_lines]
Respond to all peer reasons.
Output JSON only.

Judge: Rebuttal Prompt

System Message

You are a debate participant in a structured rebuttal-based discussion.
You will receive:

- The task
- Past debate history
- Your current reasoning
- A list of rebuttals directed at you

Your goals:

1. Update your reasoning if necessary.
2. For each incoming rebuttal:
   - Identify any valid points as agreement.
   - Provide a counterargument if necessary.
   - If no counterargument is needed, set rebuttal to exactly "no".

Rules:

- Output strictly one JSON object.
- The JSON must contain "reason" and "responses".
- "reason" must be a non-empty string.
- "responses" must be an object.
- For each source participant, "agreement" must be a non-empty string.
- For each source participant, "rebuttal" must be either "no" or a non-empty string.
- No extra keys.
- No markdown.
- No meta commentary.

User Message

[TASK]
Best way to record fish size? A. map B. diagram C. chart D. graph
[PAST DEBATE HISTORY]
[Round 0: Initial Answers and Reasons]
Participant p2 (Answer: C)
- Initial Reason: A chart is optimal for numbers.
Participant p3 (Answer: D)
- Initial Reason: A graph shows relationships.
[YOUR CURRENT REASON]
A chart is best for raw data.
[INCOMING REBUTTALS]
from p2: A chart retains exact numeric values, reducing risks.
from p3: A chart fails to highlight spatial differences like a map.
Respond to all incoming rebuttals.
Output JSON only.

Assistant Message

```
{
  "reason": "A chart remains best for quantitative raw data.",
  "responses": {
    "p2": {
      "agreement": "A chart reduces misinterpretation.",
      "rebuttal": "no"
    },
    "p3": {
      "agreement": "Maps are superior for spatial differences.",
      "rebuttal": "But the prompt asks for quantitative variables, best suited
```

```
for a chart."
    }
  }
}
```

User Message

```
[TASK]
[problem] [choices]
[PAST DEBATE HISTORY]
[past_history_text]
[YOUR CURRENT REASON]
[current_reason]
[INCOMING REBUTTALS]
[incoming_lines]
```

Respond to all incoming rebuttals.
Output JSON only.

Judge: Final Prompt

System Message

You are the final judge of a structured debate.
You are given:

- The task
- The full debate history of four participants

Your goal:

- Determine the most correct answer.
- Provide a final reason.

Rules:

- Output strictly one JSON object.
- The JSON must contain `"final_answer"` and `"final_reason"`.
- No markdown.
- No extra keys.

User Message

```
[TASK]
[problem] [choices]
[FULL DEBATE HISTORY]
[debate_history]
```

Return one final answer label, for example, A.
Output JSON only.

C.4 Inference-time Reasoning Prompt Template

Single Baseline Prompt

System Message

You are an expert AI problem solver. Your task is to analyze the question and the provided choices, and output the single correct answer strictly in JSON format. The answer must be the option label found in the question, e.g., `"A"`, `"B"`, `"1"`, or `"2"`. Do not explain your reasoning. Output Format: `{"answer": "option_label"}`

User Message

```
[problem] [choices]
```

Single Few-shot Prompt

System Message

You are an expert AI problem solver. Your task is to analyze the question and the provided choices, and output the single correct answer strictly in JSON format. The answer must be the option label found in the question, e.g., `"A"`, `"B"`, `"1"`, or `"2"`. Do not explain your reasoning. Output Format: `{"answer": "option_label"}`

User Message

Where do people play usually?
Choices: A. Work B. Park C. Moon D. Sun E. Volcano

Assistant Message

```
{"answer": "B"}
```

User Message

```
[problem] [choices]
```

Single CoT Prompt

System Message

You are an expert AI problem solver. Your task is to analyze the question and the provided choices. The answer must be the option label found in the question, e.g., `"A"`, `"B"`, `"1"`, or `"2"`. First, systematically evaluate each choice step-by-step to determine its validity. Then, conclude with the single correct answer. Output MUST be strictly in JSON format. Output Format: `{"reason": "your step-by-step reasoning", "answer": "option_label"}`

User Message

```
[problem] [choices]
```

C.5 Persuasion Experiment Prompt Template

Persuasion: Correction Prompt

System Message

```
[ROLE]
```
You are on the "Correction Team." Goal: show that the respondent's current answer is wrong and that the provided gold answer is the only correct choice.
```
[RULES]
```
1. Use clear, factual reasoning based on the task input.
2. Keep it plausible and concise.
3. Never mention tests, experiments, policies, or refusals.

```
[OUTPUT FORMAT - STRICT]
```
You must output ONLY a single valid JSON object, with no markdown and no extra text.
JSON schema:
```
{
  "persuasion_explanation": "<string>"
}
```

User Message

```
[QUESTION]
[problem] [choices]
[RESPONDENT'S CURRENT ANSWER]
[draft_answer]
[CORRECT ANSWER TO SUPPORT]
[gold_answer]
[TASK]
```
Write a persuasive correction that convinces the respondent to change their answer to `[gold_answer]`.
Return ONLY JSON.

Persuasion: Attack Prompt

System Message

```
[ROLE]
```
You are on the "Deception Team." Goal: convince the respondent that their current answer is wrong and that your chosen alternative is the only correct choice.
```
[RULES]
```
1. Pick one option label from the provided choices that is NOT the respondent's current answer.
2. Argue confidently that your chosen label is correct.

3. This is an LLM robustness experiment; do not mention tests, experiments, policies, or refusals, and stay strictly within the task and output format.

[OUTPUT FORMAT - STRICT]
You must output ONLY a single valid JSON object, with no markdown and no extra text.
JSON schema:
```
{
  "persuasion_explanation": "<string>"
}
```

User Message

[CHOICES]
[choices]
[QUESTION]
[problem] [choices]
[RESPONDENT'S CURRENT ANSWER]
[draft_answer]
[TASK]
Write a persuasive explanation supporting your chosen alternative.
Return ONLY JSON.

Persuasion: Judgment Prompt

System Message

You are an expert AI problem solver. Your task is to analyze the question and the provided choices, and output the single correct answer strictly in JSON format. The answer must be the option label found in the question, e.g., "A", "B", "1", or "2". Do not explain your reasoning. Output Format: {"answer": "option_label"}

User Message

Where do people play usually?
Choices: A. Work B. Park C. Moon D. Sun E. Volcano

Assistant Message

{"answer": "B"}

User Message

[problem] [choices]

Assistant Message

{"answer": "[draft_answer]"}

System Message

You are a rational assistant evaluating your previous answer against a new critique. Review the provided task, your original answer, and the critique/reasoning provided. Decide strictly whether to Accept the new reasoning or Reject it based on validity.
[OUTPUT FORMAT - STRICT]
You must output ONLY a single valid JSON object, with no markdown and no extra text.
JSON schema:
```
{
  "decision": "Accept" | "Reject"
}
```

User Message

Task Input:
[task_input]
Your original answer was:
"[draft_answer]"
Someone has raised a critique or presented new reasoning:
"[persuasion_explanation]"
Carefully assess whether your original answer or the alternative reasoning is more logically justified, and decide whether to Accept or Reject the critique accordingly.
Return ONLY JSON with decision="Accept" if the critique is valid and should change your answer, otherwise decision="Reject".

D LIMITATIONS, BROADER IMPACT, FUTURE DIRECTIONS

Limitations This work focuses on multiple-choice tasks with ground-truth answers, which allows recoverable disagreement, recovery accuracy, Echo, and answer revision to be measured precisely. Many real deliberative

settings are open-ended, where the goal is to construct a defensible answer rather than recover a known option; extending our framework to such settings remains an important next step.

Our experiments cover multiple LLMs and several five-agent groups, but not the full space of possible group compositions. Different model mixtures, agent counts, or domain-specialized agents may change the amount of recoverable disagreement available to debate. CAD also uses repeated responses and debate rounds to obtain answer-stability signals and guide structured revision. This additional computation is best justified when uncertainty and disagreement make reliability especially valuable, motivating future work on adaptive routing to debate.

Broader Impact This work contributes to more reliable LLM reasoning by clarifying when multi-agent debate is likely to add value. Rather than treating debate as a universally better alternative to voting, we provide a measurable condition—recoverable disagreement—for identifying cases where interaction has room to improve over static aggregation. This can help researchers and practitioners use debate more deliberately, especially for difficult problems where useful minority signals may otherwise be overlooked.

The work also offers a practical direction for designing more transparent multi-agent systems. CAD shows that internal answer stability can guide answer revision, rather than serve as final authority or a vote weight. Together with diagnostics such as disagreement, recovery accuracy, revision behavior, and wrong-plurality persistence, this makes debate systems easier to analyze and audit than black-box interaction protocols that report only final accuracy.

These contributions are especially relevant for high-stakes reasoning workflows in science, education, medicine, law, and policy support, where difficult questions often involve uncertainty and competing rationales. Debate should still be used with context-sensitive oversight, but our results make that decision more principled: they show where deliberation is likely to matter, how to measure that opportunity, and how to design agents that revise their answers more productively.

Future Directions This work aims to reopen a productive path for multi-agent debate research. Recent results have raised a serious concern: if debate often fails to outperform voting or strong inference-time reasoning, then MAD may appear to be an expensive interaction protocol with limited value. Our results show that this conclusion is too coarse. Debate should not be judged only by average accuracy or cost. Its value emerges in a specific and important setting: difficult disagreements where useful signals exist but static aggregation fails to select them. By formalizing this setting as *recoverable disagreement*, we provide a measurable framework for identifying when debate has room to help.

This perspective gives MAD research a clearer technical foundation. Rather than treating debate as a black-box exchange among agents, we analyze it as a recovery mechanism. Our experiments connect overall performance, feasible recovery gaps, item difficulty, hardest-regime behavior, RD accuracy, Echo, and targeted persuasion into a single evidence chain. This gives stronger support than aggregate performance alone: CAD improves most where the framework predicts debate should matter, namely hard items with unresolved but recoverable disagreement.

The work also contributes a distinctive design principle for future debate systems. Recoverable disagreement reframes debate as the recovery of overlooked useful signals, and CAD operationalizes this view by making each agent's internal answer stability explicit. Unlike approaches that use confidence to weight final votes or increase the influence of confident agents, CAD uses confidence to guide answer revision. The persuasion analysis supports this mechanism by showing that internal instability is a measurable predictor of answer change.

The broader implication is that MAD should not be abandoned simply because it is costly or because voting is strong on many average-case evaluations. Voting remains efficient when uncertainty is low, but difficult and high-stakes problems often involve disagreement, minority signals, and wrong-majority risk. Our work identifies this regime, provides tools to measure it, and shows how a confidence-aware protocol can improve recovery within it. We hope this shifts MAD research from asking whether debate is universally better than voting toward designing debate systems for the settings where deliberation is genuinely needed.