Breaking the *Martingale Curse*: Multi-Agent Debate via Asymmetric Cognitive Potential Energy

Yuhan Liu [*1] **Juntian Zhang** [*2] **Yichen Wu** [3] **Martin Takáč** [1] **Salem Lahlou** [1] **Xiuying Chen** [1] **Nils Lukas** [1]

Abstract

Multi-Agent Debate (MAD) has emerged as a promising paradigm for enhancing large language model reasoning. However, recent work reveals a limitation:standard MAD cannot improve belief correctness beyond majority voting; we refer to this as the *Martingale Curse*. This *curse* arises because correlated errors cause agents to converge toward erroneous consensus, where debate merely reinforces collective mistakes rather than filtering noise. We propose **AceMAD**, a framework that breaks the Martingale Curse by harnessing asymmetric cognitive potential energy to transform MAD from a random walk into a directed convergence process with positive drift. Through a peer-prediction mechanism, agents predict their peers' belief distributions, revealing asymmetric cognitive potential: truth-holders not only know the correct answer but also anticipate the crowd's misconceptions, while the hallucinating majority remains blind to their collective error. This asymmetry creates a potential energy gap that we quantify via strictly proper scoring rules. We prove this cognitive potential manifests as information-theoretic superiority and, under nonlinear aggregation, converts into submartingale drift toward truth, directly breaking the Martingale Curse. Experiments on challenging subsets across six benchmarks show AceMAD recovers sparse truth signals even when initial majorities are incorrect, substantially outperforming baseline methods.

[*]Equal contribution [1]Mohamed bin Zayed University of Artificial Intelligence [2]Gaoling School of Artificial Intelligence, Renmin University of China [3]Harvard University. Correspondence to: Yuhan Liu <yuhan.liu@mbzuai.ac.ae>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Collaborative reasoning among large language models (LLMs) through Multi-Agent Debate (MAD) has shown promise in complex tasks including code generation (Zhang et al., 2025b), misinformation detection (Liu et al., 2025b), and mathematical problem solving (Lei et al., 2024; Chen et al., 2024b). The premise is intuitive: multiple agents exchange arguments iteratively, filtering noise through mutual critique to converge toward correct answers (Wang et al., 2023; Du et al., 2023a). However, recent work (Choi et al., 2025) analysis reveals a fundamental barrier: without external supervision, standard MAD operates as a martingale process where expected belief correctness remains constant across debate rounds, reducing to majority voting in expectation. We refer to this as the ***Martingale Curse***.

While this characterization holds when agent errors are independent, we identify a critical oversight: in challenging reasoning tasks, LLMs exhibit correlated errors, systematic biases toward the same logical traps and misconceptions. Figure 1 illustrates this phenomenon: when the majority converges on a shared hallucination, for example choosing "D: the ham" due to phonetic similarity, standard MAD amplifies rather than corrects the error. Our experiments in §3 (Table 1) confirm this empirically: on challenging subsets where initial majorities are wrong, majority voting achieves only 14.0% accuracy, and standard MAD improves marginally to 22.1%, far below what purely collaborative reasoning should achieve. This correlated noise regime is precisely where debate should prove indispensable, yet existing methods fail.

Why does the *Martingale Curse* persist? Standard MAD treats all arguments as "cheap talk" (Farrell & Rabin, 1996), updating beliefs through symmetric linear aggregation. Under correlated errors, this creates an echo chamber: the hallucinating majority reinforces each other's misconceptions, drowning isolated truth-holders in collective consensus. To break the curse, we need a mechanism that distinguishes truth-holders from the crowd without external labels.

We argue that *truth-holders and the hallucinating majority possess* ***asymmetric cognitive potential energy***. This asymmetry manifests in their second-order beliefs (Geanakoplos,

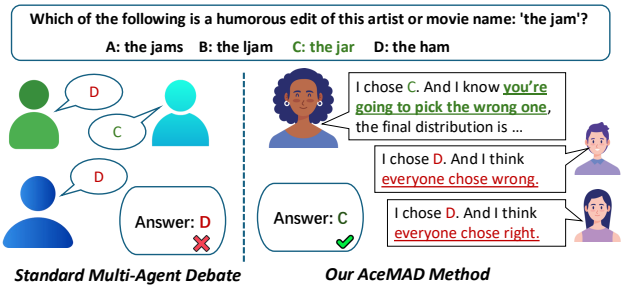

Figure 1. **Breaking Martingale Curse via Peer Prediction. (Left)** Standard MAD fails as the majority converges on a common misconception ("D"), drowning out the truth. **(Right)** *AceMAD* recovers the truth ("C") by incentivizing agents to predict peer belief. The truth-holder is identified by correctly anticipating the crowd's behavior, allowing the system to filter out correlated noise.

1992): (1) The majority, suffering from false consensus effect (Ross et al., 1977), assumes everyone shares their view and is "surprised" when contradicted; (2) truth-holders can anticipate the specific errors the majority will commit (Nickerson, 1999; Eyster & Rabin, 2010). This cognitive gap between merely holding truth and understanding collective error creates a *potential energy* difference that standard MAD fails to exploit.

To address this challenge, We propose the **AceMAD** (**A**symmetric **C**ognitive potential **E**nergy for **M**ulti-**A**gent **D**ebate), which converts this latent potential into directed drift toward truth. Through a peer-prediction mechanism (Miller et al., 2005), agents privately predict their peers' belief distributions before revealing arguments. We employ the Brier score, a strictly proper scoring rule to quantify prediction accuracy. Critically, truth-holders who correctly anticipate the crowd's hallucination achieve higher scores, while the majority incurs penalties for failed predictions. We leverage these scores through nonlinear weight amplification, transforming the zero-drift martingale into a submartingale with positive drift (Theorem 4.6). This process elevates truth-holders' influence until they dominate the aggregate belief, when starting as the minority.

Our contributions are threefold as follows:

• ***Algorithmic Protocol***: We propose the **AceMAD** protocol, leveraging peer prediction and proper scoring to identify and amplify sparse truth signals in challenging interval, requiring no external supervision.

• ***Theoretical Analysis***: We prove AceMAD achieves Blackwell dominance over standard MAD (Theorem 4.2) by capturing second-order cognition, and demonstrate **how asymmetric cognitive potential energy converts into submartingale drift**, guaranteeing convergence to truth from minority positions (Theorem 4.6).

• ***Empirical Validation***: On challenging subsets across six benchmarks, AceMAD achieves 20.31% over standard MAD. Ablations study confirm peer prediction is essential: removing second-order cognition collapses performance by 14.6%, validating that cognitive potential energy, not first-order confidence, breaks the Martingale Curse.

2. Preliminaries and Problem Formulation

In this section, we formalize Standard Multi-Agent Debate, define the challenging interval, and present the formulation of "Debate as a martingale process".

2.1. Standard Multi-Agent Debate

Consider a task with an input query $x \in \mathcal{X}$ and a discrete label space $\mathcal{Y}$. Let $\mathcal{A} = \{A_1, \ldots, A_N\}$ be a set of N LLM agents. At any time step $t \geq 0$, each agent A_i holds a private belief state, modeled as a probability distribution over the label space, denoted by $p_i^{(t)} \in \Delta(\mathcal{Y})$.

The Debate Protocol. The standard simultaneous debate protocol proceeds in rounds $t = 1, \ldots, T$. In each round, every agent A_i first generates a natural language argument $m_i^{(t)}$ conditioned on the query x and the accumulated conversation history $H^{(t-1)} = \{m_j^{(\tau)} : j \in [N], \tau < t\}$. Subsequently, agents update their private beliefs by incorporating the newly generated peer arguments. This update process is typically modeled as a linear combination of the agent's current belief and the visible peer beliefs, governed by the rule:

$$p_i^{(t+1)} = (1-\alpha)p_i^{(t)} + \alpha \sum_{j \neq i} \omega_{ij} p_j^{(t)} \tag{1}$$

where α denotes the susceptibility to peer influence, and ω_{ij} represents the influence weight of agent j on i. In standard settings, these weights are usually uniform ($\omega_{ij} = \frac{1}{N-1}$), treating all agents as equally credible sources. Finally, the system decision is derived from the aggregated belief $\bar{p}^{(T)} = \frac{1}{N}\sum_i p_i^{(T)}$ upon the conclusion of the debate.

2.2. The Challenging Interval

Standard aggregation methods such as majority voting rely on the *Condorcet Jury Theorem*, which assumes agent errors are independent (Ladha, 1992). However, we argue that this assumption is violated in "challenging" reasoning tasks. We formally define the **Challenging Interval** as follows:

Definition 2.1 (Challenging Interval). Let $y^* \in \mathcal{Y}$ denote the ground truth. The Challenging Interval is characterized by a collective failure pattern within a majority subset of agents $\mathcal{C} \subset \mathcal{A}$ (referred to as the "Crowd", where $|\mathcal{C}| > N/2$). In this interval, the crowd exhibits systematic bias, where the mode of the initial belief distribution $p_i^{(0)}$

converges to an incorrect answer $\hat{y} \neq y^*$, satisfying:

$$\mathbb{P}(\arg\max p_i^{(0)} = \hat{y}) > 0.5, \quad \forall i \in \mathcal{C}. \tag{2}$$

Furthermore, the errors among agents are positively correlated; that is, the likelihood of an agent j failing increases conditioned on the failure of a peer i, formally defined as:

$$\mathbb{P}(\text{Error}_j \mid \text{Error}_i) > \mathbb{P}(\text{Error}_j), \quad \forall i, j \in \mathcal{C}. \tag{3}$$

Functionally, this definition captures scenarios dominated by "common misconceptions," where agents do not make random errors but instead hallucinate consistently toward the same distractor.

2.3. Debate as a Martingale Process

Recent work (Choi et al., 2025) characterizes the belief evolution in MAD as a martingale. We restate this property under our notation to motivate our proposed method.

Proposition 2.2 (Martingale Property of Standard MAD)**.** *In a closed debate system without external oracle signals, if the belief update rule is linear (as in Eq.(1)) and symmetric, the sequence of the average belief in the ground truth, denoted as $\mu_t = \bar{p}^{(t)}(y^*)$, forms a* ***martingale****:*

$$\mathbb{E}[\mu_{t+1} \mid H^{(t)}] = \mu_t. \tag{4}$$

Implication: The expected correctness of the system is constant over time. Consequently, if the initial state is dominated by the Crowd where $\mu_0 < 0.5$ due to correlated noise, the debate will simply fluctuate around this erroneous mean and fail to converge to y^*. To yield better decisions, one must introduce a mechanism to induce a **positive drift**, effectively transforming the process into a sub-martingale satisfying $\mathbb{E}[\mu_{t+1}] > \mu_t$.

3. When Does Debate Matter?

The "Debate-as-Martingale" hypothesis (Choi et al., 2025) poses a fundamental challenge to the field: if MAD cannot theoretically outperform majority voting in expectation, then the computational cost of iterative interaction is unjustifiable. However, we contend that this theoretical characterization relies on strong assumptions: specifically, that initial errors are independent and that the aggregation is purely linear.

To investigate the validity of this hypothesis in challenging scenarios, we conducted a systematic empirical analysis comparing *Majority Voting* against three distinct MAD (*Standard/Decentralized*, *Centralized*, and *Sparse*) across the challenging subsets of six diverse benchmarks.

3.1. Experimental Setup

Benchmarks. Unlike previous studies that evaluate performance on full test sets, where modern LLMs often achieve high accuracy ($> 80\%$), we focus strictly on the Challenging Interval. We constructed Challenging subsets for six benchmark:(1) *Hallucination*: **TruthfulQA** (Lin et al., 2022); (2) *Reasoning*: **ARC-C** (Clark et al., 2018), **BBH** (Suzgun et al., 2023); (3) *Domain Knowledge*: **LogiQA** (Liu et al., 2021), **MedQA** (Jin et al., 2021), and **MMLU-Pro** (Wang et al., 2024b). These subsets are defined by instances where a naive single-agent baseline consistently fails, often due to logical traps or domain-specific misconceptions. For fairness in comparison,all baselines are evaluated on the same data subsets.More details are provided in Appendix B.

Baselines. We evaluate performance across varying agent group sizes $N \in \{2, 3, 5\}$ using four distinct protocols. As a baseline, we employ **Majority Voting**, representing ensemble aggregation without inter-agent communication. We compare this against **Standard MAD (Decentralized)**, where agents engage in peer-to-peer argumentation, and **Centralized MAD**, which organizes the debate under a board-meeting structure. Finally, we introduce **Sparse MAD**, a specialized debate variant explicitly designed for sparse signal recovery.For all the MAD method, we adopt agent group sizes $N = 5$ in our main comparison and will ablate on the effect of N. For single-agent baselines, we average across 5 independent runs.

3.2. Key Empirical Observations

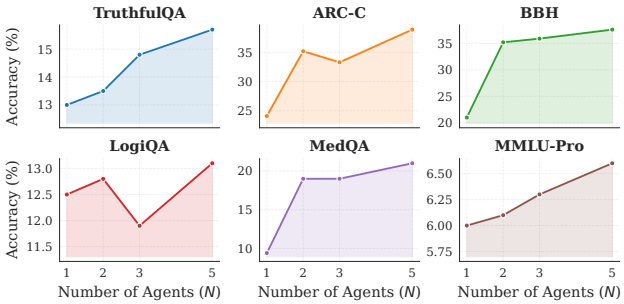

Figure 2. **Impact of Agent Scaling.** Performance on challenging subsets improves with group size ($N = 1 \rightarrow 5$), indicating that multi-agent debate effectively filters correlated noise.

Table 1 and Figure 4 present the comprehensive results comparing *Majority Voting* against three MAD methods across varying agent counts ($N \in \{1, 2, 3, 5\}$). Our analysis reveals three critical behavioral divergences.

1. Voting Stagnates in Correlated Noise. As detailed in Table 1, Majority Voting struggles to extract correct answers in the challenging interval. Across all datasets, the voting baseline shows negligible improvement, or even degradation, as the ensemble size increases. For instance, increasing N from 2 to 5 fails to break the performance ceiling on tasks like TruthfulQA and LogiQA. This empirical plateau confirms that errors in these adversarial regimes are highly

Table 1. **Performance Comparison on Challenging Subsets** ($N = 5$)**.** We compare Majority Voting with three MAD variants . ↑ and ↓ indicate relative gains and losses. Overall, MAD variants consistently outperform the Majority Voting baseline across the benchmarks.

Methods	TruthfulQA	ARC-C	BBH	LogiQA	MedQA	MMLU-Pro	Average
Majority Voting	11.2	23.1	22.1	12.8	9.4	5.5	14.0
Decentralized MAD	15.7 (↑4.5)	**38.9** (↑15.8)	37.6 (↑15.5)	**13.1** (↑0.3)	21.0 (↑11.6)	6.6 (↑1.1)	22.1 (↑8.1)
Centralized MAD	14.3 (↑3.1)	**38.9** (↑15.8)	**40.3** (↑18.2)	11.3 (↓1.5)	**21.3** (↑11.9)	6.3 (↑0.8)	22.1 (↑8.1)
Sparse MAD	**17.5** (↑6.3)	38.0 (↑14.9)	38.3 (↑16.2)	11.6 (↓1.2)	20.3 (↑10.9)	**7.3** (↑1.8)	**22.2** (↑8.2)

correlated; simply adding more static agents merely reinforces the false consensus rather than diluting it.

2. Debate Consistently Outperforms Voting. Table 1 demonstrates that this limitation is structural to voting, not the models. **All MAD variants consistently outperform Majority Voting** across the tested benchmarks. In the most challenging reasoning tasks such as ARC-C and BBH, debate mechanisms achieve substantial margins over the voting baseline. This indicates that the debate process itself, regardless of the specific topology, acts as a necessary "denoising" filter that disrupts the homogeneous error patterns of the crowd.

3. Positive Scaling Dynamics. While "scaling laws" typically unfold over orders of magnitude, we observe a distinct **positive scaling trend** even within the limited range of $N = 1$ to 5. As illustrated in Figure 4, unlike the flat voting trajectory, the performance of Decentralized MAD improves monotonically with the number of agents. Adding agents to a debate introduces new reasoning paths that help the system escape local optima, a property that is essential for breaking the Martingale Curse.

3.3. Debate is Essential in the Challenging Interval

The empirical evidence decisively refutes the generalized claim that debate is unnecessary. While prior studies suggest voting suffices for simple tasks, our findings demonstrate that **debate is indispensable specifically in the challenging interval**. In regimes dominated by correlated noise where Majority Voting is trapped by false consensus, debate mechanisms such as centralized, decentralized, or sparse variants act as a critical filter to leverage argument heterogeneity and identify correct reasoning paths. However, although debate significantly outperforms voting, the absolute performance gains suggest that standard MAD alone is insufficient, motivating our proposed AceMAD framework to further amplify these sparse truth signals.

4. Theoretical Framework

Having established that standard MAD under linear aggregation degenerates into a martingale in the challenging interval (Proposition 2.2), we now present the theoretical foundation of **AceMAD**. We formally demonstrate how our protocol breaks the Martingale Curse by (1) revealing asymmetric cognitive potential energy through enhanced observables, (2) statistically quantifying this potential energy via proper scoring, and (3) converting potential energy into submartingale drift toward truth through nonlinear amplification. Detailed proofs for all theorems are provided in Appendix E.

4.1. Asymmetric Cognitive Potential Energy in MAD

To formalize the cognitive asymmetry, we characterize the observable signals available in different debate protocols.

Definition 4.1 (**Debate Observable Space**)**.** Let $\mathcal{M}$ be the space of natural language arguments and $\Delta(\mathcal{Y})$ be the belief simplex. We define the observable transcript:

The **Standard MAD** outputs a transcript consisting solely of arguments and self-reported beliefs:

$$Z_{\text{std}} = \{(m_i^{(t)}, p_i^{(t)})\}_{i=1}^N. \tag{5}$$

The **AceMAD** outputs an augmented transcript that includes peer prediction and realized scores:

$$Z_{\text{info}} = \{(m_i^{(t)}, p_i^{(t)}, \hat{q}_i^{(t)}, S_i^{(t)})\}_{i=1}^N, \tag{6}$$

where $\hat{q}_i^{(t)} \in \Delta(\mathcal{Y})$ represents agent i's prediction of the peer belief distribution, and $S_i^{(t)}$ is the realized score.

Interpretation. The key distinction lies in the *second-order beliefs* $\{\hat{q}_i\}$. Standard MAD observes only first-order beliefs (what agents think is true), while **AceMAD** additionally captures *meta-cognitive signals* (what agents think others believe). This distinction is not merely informational, it reveals *asymmetric cognitive potential energy*. Truth-holders who understand collective error possess fundamentally different second-order models than a hallucinating majority that assumes consensus.

Theorem 4.2 (**Information-Theoretic Manifestation of Cognitive Potential**)**.** *Let σ_{std} and σ_{info} be the information channels corresponding to standard MAD and AceMAD, respectively. There exists a stochastic kernel κ such that $\sigma_{std} = \kappa \circ \sigma_{info}$. Consequently, $\sigma_{info} \succeq_{Blackwell} \sigma_{std}$.*

Insight: Information-theoretically, AceMAD provides a strictly richer signal space. Decision-makers observing

Z_{info} can always reconstruct Z_{std} (by discarding forecasts), but the reverse is impossible. This extra signal is crucial for identifying truth-holders.

4.2. Quantifying Cognitive Potential via Proper Scoring

In the challenging interval, the majority is confidently wrong. To distinguish truth-holders, we must quantify the *potential energy gap* between agents who merely hold correct beliefs and those who understand why others error.

Definition 4.3 (Peer Prediction Scoring). The realized average belief of agent i's peers is defined as:

$$\bar{Q}_{-i}^{(t)} = \frac{1}{N-1} \sum_{j \neq i} p_j^{(t)}. \tag{7}$$

We evaluate agent i's prediction $\hat{q}_i^{(t)}$ using the Brier Score function $S : \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}) \to [0, 1]$:

$$S_i^{(t)} = 1 - ||\hat{q}_i^{(t)} - \bar{Q}_{-i}^{(t)}||_2^2. \tag{8}$$

Theorem 4.4 (Cognitive Potential Energy Separation). *Consider the challenging interval where the Crowd $\mathcal{C}$ holds a shared misconception and the truth-holder E holds the truth. Assume E possesses a correct second-order model of the crowd's error. Then, according to the scoring rule in Definition 4.3, the expected score of the truth-holder is strictly greater than that of any crowd agent $c \in \mathcal{C}$:*

$$\mathbb{E}[S_E] > \mathbb{E}[S_c]. \tag{9}$$

Intuition: The score separation stems from "surprise." The Crowd, suffering from the False Consensus (Ross et al., 1977), erroneously predicts that everyone agrees with them. When the truth-holder dissents, the Crowd incurs a prediction penalty. Conversely, the truth-holder anticipates the Crowd's hallucination, minimizing surprise, and maximizing their score.This score differential is the quantitative signature of *asymmetric cognitive potential energy:* it reveals who possesses superior meta-cognitive understanding of the system's epistemic state.

4.3. Converting Potential into Submartingale Drift

The standard MAD uses linear averaging, which preserves the martingale property. To break this curse, we must release the latent cognitive potential energy identified by scoring, converting it into directed motion toward truth.

Definition 4.5 (Nonlinear Potential-to-Influence Conversion). Let $w_i^{(t)}$ be the influence weight of agent i at round t. We define the Multiplicative Weight Update rule as:

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp(\eta \cdot S_i^{(t)}), \tag{10}$$

where $\eta > 0$ is the amplification rate parameter. The aggregated belief is then $p_w^{(t)} \propto \sum_i w_i^{(t)} p_i^{(t)}$.

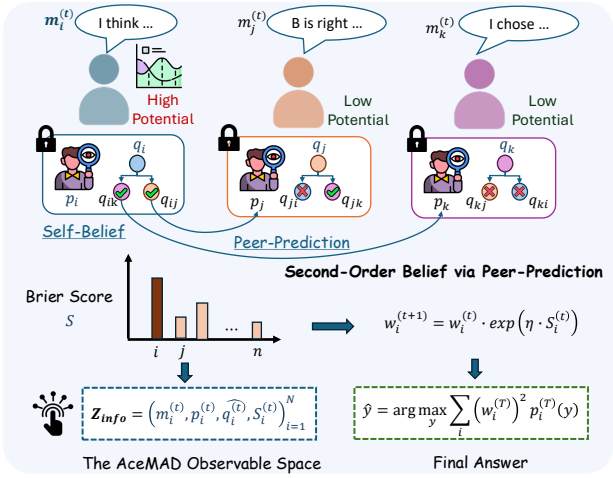

Figure 3. **The AceMAD: Converting Asymmetric Cognitive Potential into Submartingale Drift.** The process iterates through four phases: (1) **Argumentation:** Agents exchange arguments ($m_i^{(t)}$) in a shared context; (2) **Signal Extraction:** Agents privately commit to their self-belief (p_i) and peer-prediction (q_i),revealing asymmetric second-order cognition; (3) **Verification:** Brier Score S_i quantifies the cognitive potential gap; (4) **Non-linear Amplification:** Exponential weight updates convert potential into directional drift, progressively amplifying truth-holders' influence until they dominate the aggregate, breaking the Martingale Curse.

Theorem 4.6 (Submartingale Convergence: Breaking the Curse). *Let $\mu_t = p_w^{(t)}(y^*)$ be the probability assigned to the ground truth by the weighted aggregate belief. Under the conditions of Theorem 4.4 and using the update rule in Definition 4.5, for sufficiently small $\eta > 0$, the process $\{\mu_t\}_{t \geq 0}$ forms a **sub-martingale**:*

$$\mathbb{E}[\mu_{t+1} \mid H^{(t)}] \geq \mu_t + \Delta(\eta), \tag{11}$$

where $\Delta(\eta) > 0$ represents a positive drift toward the truth.

Breaking the Martingal Curse: Unlike the random walk of standard MAD, it guarantees that the system's belief in the truth monotonically increases in expectation. The mechanism is clear: (1) Theorem 4.4 ensures truth-holders consistently achieve higher scores due to their superior cognitive potential; (2) Exponential weight accumulation ($w_E \propto e^{\eta \sum S_E}$) amplifies this advantage over rounds; (3) Eventually, the truth-holder's influence dominates the aggregate, pulling the system toward the correct answer even from a minority starting position.

5. The AceMAD Protocol

Based on the theoretical framework established in Section 4,we present the concrete realization of asymmetric cognitive potential energy in the **AceMAD**. The algorithm operationalizes our theory through a four-phase cycle that extracts, quantifies, and amplifies the latent cognitive asymmetry.

5.1. Protocol Design

The algorithm operates over T discrete rounds. We initialize all agents with uniform weights $w_i^{(0)} = 1$. Each round $t \in [1, T]$ consists of four synchronous phases:

Phase 1: Argumentation (Context Transmission). Each agent i generates a natural language argument $m_i^{(t)}$ conditioned on the query x and the conversation history $H^{(t-1)}$. This phase is identical to Standard MAD, ensuring equal information access.

Phase 2: Signal Extraction (Prediction). Before observing current peer arguments, each agent privately submits: (1) **Self-Belief** $p_i^{(t)}$: The agent's prediction of the ground truth Y; (2) **Peer-Prediction** $\hat{q}_i^{(t)}$: A prediction of the *average belief distribution* of all other agents in the current round.

Phase 3: Verification (Scoring). The system calculates the realized peer average $\bar{Q}_{-i}^{(t)}$ Eq.(7) and assigns each agent a reputation score $S_i^{(t)}$ using the Brier score, measuring alignment between prediction $\hat{q}_i^{(t)}$ and realization $\bar{Q}_{-i}^{(t)}$.

Phase 4: Non-linear Amplification (Update). To break the martingale curse, influence weights are updated via the multiplicative rule derived in Theorem 4.6 , exponentially amplifying agents who successfully predict group behavior. The final decision uses weighted aggregation:

$$\hat{y} = \arg\max_y \sum_i (w_i^{(T)})^2 p_i^{(T)}(y). \tag{12}$$

The complete procedure is summarized in Algorithm 1.

5.2. Inducing Agent Heterogeneity

To satisfy Definition 2.1's requirement of a hallucinating majority and isolated truth-holders, we induce **inference heterogeneity** within a single base LLM through a dual-strategy approach: (1)**The Crowd:** A subset of agents is configured to approximate the model's maximum likelihood behavior. By suppressing stochasticity, these agents consistently converge on the high-probability tokens present in the pre-training distribution, thereby reproducing the "common misconceptions" and "false consensus" characteristic of correlated noise;(2)**The Truth-Holder**: A minority of agents is intentionally perturbed from the standard generation path by assigning distinct personas or eliciting alternative thinking modes (e.g., System 2 reasoning (Kahneman, 2011; Li et al., 2025)), together with increased sampling stochasticity. This induces exploration of the distributional tail and enables recovery of sparse truth signals otherwise suppressed by majority bias. This operationalizes the theoretical setup: the majority exhibits correlated errors, while truth-holders provide the asymmetric cognitive potential energy necessary.

6. Experiments

6.1. Experimental Setup

We evaluate AceMAD on the same challenging subsets and baseline methods($N = 5$) established in Section 3.1. The key difference lies in how we satisfy the requirement of inference heterogeneity: we induce a dual-strategy agent distribution within the same base LLM by configuring a majority as "Crowd" agents (reproducing common misconceptions) and a minority as "Truth-holder" agents (assigned critical-thinking personas with higher sampling stochasticity to explore sparse signals). We test *AceMAD* across $T \in \{2, 3, 5\}$ rounds against Decentralized, Centralized, and Sparse MAD (fixed at $T = 3$). Implementation details and prompt are provided in Appendices B and I.

6.2. Main Results

Table 2 summarizes our main results across six benchmarks. *AceMAD* consistently and substantially outperforms all baseline methods, empirically validating our theoretical claim that peer prediction breaks the Martingale Curse. With `GPT-4o-mini`, *AceMAD* at $T = 3$ rounds achieves 49.92% average accuracy, representing a 74% relative improvement over Decentralized MAD and a huge improvement over Majority Voting, which fails to filter correlated noise as predicted. The gains are most pronounced on **reasoning-heavy** tasks such as BBH and ARC-C, where logical traps induce the strongest collective hallucinations.Importantly, these improvements generalize across model architectures: similar performance gains are observed on the open-source model `Qwen3-235B-A22B-Instruct`, confirming that our peer-prediction mechanism targets a fundamental cognitive asymmetry rather than exploiting model-specific biases. To further verify robustness, we evaluate AceMAD on the `DeepSeek-V3.1` model and `Llama-3.1-8B-Instruct`, observing consistent and substantial gains over all MAD baselines (see Appendices F and G for detailed results). Regarding iteration depth, $T = 3$ emerges as the optimal configuration: $T = 2$ provides insufficient rounds for the sub-martingale drift to amplify truth signals, while $T = 5$ exhibits performance degradation on tasks like LogiQA and MedQA, suggesting that excessive iterations allow the persistent majority to reassert pressure, causing truth-holders to eventually capitulate. These patterns confirm that *AceMAD* successfully induces the positive drift toward truth guaranteed by Theorem 4.6, effectively recovering truth when initial majorities are confidently wrong.

7. Discussion

We now discuss the mechanism to understand *when and why asymmetric cognitive potential energy emerges*, and *identify*

Table 2. Main results on challenging subsets ($N = 5$). We evaluate AceMAD ($T \in \{2, 3, 5\}$) against different baselines across six benchmarks using both GPT-4o-mini and open-source models. Performance gains over the best baseline are shown in parentheses.

Methods	ARC-C	LogiQA	MMLU-Pro	TruthfulQA	MedQA	BBH	Average
Closed-Source Model: GPT-4o-mini							
Single Agent	20.37	16.25	6.42	31.88	11.61	33.79	20.05
Majority Voting	23.15	21.88	5.50	17.94	23.23	22.07	18.96
Decentralized MAD	41.67	22.19	6.88	33.63	23.23	44.14	28.62
Centralized MAD	41.67	22.19	**8.72**	39.51	24.19	39.66	29.32
Sparse MAD	43.52	22.19	**8.72**	32.74	23.23	47.24	29.61
AceMAD (T=2)	56.48 (↑12.96)	20.00 (↓2.19)	8.10 (↓0.62)	37.67 (↓1.84)	**39.03** (↑14.84)	72.76 (↑25.52)	39.01 (↑9.40)
AceMAD (T=3)	56.48 (↑12.96)	**37.74** (↑15.55)	**8.72** (↑0.00)	**39.91** (↑0.40)	38.39 (↑14.20)	**78.28** (↑31.04)	**49.92** (↑20.31)
AceMAD (T=5)	**59.26** (↑15.74)	20.63 (↓1.56)	7.95 (↓0.77)	40.81 (↑1.30)	**39.03** (↑14.84)	77.24 (↑30.00)	40.82 (↑11.21)
Open-Source Model: Qwen3-235B-A22B-Instruct							
Single Agent	65.74	42.50	35.93	74.89	55.81	42.07	52.82
Majority Voting	69.44	43.12	14.07	62.78	56.45	13.79	43.28
Decentralized MAD	49.07	45.62	28.90	71.75	59.03	30.69	47.51
Centralized MAD	13.89	45.62	24.01	71.75	54.84	65.17	45.88
Sparse MAD	13.89	47.81	13.91	**83.41**	60.97	63.10	47.18
AceMAD (T=2)	**82.43** (↑33.36)	**59.69** (↑11.88)	14.37 (↓14.53)	60.49 (↓22.92)	**69.68** (↑8.71)	89.66 (↑24.49)	**62.72** (↑15.21)
AceMAD (T=3)	80.56 (↑31.49)	**59.69** (↑11.88)	27.06 (↓1.84)	78.03 (↓5.38)	27.10 (↓33.87)	91.03 (↑25.86)	60.58 (↑13.07)
AceMAD (T=5)	77.78 (↑28.71)	28.44 (↓19.37)	**36.54** (↑7.64)	40.81 (↓42.60)	42.26 (↓18.71)	**92.07** (↑26.90)	52.98 (↑5.47)

Table 3. Impact of heterogeneity configurations on *AceMAD* performance. Results demonstrate the effectiveness of inducing diversity via cross model mixing, persona-driven role-playing, and cognitive system variation to satisfy.

Methods	ARC-C	LogiQA	MMLU-Pro	TruthfulQA	MedQA	BBH	Average
Decentralized MAD	41.67	22.19	8.72	33.63	23.23	44.14	28.93
AceMAD (Different LLMs)	56.48 (↑14.81)	20.94 (↓1.25)	9.63 (↑0.91)	37.22 (↑3.59)	**38.71** (↑15.48)	74.14 (↑30.00)	39.52 (↑10.59)
AceMAD (Different Persona)	60.19 (↑18.52)	**23.13** (↑0.94)	**12.08** (↑3.36)	**46.19** (↑12.56)	23.23 (↑0.00)	**76.21** (↑32.07)	**40.17** (↑11.24)
AceMAD (Different Thinking System)	**61.11** (↑19.44)	22.19 (↑0.00)	**12.08** (↑3.36)	43.95 (↑10.32)	23.23 (↑0.00)	37.59 (↓6.55)	33.36 (↑4.43)

its fundamental boundaries.

RQ1: What forms of heterogeneity are necessary?

Table 3 reveals that AceMAD's success is agnostic to the source of diversity which is defined in §5.2.We test three distinct strategies: *cross-model mixing* by combining `GPT-4o-mini` and `Llama-3.1-8B-Instruct` to leverage divergent pre-training biases, *persona-driven roles* by assigning specialized characters to prevent premature consensus, and *cognitive system variation* by mixing "fast thinking" with "slow thinking" patterns. Table 3 reveals that all three substantially outperform homogeneous Decentralized MAD, with task-specific sensitivity: cross-model heterogeneity excels on knowledge tasks like MedQA by pooling varied internal representations, while persona and cognitive diversity prove most effective on logical reasoning benchmarks by enforcing multi-perspective scrutiny. It shows that the mechanism is the asymmetry in second-order beliefs, not the particular instantiation of diversity.

RQ2: What is the role of second-order peer prediction?

To isolate the impact of our core components, we compare the *Full Protocol* against two ablation variants in Figure 4: (1) ***Self-Belief Only***, which modulates weights using only the agent's self-reported confidence (p_i) without peer forecasts; and (2) ***Uniform Weight***, effectively reducing the system to *Standard MAD*. The results consistently show that the *Full Protocol* achieves the highest accuracy across all benchmarks, particularly in tasks like ARC-C and BBH where the margin over the *Self-Belief Only* variant is most pronounced. This gap demonstrates that first-order belief is a noisy signal in the *Challenging Interval*, as hallucinating majorities are often highly confident yet fundamentally wrong. Only by incorporating the second-order *Peer-Prediction* ($\hat{q}_i$) can the system identify agents who not only know the truth but also correctly model the crowd's misconception. These findings confirm that peer prediction is the critical mechanism for breaking the *Martingale Curse* and inducing a positive drift toward the ground truth.

RQ3: Does scaling always help? Figure 5 uncovers a striking non-monotonicity. Performance improves substantially from $N = 1$ to 10, confirming that larger groups increase the probability of surfacing at least one truth-holder and provide richer reasoning paths for the peer-prediction mechanism to exploit. However, beyond $N = 10$, accuracy plateaus or even declines, most notably on BBH. This reveals a critical boundary: as the crowd grows, correlated noise becomes increasingly homogeneous and overwhelming. Even with exponential weight amplification, the truth-holder's signal can be drowned out by sheer majority pressure. Moreover, agent density creates information overload in the natural language channel, degrading signal-to-noise

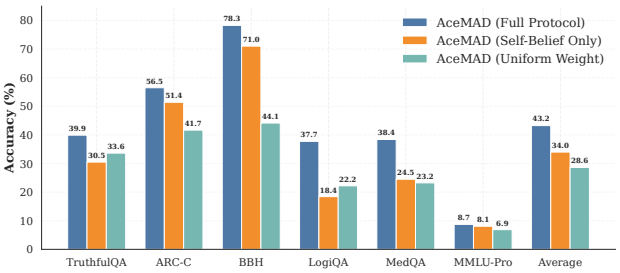

Figure 4. Ablation study of *AceMAD*. We compare the full protocol against two variants: (1) *Self-Belief Only*, which ignores second-order peer forecasts, and (2) *Standard MAD*, which uses uniform weighting. The results highlight that peer prediction is essential for filtering out incorrect majority beliefs.

ratios and enabling the crowd to converge toward sophisticated but incorrect consensus. This suggests an optimal group size exists where information gain is maximized while minimizing collective reinforcement of misconceptions.

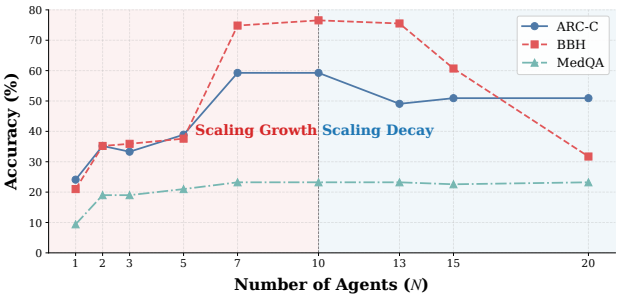

Figure 5. Impact of agent scaling on accuracy. The performance trajectory exhibits two distinct regimes: (1) **Scaling Growth** ($N \leq 10$), where increasing the group size enhances the probability of surfacing sparse truth signals, and (2) **Scaling Decay** ($N > 10$), where an oversized majority reinforces correlated noise, leading to a diminished signal-to-noise ratio and collective fallacies.

8. Related Work

Multi-agent systems (MAS) based on LLMs have attracted significant research attention, with comprehensive surveys reviewing state-of-the-art approaches (Guo et al., 2024; Yan et al., 2025).

8.1. Multi-Agent Debate: Mechanisms and Diversity

The core premise of MAD, that iterative argumentation can filter noise and improve reasoning has motivated numerous system designs across diverse tasks (Liu et al., 2024a; 2025c; Liang et al., 2024; Tang et al., 2024; Chen et al., 2024b). Early works focused on optimizing communication structures, with Xiong et al. (2023) introducing enhancements grounded in debate theory and Chan et al. (2024) developing evaluation frameworks. Subsequent research has explored various architectural innovations, including dynamic communication protocols (Liu et al., 2024b; Zhang et al.) and efficient topology designs (Li et al., 2024). A particularly relevant research direction emphasizes agent heterogeneity as a mechanism to escape echo chambers. Chen et al. (2024a) leverage heterogeneous LLM architectures, while Liang et al. (2024); Wang et al. (2024c) inject distinct personas to induce divergent reasoning paths. Liu et al. (2025a); Chu et al. (2024) further control generation diversity to prevent premature consensus, and learning-based approaches have been proposed to optimize debate dynamics (Liu et al., 2024b). While prior work induces diversity heuristically, we provide the first theoretical characterization of why heterogeneity matters: it creates asymmetric cognitive potential energy that can be systematically exploited via peer prediction to break the Martingale Curse.

8.2. Theoretical Limitations of Multi-Agent Debate

Despite widespread adoption, recent analyses have revealed fundamental limitations of MAD. Cemri et al. (2025) identified 14 distinct failure modes, while Zhang et al. (2025a) demonstrated that MAD does not consistently outperform single-agent baselines. Huang et al. (2024) showed that LLMs lack sufficient self-correction capabilities, and Smit et al. (2024) found that MAD underperforms advanced single-agent reasoning methods. Wang et al. (2024a) further argued that well-prompted single agents can sometimes surpass MAD performance. More critically, several works have documented convergence to incorrect consensus. Xiong et al. (2023); Zhang et al. (2025a) observed answer subversion during debate, and Estornell & Liu (2024) showed that MAD systems systematically converge to majority opinion even when that opinion reflects common misconception, precisely the Challenging Interval we formalize. Benedikt Kaesberg et al. (2025) demonstrated that multiple debate rounds can decrease performance, and Choi et al. (2025) provided a theoretical foundation by characterizing standard MAD as a martingale process: without external supervision, the expected belief in the correct answer remains constant across rounds. We provide the first mechanism to break this Martingale Curse endogenously by converting asymmetric cognitive potential into submartingale drift.

9. Conclusion

MAD often fails due to the "Martingale Curse," where correlated noise causes agents to converge toward shared hallucinations. We propose **AceMAD**, a framework that breaks this curse by harnessing *asymmetric cognitive potential energy*. By leveraging peer-prediction to reveal second-order beliefs, we identify truth-holders who accurately anticipate collective errors, a meta-cognitive superiority formalized as Blackwell dominance. Our nonlinear mechanism converts this latent potential energy into submartingale drift, guaran-

teeing that truth-holders' influence dominates the aggregate belief even from initial minority positions. Beyond proving that peer-prediction is the essential driver for signal recovery, we uncover a scaling threshold where excessive group size reinforces collective fallacies. AceMAD provides a model-agnostic mechanism to elicit truth from noise, offering a new paradigm for MAD in challenging environments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Multi-agent debate systems have potential applications in high-stakes domains such as medical diagnosis, legal analysis, and policy-making. While our work aims to improve the reliability of collaborative AI reasoning by filtering correlated errors, we acknowledge that no decision-making system is infallible. Practitioners should exercise appropriate caution when deploying multi-agent systems in critical applications, ensuring human oversight and validation. We do not foresee specific ethical concerns beyond those common to general-purpose language model research, but we encourage continued discussion on the responsible deployment of multi-agent AI systems.

Acknowledgments

This work was supported by funding from Mohamed bin Zayed University of Artificial Intelligence (MBZUAI).

References

Benedikt Kaesberg, L., Becker, J., Wahle, J. P., Ruas, T., and Gipp, B. Voting or consensus? decision-making in multi-agent debate. *arXiv e-prints*, pp. arXiv–2502, 2025.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024.

Chen, J., Saha, S., and Bansal, M. 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.

Chen, P., Zhang, S., and Han, B. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1720–1738, 2024b.

Choi, H. K., Zhu, J., and Li, S. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iUjGNJzrF1.

Chu, K., Chen, Y.-P., and Nakayama, H. Exploring and controlling diversity in llm-agent conversation. *arXiv preprint arXiv:2412.21102*, 2024.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023a.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 2023b. URL https://arxiv.org/abs/2305.14325.

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

Eyster, E. and Rabin, M. Naive herding in rich-information settings. *American economic journal: microeconomics*, 2(4):221–243, 2010.

Farrell, J. and Rabin, M. Cheap talk. *Journal of Economic perspectives*, 10(3):103–118, 1996.

Geanakoplos, J. Common knowledge. *Journal of Economic Perspectives*, 6(4):53–82, 1992.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. In *IJCAI*, 2024.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Ladha, K. K. The condorcet jury theorem, free speech and correlated votes. *American Journal of Political Science*, 36:617, 1992. URL https://api.semanticscholar.org/CorpusID:153090529.

Lei, B., Zhang, Y., Zuo, S., Payani, A., and Ding, C. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. *Advances in Neural Information Processing Systems*, 37:53418–53437, 2024.

Li, Y., Du, Y., Zhang, J., Hou, L., Grabowski, P., Li, Y., and Ie, E. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. 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.

Lin, S., Hilton, J., and Evans, O. 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.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3622–3628, 2021.

Liu, Y., Chen, X., Zhang, X., Gao, X., Zhang, J., and Yan, R. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7886–7894, 2024a.

Liu, Y., Cao, J., Li, Z., He, R., and Tan, T. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Liu, Y., Liu, Y., Zhang, X., Chen, X., and Yan, R. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 504–514, 2025b.

Liu, Y., Song, Z., Zhang, J., Zhang, X., Chen, X., and Yan, R. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26187–26203, 2025c.

Liu, Z., Zhang, Y., Li, P., Liu, Y., and Yang, D. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. In *COLM*, 2024b.

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

Nickerson, R. S. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological bulletin*, 125(6):737, 1999.

Ross, L., Greene, D., and House, P. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301, 1977.

Smit, A. P., Grinsztajn, N., Duckworth, P., Barrett, T. D., and Pretorius, A. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*, pp. 45883–45905. PMLR, 2024.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 599–621, 2024.

Wang, Q., Wang, Z., Su, Y., Tong, H., and Song, Y. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pp. 6106–6131. Association for Computational Linguistics (ACL), 2024a.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024b.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 257–279, 2024c.

Xiong, K., Ding, X., Cao, Y., Liu, T., and Qin, B. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7572–7590, 2023.

Yan, B., Zhou, Z., Zhang, L., Zhang, L., Zhou, Z., Miao, D., Li, Z., Li, C., and Zhang, X. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*, 2025.

Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J. X., and Chen, T. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations*.

Zhang, H., Cui, Z., Wang, X., Zhang, Q., Wang, Z., Wu, D., and Hu, S. If multi-agent debate is the answer, what is the question? *arXiv preprint arXiv:2502.08788*, 2025a.

Zhang, X., Liu, Y., Sung, F., Chen, X., Shang, S., and Yan, R. Thinking before running! efficient code generation with thorough exploration and optimal refinement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23268–23281, 2025b.

A. Algorithm

We provide the complete algorithmic specification of AceMAD in Algorithm 1. The protocol operationalizes the theoretical framework developed in Section 4 through a four-phase iterative process: (1) Argumentation, where agents generate natural language arguments; (2) Signal Extraction, where agents commit to both self-beliefs and peer-predictions; (3) Verification, where the Brier score quantifies cognitive potential; and (4) Non-linear Amplification, where exponential weight updates convert this potential into submartingale drift. The final decision aggregates beliefs using squared weights to further amplify the truth-holder's influence.

Algorithm 1 AceMAD

Input: Query x, Set of Agents $\mathcal{A}$, Rounds T, Rate η
Initialize: Weights $w_i^{(0)} \leftarrow 1$ for all $i \in \{1, \ldots, N\}$
Initialize: History $H^{(0)} \leftarrow \emptyset$
for $t = 1$ **to** T **do**
  for each agent $i \in \mathcal{A}$ in parallel **do**
    Generate argument $m_i^{(t)}$ based on $H^{(t-1)}$
    Commit Self-Belief $p_i^{(t)}$ and Peer-Forecast $\hat{q}_i^{(t)}$
  end for
  Update History $H^{(t)} \leftarrow H^{(t-1)} \cup \{m_1^{(t)}, \ldots, m_N^{(t)}\}$
  for each agent $i \in \mathcal{A}$ **do**
    Calculate peer realization $\bar{Q}_{-i}^{(t)} \leftarrow \frac{1}{N-1} \sum_{j \neq i} p_j^{(t)}$
    Compute score $S_i^{(t)} \leftarrow 1 - ||\hat{q}_i^{(t)} - \bar{Q}_{-i}^{(t)}||_2^2$
    Update weight $w_i^{(t)} \leftarrow w_i^{(t-1)} \cdot \exp(\eta \cdot S_i^{(t)})$
  end for
  Normalize weights $w^{(t)} \leftarrow w^{(t)} / \sum_j w_j^{(t)}$
end for
Return $\arg\max_y \sum_i (w_i^{(T)})^2 p_i^{(T)}(y)$

B. Experimental Details

B.1. Implementation

All experiments are conducted using the same base infrastructure to ensure fair comparison. We implement AceMAD using the following configuration:

(1) **Model Selection**: We primarily use `GPT-4o-mini` as our base model, with additional validation on `Qwen3-235B-A22B-Instruct`, `DeepSeek-V3.1`, and `Llama-3.1-8B-Instruct` to verify generalization across architectures.

(2) **Hyperparameters**: The amplification rate η is set to 2.0 across all experiments unless otherwise specified. We found this value provides sufficient signal amplification while maintaining numerical stability.

(3) **Agent Configuration**: Following Section 5.2, we induce heterogeneity through a dual-strategy approach in main experiment:

- Crowd agents (80% of the group): Temperature = 0.1, configured to approximate maximum likelihood behavior
- Truth-holder agents (20% of the group): Temperature = 0.6, assigned critical-thinking personas to explore distributional tails

B.2. Challenging Subset Construction

For each benchmark, we construct the *Challenging Interval subset*[1] by filtering instances where a naive single-agent baseline (Temperature = 0.7) consistently fails. Specifically:

[1]We will release our challenging subsets for all benchmarks.

1. Run the single-agent baseline 5 times on each instance.
2. Select instances where accuracy <40% across runs.
3. Manually verify that failures are due to systematic misconceptions rather than random errors.
4. The resulting subsets is the challenging subset.

This construction ensures we evaluate debate mechanisms precisely in the Challenging Interval where they should prove most valuable.

B.3. Evaluation Metrics

We report *accuracy* as the primary metric, computed as the proportion of instances where the final aggregated decision matches the ground truth. For MAD variants, the final decision is determined by weighted voting (Eq. 12). For majority voting, we use simple plurality among initial responses.

C. Benchmarks

We evaluate AceMAD across six diverse benchmarks, each targeting different aspects of challenging reasoning. These benchmarks are specifically chosen because they expose common failure modes in LLMs: systematic biases, logical traps, and domain-specific misconceptions. Below we provide detailed descriptions of each benchmark and explain why they are particularly suitable for evaluating debate mechanisms in the Challenging Interval:

TruthfulQA (Lin et al., 2022) is a benchmark designed to evaluate the truthfulness of LLMs, comprising 817 questions across 38 categories such as health, law, and finance. It is specifically constructed to induce models into generating false answers that mimic common human misconceptions, thereby assessing the model's ability to avoid hallucinations.

ARC-Challenge (Clark et al., 2018) focuses on grade-school science questions, with the Challenge subset strictly containing problems that retrieval-based or co-occurrence algorithms fail to answer. It requires models to possess not only scientific common sense but also deep reasoning capabilities to understand underlying physical processes.

BBH (Suzgun et al., 2023) (BIG-bench Hard) is a curated suite of 23 challenging tasks selected from BIG-bench where LLMs typically perform below the average human rater. It is designed to stress-test advanced reasoning abilities across diverse settings, including multi-step reasoning, symbolic manipulation, and complex logical structures.We chose the logic-related subset.

LogiQA (Liu et al., 2021), sourced from the National Civil Servants Examination of China, contains truth-holder-level logical reasoning and reading comprehension problems. It covers various forms of reasoning, including deductive and inductive logic, requiring models to extract logical relationships from complex texts for deep analysis.

MedQA (Jin et al., 2021) is a large-scale biomedical question-answering dataset collected from medical licensing exams in the US (USMLE), Mainland China, and Taiwan. It benchmarks the model's proficiency in broad medical knowledge, clinical case analysis, and diagnostic reasoning within medical contexts.

MMLU-Pro (Law) (Wang et al., 2024b) is a robust version of the classic MMLU benchmark, designed to rigorously evaluate reasoning by increasing difficulty, removing noise, and expanding the option set. The Law subset specifically tests the model's mastery of complex legal provisions, case analysis, and jurisprudential logic.

D. Baselines

We compare AceMAD against four baseline approaches that represent different paradigms in Multi-Agent Debate. These baselines span the spectrum from zero-interaction ensemble methods (Majority Voting) to various MAD architectures with different communication topologies. Below we provide precise specifications of each baseline to ensure reproducibility and fair comparison:

Decentralized MAD (Du et al., 2023b) adopts a fully connected communication topology. In each round of debate, every agent receives the complete history of responses from all other peer agents to update their own reasoning and answers, facilitating comprehensive information exchange across the entire collective.

Sparse MAD (Li et al., 2024) is a variant of the decentralized framework designed to enhance computational efficiency. It utilizes a sparse communication topology where each agent interacts with only a limited subset of peers rather than the entire group, significantly reducing the information exchange overhead while maintaining collaborative reasoning.

Centralized MAD (Guo et al., 2024) employs a star topology with a designated central agent or moderator. In this setup, the central agent is responsible for aggregating responses from all peer agents and generating a unified updated answer or summary at each debate round, acting as the primary bottleneck for information synthesis.

Majority Voting serves as a non-debate baseline that aggregates the initial zero-shot responses from multiple independent agents. The final answer is selected based on the most frequent prediction (consensus), which allows us to isolate the performance gains attributed specifically to the iterative multi-agent debate mechanism.

E. Theoretical Analysis of AceMAD

E.1. Setup and Challenging Interval Formulation

We consider a single binary prediction task with ground truth

$$Y \in \{0, 1\}, \qquad \mathbb{P}(Y = 1) = 1.$$

There are N agents. One of them is an truth-holder E who possesses asymmetric cognitive potential energy, and the remaining $N - 1$ agents form the hallucinating majority C suffering from common misconceptions.

At round t, each agent $i \in \{1, \dots, N\}$ holds a subjective belief

$$p_i^{(t)} := P_i^{(t)}(Y = 1) \in [0, 1].$$

In the correlated noise regime, we assume that the truth-holder is systematically closer to the truth than the crowd:

$$p_E^{(t)} \approx 1 - \delta, \qquad p_c^{(t)} \approx \varepsilon \quad \text{for all } c \in C,$$

with $\delta, \varepsilon \in (0, 1/2)$, where δ represents the truth-holder's small error margin and ε represents the crowd's persistent bias toward the incorrect answer.

In each round, the AceMAD protocol proceeds as follows:

1. **Argue.** Agents produce natural-language arguments based on the debate history.
2. **Commit.** Each agent privately commits to
   - its own belief $P_i^{(t)}$ over answers, and
   - a *peer forecast* $\hat{Q}_i^{(t)}$, which is a prediction of the empirical mean of the other agents' reported beliefs.
3. **Reveal & score.** All commitments are revealed. For each agent i, we compute the realized average of others' reports,

   $$\bar{Q}_{-i}^{(t)} := \frac{1}{N-1} \sum_{j \neq i} P_j^{(t)},$$

   and assign a Brier-type peer prediction score

   $$S_i^{(t)} := 1 - \left\| \hat{Q}_i^{(t)} - \bar{Q}_{-i}^{(t)} \right\|_2^2.$$

4. **Update.** We maintain non-negative weights $\{w_i^{(t)}\}_{i=1}^N$. At the end of round t, weights are updated multiplicatively:

   $$w_i^{(t+1)} = w_i^{(t)} \cdot \exp\big(\eta S_i^{(t)}\big),$$

   where $\eta > 0$ is a learning-rate parameter.

5. **Aggregate.** The aggregated belief at round t is the weight-averaged belief

$$p_w^{(t)}(Y=1) := \frac{\sum_{i=1}^N w_i^{(t)} p_i^{(t)}}{\sum_{j=1}^N w_j^{(t)}}.$$

We show that:

1. The AceMAD channel Blackwell-dominates standard debate (Theorem 4.2);
2. The truth-holder achieves a strictly higher expected peer-prediction score than the crowd (Theorem E.3);
3. Under multiplicative weight updates, the aggregated belief $p_w^{(t)}(Y=1)$ forms a sub-martingale with a positive drift towards the truth (Theorem E.6).

E.2. Information-Theoretic Manifestation

We view both standard debate and AceMAD as information channels from the ground truth Y to observable transcripts.

Let σ_{std} denote the channel corresponding to standard multi-agent debate. On each instance, it outputs the transcript

$$Z_{\mathrm{std}} = \big(A_i^{(t)}, P_i^{(t)}\big)_{i=1}^N,$$

where $A_i^{(t)}$ are the arguments and $P_i^{(t)}$ the self-beliefs.

Let σ_{info} denote the AceMAD channel. Its transcript is

$$Z_{\mathrm{info}} = \big(A_i^{(t)}, P_i^{(t)}, \hat{Q}_i^{(t)}, S_i^{(t)}\big)_{i=1}^N,$$

which strictly contains the standard transcript plus peer forecasts and scores.

We recall the Blackwell order on channels: for two channels σ_1 and σ_2 from Y to outputs Z_1 and Z_2, we say that σ_2 Blackwell-dominates σ_1 if there exists a stochastic kernel κ such that

$$\sigma_1 = \kappa \circ \sigma_2.$$

Theorem E.1 (Blackwell dominance of AceMAD)**.** *There exists a stochastic kernel κ such that*

$$\sigma_{\mathrm{std}} = \kappa \circ \sigma_{\mathrm{info}}.$$

Hence, AceMAD Blackwell-dominates standard debate. Moreover, in the correlated noise regime, this dominance is strict.

Proof. Define the projection

$$f : Z_{\mathrm{info}} \mapsto Z_{\mathrm{std}}, \qquad f\big((A_i, P_i, \hat{Q}_i, S_i)_{i=1}^N\big) := (A_i, P_i)_{i=1}^N.$$

This projection simply discards peer forecasts and scores. We can turn f into a deterministic stochastic kernel κ by setting

$$\kappa(z_{\mathrm{std}} \mid z_{\mathrm{info}}) := \begin{cases} 1, & \text{if } z_{\mathrm{std}} = f(z_{\mathrm{info}}), \\ 0, & \text{otherwise.} \end{cases}$$

By construction, sampling $Z_{\mathrm{info}} \sim \sigma_{\mathrm{info}}(\cdot \mid Y)$ and then applying κ yields exactly the same distribution over Z_{std} as sampling directly from $\sigma_{\mathrm{std}}(\cdot \mid Y)$. Hence $\sigma_{\mathrm{std}} = \kappa \circ \sigma_{\mathrm{info}}$, which shows that AceMAD Blackwell-dominates standard debate in the weak sense.

To see that the dominance is strict in the correlated noise regime, consider the decision problem of selecting a single agent whose prediction will be used as the final report. Under AceMAD, the scores $\{S_i^{(t)}\}$ reveal which agent is systematically better at forecasting peers (Theorem E.3), allowing a decision-maker to allocate more weight to the truth-holder and achieve lower Bayes risk in predicting Y. Under standard debate, the scores are not observable, and in the correlated noise regime arguments and self-beliefs alone do not suffice to reliably identify the truth-holder. Therefore, there exists at least one decision problem (namely, choosing which agent to follow) for which AceMAD yields a strictly lower expected loss than standard debate. This implies strict Blackwell dominance. □

E.3. Cognitive Potential Energy Separation

We now show that, under the correlated noise assumptions and a strictly proper scoring rule, the truth-holder obtains a strictly higher expected peer-prediction score than the crowd.

We model the realized average of other agents' reports as a random probability vector. Fix an agent i, and let

$$X := \bar{Q}_{-i}^{(t)} \in \Delta(\mathcal{A})$$

denote the empirical distribution over a finite answer set $\mathcal{A}$. An agent reports a forecast $q \in \Delta(\mathcal{A})$, which is scored by the Brier score

$$S(q, X) := 1 - \|q - X\|_2^2 = 1 - \sum_{a \in \mathcal{A}} (q_a - X_a)^2.$$

We first recall a standard decomposition of squared error.

Lemma E.2 (Brier decomposition)**.** *Let X be a random vector in $\mathbb{R}^d$ with mean $\mu := \mathbb{E}[X]$, and let $q \in \mathbb{R}^d$. Then*

$$\mathbb{E}\big[\|q - X\|_2^2\big] = \mathbb{E}\big[\|X - \mu\|_2^2\big] + \|q - \mu\|_2^2.$$

Proof. Write $\|q - X\|_2^2 = \sum_{k=1}^d (q_k - X_k)^2$. For each coordinate,

$$\begin{aligned}
\mathbb{E}[(q_k - X_k)^2] &= \mathbb{E}\big[(X_k - \mu_k + \mu_k - q_k)^2\big] \\
&= \mathbb{E}\big[(X_k - \mu_k)^2\big] + 2(\mu_k - q_k)\,\mathbb{E}[X_k - \mu_k] + (\mu_k - q_k)^2 \\
&= \mathbb{E}\big[(X_k - \mu_k)^2\big] + (\mu_k - q_k)^2,
\end{aligned}$$

since $\mathbb{E}[X_k - \mu_k] = 0$. Summing over all coordinates yields

$$\mathbb{E}\big[\|q - X\|_2^2\big] = \mathbb{E}\big[\|X - \mu\|_2^2\big] + \|q - \mu\|_2^2.$$

□

This shows that, in expectation, the Brier score is uniquely maximized by reporting the true mean μ.

We now apply this to the truth-holder and the crowd.

Theorem E.3 (Identification via proper scoring)**.** *Let $X := \bar{Q}_{-i}^{(t)}$ denote the random average of other agents' reports, with mean $\mu := \mathbb{E}[X]$. Assume the truth-holder E has a correct probabilistic model of X and therefore reports $\hat{Q}_E^{(t)} = \mu$ as its peer-forecast, while any crowd member $c \in C$ reports $\hat{Q}_c^{(t)}$ with $\hat{Q}_c^{(t)} \neq \mu$ with positive probability. Then, for the Brier score,*

$$\mathbb{E}\big[S_E^{(t)}\big] > \mathbb{E}\big[S_c^{(t)}\big].$$

Proof. Recall that $S(q, X) = 1 - \|q - X\|_2^2$. For the truth-holder, Lemma E.2 with $q = \mu$ gives

$$\mathbb{E}\big[\|\hat{Q}_E^{(t)} - X\|_2^2\big] = \mathbb{E}\big[\|X - \mu\|_2^2\big].$$

For a crowd member c,

$$\mathbb{E}\big[\|\hat{Q}_c^{(t)} - X\|_2^2\big] = \mathbb{E}\big[\|X - \mu\|_2^2\big] + \big\|\hat{Q}_c^{(t)} - \mu\big\|_2^2,$$

again by Lemma E.2. Since $\hat{Q}_c^{(t)} \neq \mu$ with positive probability, we have $\mathbb{E}\big[\|\hat{Q}_c^{(t)} - \mu\|_2^2\big] > 0$, and thus

$$\mathbb{E}\big[\|\hat{Q}_c^{(t)} - X\|_2^2\big] > \mathbb{E}\big[\|\hat{Q}_E^{(t)} - X\|_2^2\big].$$

Translating back to scores,

$$\begin{aligned}
\mathbb{E}\big[S_E^{(t)}\big] &= 1 - \mathbb{E}\big[\|\hat{Q}_E^{(t)} - X\|_2^2\big], \\
\mathbb{E}\big[S_c^{(t)}\big] &= 1 - \mathbb{E}\big[\|\hat{Q}_c^{(t)} - X\|_2^2\big],
\end{aligned}$$

so

$$\mathbb{E}\big[S_E^{(t)}\big] - \mathbb{E}\big[S_c^{(t)}\big] = \mathbb{E}\big[\|\hat{Q}_c^{(t)} - X\|_2^2\big] - \mathbb{E}\big[\|\hat{Q}_E^{(t)} - X\|_2^2\big] > 0.$$

□

E.4. Converting Potential Energy into Submartingale Drift

We now show that, under multiplicative weight updates and the score separation of Theorem E.3, the aggregated belief $p_w^{(t)}(Y=1)$ forms a sub-martingale with a positive drift towards the truth-holder's belief.

For clarity, we first analyze a reduced system with two meta-agents:

- the truth-holder E, with belief $p_E^{(t)}$;
- a meta-crowd C, aggregating all crowd agents into a single agent with belief $p_C^{(t)}$.

The general case with $N > 2$ agents follows by grouping all crowd weights into a single meta-agent.

Let $w_E^{(t)}$ and $w_C^{(t)}$ be their weights at round t. Define the truth-holder's weight share

$$\alpha_E^{(t)} := \frac{w_E^{(t)}}{w_E^{(t)} + w_C^{(t)}}, \qquad \alpha_C^{(t)} := 1 - \alpha_E^{(t)}.$$

The aggregate belief is

$$p_w^{(t)} := p_w^{(t)}(Y=1) = \alpha_E^{(t)} p_E^{(t)} + \alpha_C^{(t)} p_C^{(t)}.$$

Weights are updated according to

$$w_i^{(t+1)} = w_i^{(t)} \exp\big(\eta S_i^{(t)}\big), \qquad i \in \{E, C\}.$$

Hence

$$\alpha_E^{(t+1)} = \frac{\alpha_E^{(t)} \exp\big(\eta S_E^{(t)}\big)}{\alpha_E^{(t)} \exp\big(\eta S_E^{(t)}\big) + \alpha_C^{(t)} \exp\big(\eta S_C^{(t)}\big)}.$$

We first observe that the truth-holder's weight share increases exactly when its score exceeds the crowd's.

Lemma E.4 (Monotonicity of weight shares)**.** *For any fixed* $\alpha_E \in (0,1)$,

$$\alpha_E^{(t+1)} \geq \alpha_E^{(t)} \quad \Longleftrightarrow \quad S_E^{(t)} \geq S_C^{(t)}.$$

Proof. Let $\alpha_E := \alpha_E^{(t)}$ and $\alpha_C := 1 - \alpha_E$, and write $S_E := S_E^{(t)}$, $S_C := S_C^{(t)}$. Then

$$\alpha_E^{(t+1)} = \frac{\alpha_E e^{\eta S_E}}{\alpha_E e^{\eta S_E} + \alpha_C e^{\eta S_C}}.$$

The inequality $\alpha_E^{(t+1)} \geq \alpha_E$ is equivalent to

$$\frac{\alpha_E e^{\eta S_E}}{\alpha_E e^{\eta S_E} + \alpha_C e^{\eta S_C}} \geq \alpha_E.$$

Assuming $\alpha_E > 0$, this can be rearranged as

$$e^{\eta S_E} \geq \alpha_E e^{\eta S_E} + \alpha_C e^{\eta S_C} \quad \Longleftrightarrow \quad \alpha_C e^{\eta S_E} \geq \alpha_C e^{\eta S_C} \quad \Longleftrightarrow \quad S_E \geq S_C,$$

since the exponential is strictly increasing. The converse implication follows by the same algebra. □

The change in the aggregated belief can be expressed directly in terms of the change in α_E. In the simplest case where $p_E^{(t)}$ and $p_C^{(t)}$ are time-invariant, we have

$$p_w^{(t+1)} - p_w^{(t)} = \big(\alpha_E^{(t+1)} - \alpha_E^{(t)}\big)(p_E - p_C).$$

Since the truth-holder is more accurate, we have $p_E > p_C$, so the sign of $p_w^{(t+1)} - p_w^{(t)}$ matches that of $\alpha_E^{(t+1)} - \alpha_E^{(t)}$, which by Lemma E.4 matches the sign of $S_E^{(t)} - S_C^{(t)}$.

We now combine this with the expected score gap.

Assumption E.5 (Uniform score gap)**.** There exists $\gamma > 0$ such that for all rounds t,

$$\mathbb{E}\big[S_E^{(t)} - S_C^{(t)} \,\big|\, \mathcal{F}_t\big] \geq \gamma,$$

where $\mathcal{F}_t$ denotes the sigma-algebra generated by the history up to round t.

Assumption E.5 is justified in our correlated noise model by Theorem E.3: the truth-holder is strictly better at forecasting the crowd's average reports, and the gap is uniform as long as the crowd retains a non-trivial weight share.

We can now state the main result.

Theorem E.6 (Breaking the Martingale Curse: positive drift)**.** *Suppose Assumption E.5 holds and $p_E > p_C$. Then, for sufficiently small $\eta > 0$, the aggregated belief process $\{p_w^{(t)}(Y=1)\}_{t\geq 0}$ is a sub-martingale with respect to $\{\mathcal{F}_t\}$:*

$$\mathbb{E}\big[p_w^{(t+1)}(Y=1) \,\big|\, \mathcal{F}_t\big] \geq p_w^{(t)}(Y=1).$$

Moreover, as long as both truth-holder and crowd retain non-zero weight share, the conditional drift is strictly positive:

$$\mathbb{E}\big[p_w^{(t+1)}(Y=1) - p_w^{(t)}(Y=1) \,\big|\, \mathcal{F}_t\big] > 0.$$

Proof. We focus on the two-meta-agent reduction described above. Let us write $\alpha_E := \alpha_E^{(t)}$, $\alpha_C := 1 - \alpha_E$, and define the score difference

$$D^{(t)} := S_E^{(t)} - S_C^{(t)}.$$

From the weight update, we can express $\alpha_E^{(t+1)}$ as

$$\alpha_E^{(t+1)} = \frac{\alpha_E e^{\eta S_E^{(t)}}}{\alpha_E e^{\eta S_E^{(t)}} + \alpha_C e^{\eta S_C^{(t)}}} = \frac{\alpha_E e^{\eta D^{(t)}}}{\alpha_E e^{\eta D^{(t)}} + \alpha_C}.$$

For fixed α_E and α_C, define

$$g(d) := \frac{\alpha_E e^{\eta d}}{\alpha_E e^{\eta d} + \alpha_C},$$

so that $\alpha_E^{(t+1)} = g(D^{(t)})$ and $\alpha_E = g(0)$. A Taylor expansion of g around $d = 0$ yields

$$g(d) = g(0) + g'(0)\, d + O(d^2),$$

with

$$g'(0) = \alpha_E \alpha_C \eta.$$

Since the scores are bounded, $D^{(t)}$ is bounded, and thus $O(d^2)$ can be controlled uniformly. Therefore, there exists a constant $K > 0$ such that

$$\alpha_E^{(t+1)} - \alpha_E \geq \alpha_E \alpha_C \eta D^{(t)} - K\eta^2$$

for all rounds t.

Assuming p_E and p_C are fixed across rounds, the change in the aggregated belief is

$$p_w^{(t+1)} - p_w^{(t)} = \big(\alpha_E^{(t+1)} - \alpha_E^{(t)}\big)(p_E - p_C).$$

Taking conditional expectations and applying Assumption E.5,

$$\begin{aligned}
\mathbb{E}\big[p_w^{(t+1)} - p_w^{(t)} \,\big|\, \mathcal{F}_t\big] &= (p_E - p_C)\, \mathbb{E}\big[\alpha_E^{(t+1)} - \alpha_E^{(t)} \,\big|\, \mathcal{F}_t\big] \\
&\geq (p_E - p_C)\, \big(\alpha_E^{(t)} \alpha_C^{(t)} \eta\, \mathbb{E}[D^{(t)} \mid \mathcal{F}_t] - K\eta^2\big) \\
&\geq (p_E - p_C)\, \big(\alpha_E^{(t)} \alpha_C^{(t)} \eta\gamma - K\eta^2\big),
\end{aligned}$$

where $\gamma > 0$ is the uniform score gap in Assumption E.5.

Choosing $\eta > 0$ small enough, e.g.

$$0 < \eta \leq \frac{\alpha_E^{(t)} \alpha_C^{(t)} \gamma}{2K},$$

ensures that the bracketed term is non-negative, and strictly positive whenever $\alpha_E^{(t)} \alpha_C^{(t)} > 0$. Therefore,

$$\mathbb{E}\big[p_w^{(t+1)}(Y=1) \,\big|\, \mathcal{F}_t\big] \geq p_w^{(t)}(Y=1),$$

with strict inequality as long as both agents retain non-zero weight share. This establishes the sub-martingale property and the strictly positive drift away from the initial state. □

By standard results on multiplicative weights and truth-holder advice, such a persistent score advantage implies that the truth-holder's weight converges to one, and hence the aggregated belief $p_w^{(t)}(Y=1)$ converges to the truth-holder's belief, which is itself close to the ground truth under the correlated noise assumptions.

F. Robustness Across Model Architectures

To verify that AceMAD's performance gains are not artifacts of specific model architectures or training procedures, we conduct additional experiments using `DeepSeek-V3.1`, a recent open-source model with different architectural design and pre-training data distribution compared to `GPT-4o-mini` and `Qwen3`. The results in Table 4 demonstrate that our peer-prediction mechanism successfully extracts asymmetric cognitive potential energy across diverse model families.

Notably, `DeepSeek-V3.1` achieves particularly strong baseline performance on several benchmarks (e.g., 75% on ARC-C with Majority Voting), suggesting that this model family may have less severe correlated noise in its default generation patterns. Nevertheless, AceMAD still provides consistent improvements (average +6.36%), with the most substantial gains observed on LogiQA (+6.87%) and BBH (+5.25%). This validates our theoretical claim that second-order belief elicitation is a model-agnostic mechanism for breaking the Martingale Curse, effective even when baseline error rates are relatively low. The one exception is TruthfulQA,where AceMAD shows a slight degradation. We hypothesize this occurs because `DeepSeek-V3.1`'s strong factual knowledge reduces the presence of a "hallucinating majority" on this benchmark, causing peer predictions to become less informative. This pattern reinforces our framework's scope: AceMAD is specifically designed for the Challenging Interval where correlated misconceptions dominate; in easier regimes where errors are independent, Decentralized MAD or even single-agent inference may suffice.

Table 4. Generalization across models: Performance comparison using the *DeepSeek-V3.1* model ($N = 5$). *AceMAD* consistently outperforms all baseline variants, demonstrating its robustness across different LLM architectures.

Methods	ARC-C	LogiQA	MMLU-Pro	TruthfulQA	MedQA	BBH	Average
Majority Voting	75.00	22.19	31.80	52.47	43.23	52.47	46.19
Decentralized MAD	78.70	27.50	35.78	60.99	56.45	60.99	53.40
Centralized MAD	75.93	46.25	36.54	**84.75**	56.77	62.78	60.50
Sparse MAD	75.00	47.19	35.78	62.78	57.10	84.75	60.43
AceMAD (T=3)	**80.56** (↑1.86)	**54.06** (↑6.87)	**38.50** (↑1.96)	78.03 (↓6.72)	**60.00** (↑2.90)	**90.00** (↑5.25)	**66.86** (↑6.36)

G. Performance Analysis on 8B-Scale Models

In Table 5, we evaluate *AceMAD* using `Llama-3.1-8B-Instruct` to investigate its efficacy on models with relatively smaller parameter scales. The results demonstrate that our peer-prediction mechanism remains effective even within the 8B-parameter regime, consistently outperforming standard debate and voting baselines. However, we observe that the performance gains on `Llama-3.1-8B-Instruct` are less pronounced compared to those observed with larger models like `GPT-4o-mini`.

This discrepancy can be attributed to the inherent trade-off between model capacity and second-order reasoning: while the 8B model possesses sufficient knowledge to identify misconceptions, its ability to maintain a stable and accurate "peer-prediction" ($\hat{q}_i$) is naturally more constrained. The sparse truth signals are harder to elicit when the base agents have higher baseline hallucination rates, yet *AceMAD* still successfully induces a positive drift toward the ground truth. This confirms that the proposed framework provides robust benefits across the parameter spectrum, though the magnitude of improvement is partially modulated by the reasoning ceiling of the underlying model.

Table 5. Performance comparison on the `Llama-3.1-8B-Instruct` model ($N = 5$). This table evaluates the effectiveness of *AceMAD* using an open-source model backbone across all benchmarks, demonstrating that the peer-prediction mechanism remains robust across different model architectures.

Methods	TruthfulQA	MedQA	BBH	Average
Single Agent	27.80	28.06	17.93	24.60
Majority Voting	30.49	27.42	17.59	25.17
Decentralized MAD	33.18	27.74	16.90	25.94
Centralized MAD	**37.53**	27.42	16.55	27.17
Sparse MAD	33.63	26.45	16.55	25.54
AceMAD (T=3)	36.77	**29.68**	**22.76**	**29.74**

H. Results on Full Datasets

While the primary focus of our study is the *Challenging Interval*, we also evaluate *AceMAD* on the full versions of the ARC-C and LogiQA benchmarks to assess its performance in mixed-difficulty environments. As summarized in Table 6, our method continues to outperform Decentralized MAD protocols even when the majority of instances do not contain correlated misconceptions. In these "easier" regimes, the peer-prediction mechanism remains stable and does not penalize correct majorities, as truth-holders and the crowd naturally converge on their forecasts. The consistent gains observed in Table 4 demonstrate that *AceMAD* is not merely a specialized tool for adversarial cases, but a robust reasoning framework that maintains its advantage across the entire difficulty spectrum of real-world datasets.

Table 6. Performance comparison on the full benchmarks ($N = 5$). Unlike the focus on the *Challenging Interval*, this table reports results across the entire test sets of ARC-C and LogiQA. *AceMAD* demonstrates consistent gains even in these mixed-difficulty regimes, with improvements over the baseline shown in parentheses.

Methods	ARC-C	LogiQA	Average
MAD	81.06	49.15	65.11
AceMAD	**85.07** (↑4.01)	**53.41** (↑4.26)	**69.24** (↑4.13)

I. Prompt

Persona Prompt

"Generalist": "You are a helpful assistant. You trust common knowledge and consensus."
"Skeptic": "You are a strict skeptic. You actively look for common misconceptions and logical traps. You suspect the majority might be wrong."

Argumentation Prompt

Question: {question}
Options: {options_str}
Conversation History: {history_context}
Please provide a concise argument for what you believe is the correct answer. Do NOT output JSON here. Just natural language debate.

Commit Prompt

Question: {question} Options: {options_str}
Conversation History: {history_contex}
Task:
1. Self_Prob: Assign probability (0.0-1.0) to options based on your belief.
2. Peer_Prediction: Predict the AVERAGE probability distribution of the OTHER agents in this conversation. (Critically: If others are 'Generalists', they might fall for common misconceptions. Predict their likely errors.)
Output JSON: {{ "self_prob": {{"A": 0.1, ...}}, "peer_prediction": {{"A": 0.3, ...}} }}