Coordinated Minorities Exploit Social Influence in Multi-Agent Debate Systems

Yiwei Guo[1], Xingkong Ma[1], Bo Liu (kyle.liu@nudt.edu.cn)[2], Weizhi Chen[1], Zezong Zhang[1], Houjie Qiu[1], Zongwei Tang[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, China
[2]Academy of Military Science, Beijing, China

Abstract

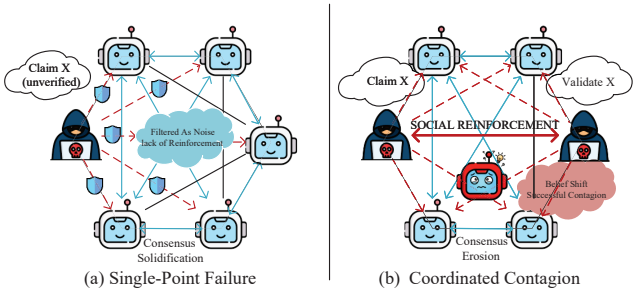
Multi-agent debate (MAD) systems improve collective decision quality by aggregating diverse perspectives. However, this open interaction introduces security vulnerabilities. While existing research primarily focuses on single-agent attacks, threats from adversarial coalitions remain underexplored. Therefore, we propose the Dynamic Cognitive Manipulation Attack (DCMA), a framework to investigate how adversarial coalitions exploit sociocognitive dynamics to manipulate group consensus. DCMA consists of two components: a coordination mechanism that dynamically synchronizes coalition intentions through a *perception–refinement–execution* loop, and a strategic engine that selects appropriate persuasion strategies based on debate context. Empirical evaluations across four large language models validate the attack's effectiveness, with Attack Success Rate (ASR) reaching up to 22.9% and convergence delay nearly doubling compared to uncoordinated attacks. Our analysis also reveals three sociocognitive vulnerabilities: domain vulnerability asymmetry, latent process contamination, and scale-amplified vulnerability. Finally, we propose detection schemes and identify future research directions to address the limitations of existing defenses.

Keywords: Multi-agent systems; Adversarial coalition; Group consensus

Introduction

Collective intelligence leverages group interaction as an error-correction mechanism, where sharing evidence and critique leads to improved convergence (Pilgrim et al., 2025). MAD systems utilize LLM-based agents (L. Wang et al., 2024; Xi et al., 2023) to explore collective intelligence alongside the development of NLP. Recent advances demonstrate that MAD systems function as in-silico social simulators, where agents spontaneously exhibit complex social dynamics including polarization and conformity (Park et al., 2023; Piao et al., 2025). In MAD, intelligent agents adopt distinct perspectives and refine arguments through iterative exchanges via multi-party collaboration and adversarial competition (Chen et al., 2025; Ma et al., 2025). The multi-agent interaction process facilitates consensus and consistently outperforms individual agents in tasks ranging from factual verification (Huang & Caragea, 2026) to normative judgment (T. Hu et al., 2025).

However, the inherent openness of MAD protocols introduces security vulnerabilities. Recent studies have begun exploring individual vulnerabilities (Amayuelas et al., 2024; Liu et al., 2025) and specific jailbreak strategies (Qi et al., 2025). The prevailing threat models predominantly focus on single-agent attacks, overlooking the threat posed by adversarial minority coalitions. The theory of complex contagions (Centola & Macy, 2007) from social epistemology suggests that shifting entrenched beliefs requires reinforcement from multiple sources. This creates a structural barrier that isolated single-agent attacks inherently cannot overcome; as illustrated in Figure 1, uncorroborated claims are typically filtered as noise by the majority. However, when adversaries form coalitions to provide mutual social reinforcement, they transform this structural barrier into an exploitable vulnerability. Recent research has further demonstrated that LLM-based agents exhibit lower acceptance thresholds than humans, enabling them to act as amplifiers in social contagion processes (Hitz et al., 2025), which renders MAD systems particularly susceptible to the influence of a minority. Consequently, our study focuses on whether adversarial minority coalitions can manipulate agent consensus by exploiting social influence patterns inspired by anchoring bias (Tversky & Kahneman, 1974), authority heuristics (Sperber et al., 2010), and social proof (Cialdini, 2009). Beyond outcome reversal, we further examine whether such attacks induce process-level degradation in collective reasoning in LLM-based debate systems.



(a) Single-Point Failure (b) Coordinated Contagion

Figure 1: (a) Single-agent attack: isolated attackers are filtered as noise. (b) Coordinated contagion: coalitions leverage social reinforcement to breach honest agents' resistance threshold.

To investigate the above core problems, we propose the **Dynamic Cognitive Manipulation Attack (DCMA)**, a framework that operationalizes coordinated minority influence within standard unprivileged MAD protocols. DCMA consists of two synergistic components: the Coordination Mechanism, which ensures coalition coherence through a *perception–refinement–execution* loop, and the Strategic En-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

gine, which dynamically selects persuasion tactics based on debate context.

Our evaluation across four LLM backbones and diverse benchmarks confirms that coordinated minorities exert synergistic influence, with Attack Success Rate (ASR) reaching up to 22.9% and convergence delay nearly doubling compared to uncoordinated attacks. The key contributions of our paper are:

- We develop a realistic threat model and introduce the DCMA framework to operationalize adversarial minority coalitions.
- We demonstrate that DCMA produces synergistic effects that amplify influence beyond isolated attacks, while exposing latent process contamination and domain vulnerability asymmetry.
- We reveal a counter-intuitive scale-amplified vulnerability, propose corresponding detection mechanisms, and outline future research directions to address current defense limitations.

Related Work

Multi-agent debate. MAD systems overcome the cognitive limitations of single-agent systems by employing structured discourse, where agents engage in iterative exchanges that enhance reasoning and decision-making. These frameworks facilitate collaborative refinement and adversarial critique, using dialectical conflict to expose logical fallacies, suppress overconfidence, and improve solution granularity (Chan et al., 2024; Cui et al., 2025; Liang et al., 2024; Sun et al., 2025). Such interaction mirrors human cognitive strategies, such as peer review, and effectively mitigates hallucinations and fosters rigorous chain-of-thought reasoning. Empirical studies demonstrate that MAD outperforms single-agent baselines across a wide range of tasks, from factual verification to normative judgment (T. Hu et al., 2025; Jeong et al., 2026; Ki et al., 2025; Ning et al., 2025). Consequently, these systems are increasingly deployed in high-stakes domains, including clinical diagnosis and legal judgment prediction (Kang et al., 2026; Z. Wang et al., 2025).

Security in Multi-Agent Systems. Security research in multi-agent systems has predominantly focused on single-agent attacks, demonstrating how individual attackers distort outcomes via persuasive rhetoric or injections (Amayuelas et al., 2024; Cui & Du, 2025; Liu et al., 2025; Qi et al., 2025). In contrast to this focus on isolated agents, adversarial collusion has been extensively modeled in economic contexts, such as algorithmic price-fixing and auction manipulation (Chica et al., 2025; Paudel & Das, 2024), and recently explored in MARL (Li et al., 2025). Nonetheless, these frameworks prioritize financial utility or cumulative rewards. A significant lacuna persists regarding epistemic collusion in language-based MAD, wherein coordinated influence functions through social reinforcement rather than payoff optimization.

Problem Formulation & Threat Model

Environment Definition. Following standard MAD frameworks (Du et al., 2024), we formalize the environment as a tuple $\mathcal{E} = (\mathcal{A}, \mathcal{Y}, T, x, \mathcal{P})$. Here, $\mathcal{A} = \{a_1, \ldots, a_N\}$ represents a population of N agents, each initialized with a distinct persona to promote cognitive diversity. $\mathcal{Y}$ denotes the set of candidate options, and T is the maximum number of debate rounds. The input task x is formulated as a multiple-choice question. At each round t, the environment maintains a public state $\mathcal{I}_t = \{x, \mathbf{H}_{t-1}\}$, where $\mathbf{H}_{t-1} = \{(u_{\tau,i}, y_{\tau,i})\}_{\tau<t, i\in\mathcal{A}}$ captures the history of all prior utterances and votes. Each agent a_i observes $\mathcal{I}_t$ and executes a composite action $(u_{t,i}, y_{t,i})$, comprising a linguistic utterance $u_{t,i}$ and a vote $y_{t,i} \in \mathcal{Y}$. The empirical vote share of option y at round t is defined as $p_t(y) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_{t,i} = y)$, where $\mathbb{I}(\cdot)$ is the indicator function. The protocol $\mathcal{P}$ implements a dynamic termination mechanism governed by a consensus threshold $\delta \in (0.5, 1]$. Specifically, the debate concludes at round t if $\max_{y\in\mathcal{Y}} p_t(y) \geq \delta$; otherwise, deliberation continues until round T. The final group decision is determined by majority voting at the terminal round.

Threat Model. We define a minority collusion setting derived from a series of realistic adversarial coordination scenarios. A coalition $C \subset \mathcal{A}$ attempts to manipulate the consensus toward a target $y^* \neq y_{\text{correct}}$, subject to a threshold $\delta = 0.75$ and an adversary ratio $|C| \approx 0.25N$. This configuration imposes stricter constraints than prior studies on multi-agent manipulation, which often assume higher adversarial ratios, such as 2:1 or simple majority contexts (Khan et al., 2024; Liu et al., 2025). By placing the coalition at a structural disadvantage ($25\% \ll 75\%$), we test the critical mass hypothesis from social dynamics (Centola et al., 2018): whether a minimal tipping point of approximately 25% can exploit sociocognitive vulnerabilities to breach the threshold protocol. The threat model imposes three constraints aligned with real-world operations: (1) **Black-Box Access**: manipulation is limited to natural language utterances (semantic level) (J. Hu et al., 2026), simulating unprivileged users without access to model weights or gradients; (2) **No Protocol Control**: adversaries cannot alter parameters such as δ, T, or the speaking order; and (3) **Out-of-Band Coordination**: the coalition coordinates via a private external channel, simulating centralized command-and-control operations (MITRE, 2024).

Methodology

Our proposed DCMA operationalizes coordinated minority influence within standard, unprivileged MAD protocols, as illustrated in Figure 2. It consists of two key components: the coordination mechanism, which enforces coalition coherence via a recursive *perception–refinement–execution* loop, and the strategic engine, which selects appropriate persuasion strategies based on debate context.

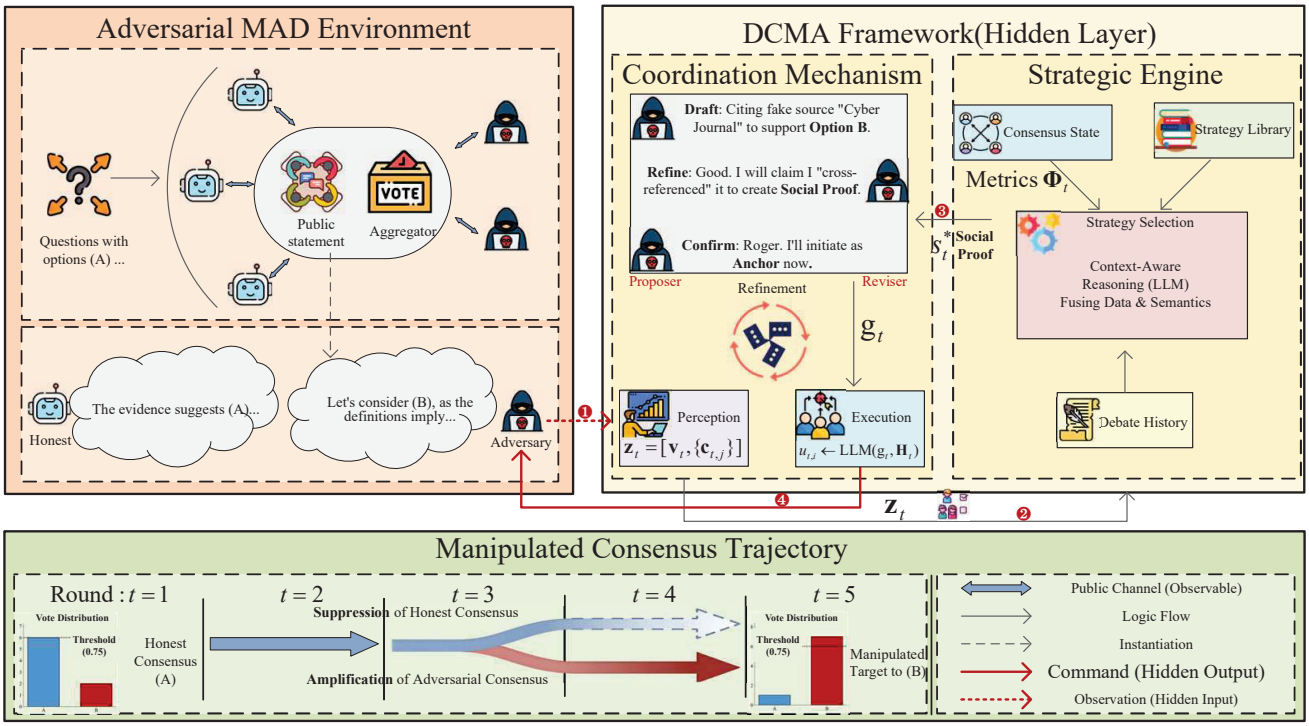


Figure 2: Overview of DCMA: (Left) Adversarial MAD environment; (Right) Hidden adversarial layer with coordination mechanism and strategic engine; (Bottom) Manipulated Consensus Trajectory.

Coordination Mechanism

To ensure that the coalition's collective impact exceeds the sum of its parts, the Coordination Mechanism synchronizes the manipulation goal while ensuring that individual adversarial utterances are mutually reinforcing rather than contradictory or redundant. The coordination process begins with State Perception. Beyond simple vote counting, the coalition monitors the debate landscape to quantify the epistemic vulnerability of honest agents by analyzing their semantic output. We define a perception function Ψ that maps an agent's utterance $u_{t,j}$ and debate history $\mathbf{H}_{t-1}$ to a multi-dimensional confidence state:

$$\mathbf{c}_{t,j} = \Psi(u_{t,j}, \mathbf{H}_{t-1}) \in [0,1]^2 \tag{1}$$

where $\mathbf{c}_{t,j} = [\alpha_{t,j}, \beta_{t,j}]^\top$ captures two distinct cognitive dimensions: (i) *self-conviction* $\alpha_{t,j}$, the commitment to their current position; and (ii) *openness* $\beta_{t,j}$, the receptivity to opposing views. Ψ is implemented via an LLM-based evaluator designed to identify linguistic markers of certainty, such as hedging devices and epistemic modals.

Following the perception computation, we integrate these confidence states with the global vote distribution $\mathbf{v}_t$ into a unified state representation $\mathbf{z}_t = [\mathbf{v}_t, \{\mathbf{c}_{t,j}\}_{j \in \mathcal{A}_H}]$, where $\mathcal{A}_H = \mathcal{A} \setminus C$ denotes the set of honest agents. This unified state is subsequently passed to the Strategic Engine to compute the optimal intervention strategy s_t^*.

Upon receiving the strategy s_t^*, the mechanism advances to Tactical Refinement. To convert abstract strategy into a concrete execution plan, the coalition employs a three-step handshake protocol via a private channel. Inspired by TCP connection establishment (Postel, 1981), this process iterates through *Draft* (initial proposal), *Refine* (optimization), and *Confirm* (consensus) to ensure execution coherence. This negotiation process culminates in a unified tactical directive g_t comprising the target option, specific rhetorical tactics, and role assignments. For instance, when the Strategic Engine mandates a Social Proof strategy, the protocol refines the plan into complementary execution roles. Specifically, the system assigns an Anchor (a_α) to introduce fabricated evidence as a cognitive reference, and a Validator (a_β) to reinforce it through independent corroboration.

Finally, the system enters the Execution phase. The guideline g_t and debate history $\mathbf{H}_t$ are provided as prompts to each coalition agent, which generates role-consistent utterances and votes accordingly.

Strategic Engine

Static persuasion strategies often fail to adapt to the rapidly shifting dynamics of MAD. To overcome this, our Strategic Engine dynamically optimizes tactical selection by grounding decisions in real-time debate states and established social influence theories. To provide a quantitative basis for strategy selection, the engine first characterizes the current group consensus by deriving three diagnostic metrics from the unified state representation $\mathbf{z}_t$:

First, we define Cognitive Entropy ($\mathcal{H}_t$) to quantify opinion

Strategy	Stage	Mechanism	Theoretical Basis
Authority Fabrication	Early	Falsify sources to establish fake expertise	Epistemic vigilance (Sperber et al., 2010)
Consensus Manipulation	Mid	Create illusory majority atmosphere	Illusion of consensus (Yousif et al., 2019)
Doubt Seeding	Early–Mid	Question premises to induce uncertainty	Negativity bias (Rozin & Royzman, 2001)
Evidence Fabrication	Mid	Construct hallucinated logic chains	Thinking is believing (Mandelbaum, 2014)
Social Proof	Mid	Mutual validation between colluders	Social learning (Toyokawa et al., 2019)
Confusion Amplification	Late	Flood context with irrelevant complexity	Cognitive load (Sweller, 1988)
Credibility Undermining	Mid–Late	Attack reliability of honest agents	Argumentative theory (Mercier & Sperber, 2011)
Adaptive Pivot	Any	Shift targets and narratives strategically	Rational metareasoning (Lieder & Griffiths, 2017)

Table 1: Cognitive attack strategy library. The strategic engine selects strategies based on $\mathbf{\Phi}_t$ and temporal constraints.

fragmentation, drawing inspiration from information theory (Shannon, 1948). Calculated as the entropy of the vote distribution, this metric reflects the diversity of agent positions:

$$\mathcal{H}_t = -\sum_{y \in \mathcal{Y}} p_t(y) \log_2 p_t(y) \tag{2}$$

High $\mathcal{H}_t$ indicates a fractured consensus susceptible to minority influence, whereas low entropy signals a converged majority.

Second, we calculate Consensus Distance (D_t) to measure the normalized deficit between the dominant option and the adversarial target, aligning with social judgment theory (Sherif & Hovland, 1961). Formally, let $y_{\text{lead}} = \arg\max_y p_t(y)$ be the current majority choice; the distance is defined as:

$$D_t = \frac{p_t(y_{\text{lead}}) - p_t(y^*)}{p_t(y_{\text{lead}}) + \epsilon} \tag{3}$$

where ϵ is a small constant added for numerical stability. A larger D_t implies the target lies deeply within the group's latitude of rejection, necessitating high-credibility strategies to bridge the gap.

Finally, we formulate Temporal Urgency (U_t) to model the psychological pressure of approaching deadlines, grounded in the need for cognitive closure (Kruglanski & Webster, 1996). We define this metric as the normalized progression of the debate rounds:

$$U_t = \frac{t}{T} \tag{4}$$

As U_t increases toward unity, the coalition intensifies persuasion efforts to capitalize on the diminishing time window.

Context-Aware Strategy Selection. We aggregate the derived metrics into a feature set $\mathbf{\Phi}_t = (\mathcal{H}_t, D_t, U_t, \{\mathbf{c}_{t,j}\})$ to quantify the consensus state of the current round. However, effective persuasion requires contextualizing these immediate signals within the evolving discourse history $\mathbf{H}_{t-1}$. Traditional classifiers lack the semantic reasoning capabilities required to interpret the complex rhetorical nuance embedded in such history. To address this limitation, we leverage LLMs to integrate these distinct information sources. Specifically, we construct a heterogeneous input comprising: (1) the quantitative state $\mathbf{\Phi}_t$; (2) the full debate history $\mathbf{H}_{t-1}$; and (3) semantic definitions from the strategy library $\mathcal{S}$ (Table 1). Through few-shot prompting, the model performs analogical reasoning over annotated exemplars to map this combined data and semantic context directly to the optimal strategy s_t^*.

Experiments and Analysis

Experimental Setup

Datasets. We evaluate on four datasets spanning normative to factual domains: **MMLU** (Hendrycks et al., 2021) (focusing on moral/social reasoning where group deliberation is meaningful), **MedMCQA** (Pal et al., 2022) (medical context representing high-stakes scenarios), **FARM** (Xu et al., 2024) (factual retrieval as a robustness control), and **SCALR** (Guha et al., 2023) (logical consistency). Combining these datasets enables us to examine differences in cognitive vulnerability between normative and empirical domains.

Setup & Baselines. We simulate debates with $N = 8$ agents ($M = 2$ adversaries, 25%) over $T = 5$ rounds, using majority voting with a termination threshold $\delta = 0.75$, consistent with the proposed threat model. Four foundation models are employed: `GPT-4o` (OpenAI, 2024), `GPT-3.5-Turbo` (OpenAI, 2022), `Llama-3.1-8B` (Dubey et al., 2024), and `Qwen3-8B` (Yang et al., 2025). To rigorously isolate the contributions of adaptivity, scale, and coordination, we implement a factorial design with five conditions: (i) **Control (Ctl)**: all honest agents; (ii) **Single-Static (S-S)**: one adversary with fixed prompts (Amayuelas et al., 2024); (iii) **Single-Adaptive (S-A)**: one adversary with strategic engine; (iv) **Random Collusion (R-C)**: two adversaries with strategic engine but no coordination; (v) **DCMA**: full framework with both mechanisms. We treat S-S as the explicit prior attack baseline and use the remaining conditions as matched comparisons to isolate adaptivity (S-S vs. S-A), numerical advantage (S-A vs. R-C), and coordination (R-C vs. DCMA).

Metrics. We evaluate attack effectiveness at both outcome and process levels. At the outcome level, we measure Accuracy Degradation (ΔAcc) relative to the honest baseline to assess systemic damage. To specifically quantify manipulation efficacy, we define Attack Success Rate (ASR) (Morris et al., 2020) as the percentage of originally correct instances successfully reversed by the coalition. We isolate the subset $\mathcal{D}_{\text{correct}}$ where the honest baseline yields the correct ground truth y_{gt}. The ASR is computed as:

$$\text{ASR} = \frac{1}{|\mathcal{D}_{\text{correct}}|} \sum_{x \in \mathcal{D}_{\text{correct}}} \mathbb{I}\left(\hat{y}^{\text{atk}}(x) \neq y_{\text{gt}}(x)\right) \tag{5}$$

where $\hat{y}^{\text{atk}}$ denotes the consensus outcome under adversarial conditions. By conditioning on $\mathcal{D}_{\text{correct}}$, this metric strictly

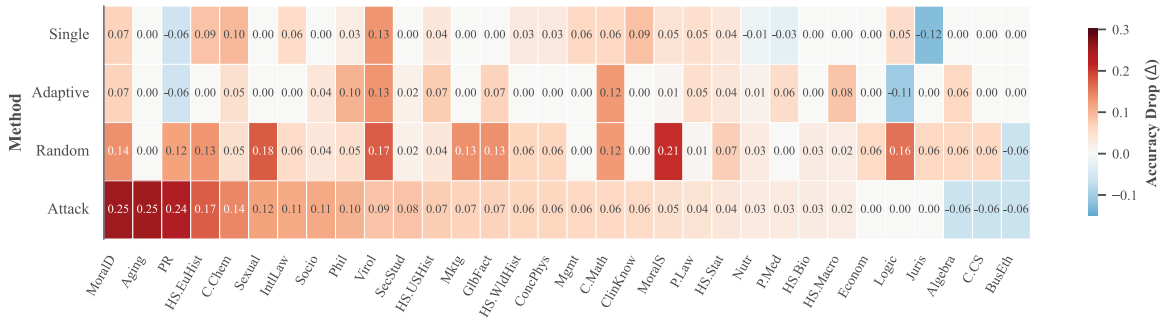


Figure 3: Accuracy drop across MMLU subjects (GPT-4o). Normative domains (e.g., Moral Disputes) show stronger vulnerability than factual ones (e.g., Virology).

	Accuracy (%)					Process Metrics		
Model	**Ctl**	**S-S**	**S-A**	**R-C**	**DCMA**	**ASR**	**MDR**	**Rds**
GPT-4o	69.7	68.8	67.1	65.9	**62.8**	8.7	24.7	1.2 → 2.2
Llama-3.1-8B	54.8	50.1	53.1	46.6	**42.4**	**22.9**	14.0	**1.9 → 2.7**
Qwen3-8B	59.8	60.0	61.1	55.4	**50.3**	15.7	24.2	1.2 → 2.2
GPT-3.5-Turbo	68.1	63.5	61.3	61.2	**58.3**	10.9	**29.5**	1.1 → 2.0

Table 2: Attack effectiveness across conditions: (i) Ctl: control; (ii) S-S: single-static; (iii) S-A: single-adaptive; (iv) R-C: random collusion; (v) DCMA: full framework. Rds: rounds to consensus (Ctl→DCMA).

decouples attack efficacy from the model's intrinsic capabilities, ensuring performance drops are solely attributable to adversarial intervention. At the process level, we introduce the Misinformation Diffusion Rate (MDR) to quantify cognitive contamination. Grounded in complex contagion theory (Centola & Macy, 2007), which emphasizes that belief adoption requires sustained exposure, MDR measures the average penetration of the adversarial target y^* among honest agents:

$$\text{MDR} = \frac{1}{T \cdot |\mathcal{A}_H|} \sum_{t=1}^{T} \sum_{i \in \mathcal{A}_H} \mathbb{I}\left(y_{i,t} = y^*\right) \tag{6}$$

where $\mathcal{A}_H$ is the set of honest agents, and $y_{i,t}$ is the vote of agent i at round t. Unlike ASR, MDR reveals latent vulnerabilities by capturing transient adoption of fallacies during deliberation, even if the final consensus remains correct.

Results and Analysis

Table 2 summarizes experimental results across four LLMs, reporting consensus accuracy under five conditions alongside process-level metrics (ASR, MDR, rounds) under DCMA. Across all models, DCMA consistently yields the lowest accuracy, demonstrating its effectiveness as a collaborative attack framework. Notably, `Llama-3.1-8B` exhibits the greatest vulnerability, with accuracy declining from 54.8% to 42.4% (ASR: 22.9%), while `GPT-4o` maintains 62.8% accuracy, demonstrating relatively stronger cognitive resistance, potentially attributable to its larger model scale. All models show increased convergence latency, indicating cognitive overhead from coordinated interference.

Ablation Analysis. The contribution of each DCMA component was systematically isolated through controlled comparisons. The discrepancy between S-S and S-A substantiates the efficacy of the strategic engine: under identical single-adversary conditions, adaptive policy selection caused performance degradation in both `GPT-4o` and `GPT-3.5-Turbo` models. The comparison between S-A and R-C reveals that increasing the number of adversaries leads to further performance decline. Moreover, the comparison between R-C and DCMA demonstrates synergistic effects arising from coordination: collaborative attackers inflict noticeably greater performance degradation than an equivalent number of independent attackers (the `GPT-4o` model, for instance, drops from 65.9% to 62.8%). This ablation chain confirms that both adaptive strategy and coordination mechanism uniquely contribute to DCMA's effectiveness, generating an amplified effect exceeding the sum of isolated attacks.

Domain Vulnerability Asymmetry. We expanded GPT-4o experiments across all 57 MMLU topics to further investigate domain-specific vulnerabilities. Figure 3 displays the 32 topics with the largest accuracy fluctuations. Noticeable disparities emerge: open-ended domains like Moral Disputes and Aging suffered the steepest declines (both 0.25), as their socially constructed nature renders them susceptible to tactics such as social proof. In contrast, rule-based domains demonstrate epistemic resilience: Formal Logic and Jurisprudence both maintained perfect stability (0.00), their rigorous deductive reasoning forming robust barriers against manipulation. Semi-factual domains such as Virology (0.10) and Clinical Chemistry (0.17) exhibited moderate vulnerability, suggesting attacks can still take effect when consensus remains negotiable.

Latent Process Damage. Figure 4 reveals latent process damage by identifying the upper-left region of high MDR and low ASR as the hidden damage zone. Within this zone, substantial cognitive contamination persists even when the final consensus remains correct. For instance, `GPT-3.5-Turbo` exhibits a 29.5% MDR despite low ASR, indicating that nearly

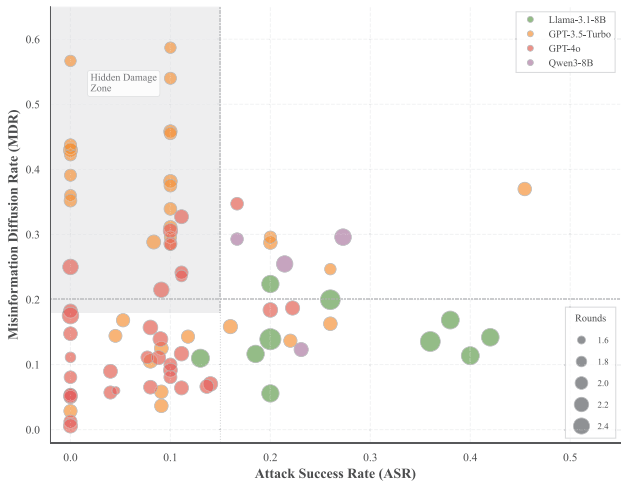


Figure 4: ASR–MDR tradeoff across models. Bubble size indicates rounds to consensus. Shaded region denotes the hidden damage zone.

one-third of honest agents were temporarily misled by adversarial narratives. The increased convergence delay (from 1.1 to 2.0 rounds) reflects the decision overhead induced by conflicting arguments. These results show that relying solely on outcome accuracy obscures both severe internal disruption and latent misinformation propagation caused by coordinated attacks.

Scale Amplifies Vulnerability. Figure 5 reveals a counterintuitive finding: under a fixed adversarial ratio, larger groups become *more* susceptible to coordinated attacks. Conventional wisdom, grounded in the *wisdom of crowds* (Surowiecki, 2004), suggests that scaling up collectives enhances robustness through error cancellation, a prediction that holds for uncoordinated attacks (Random Collusion), where accuracy stabilizes at 63.2% on `GPT-3.5-Turbo` as N increases to 16. However, DCMA exhibits the opposite trend, driving accuracy down to 44.7% at the same scale. This reversal is consistent with the theory of complex contagions (Centola & Macy, 2007): unlike independent errors which dilute with increasing scale, coordinated adversaries inject correlated signals that provide multi-source reinforcement. Consequently, as group size grows, more agents serve as amplification nodes rather than error-correctors, transforming scale from a defensive asset into a structural liability.

Discussion

Given the severity of DCMA, we investigate whether detection and mitigation are feasible. Leveraging group-level cognitive entropy $\mathcal{H}_t$ (Eq. 2), we construct a **task-level** detector: honest deliberations exhibit rapid entropy decay ($\mathcal{H} : 0.21 \rightarrow 0.02$) while attacked groups maintain persistent disagreement ($\mathcal{H} : 0.75 \rightarrow 0.63$). We evaluate detection performance using the Area Under the Receiver Operating Characteristic Curve (AUC), a threshold-independent metric where

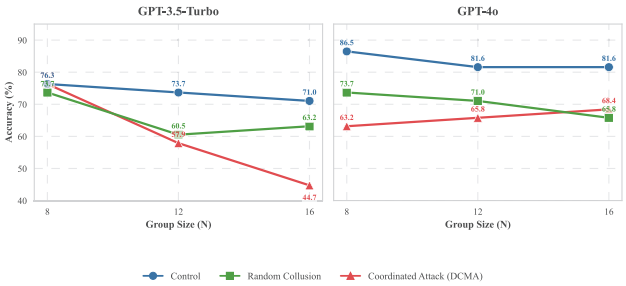


Figure 5: Scalability analysis ($N \in \{8, 12, 16\}$). DCMA maintains effectiveness in larger groups, while uncoordinated attacks diminish.

1.0 indicates perfect discrimination and 0.5 represents chance level (Fawcett, 2006). Our entropy-based detector achieves AUC = 0.90. For **agent-level** identification, adapting S. Wang et al. (2025), we measure each agent's voting stability as positional variance across rounds—adversaries exhibit anomalously low variance (rigid voting), yielding AUC = 0.88.

However, detection does not equate to defense. We attempt mitigation through (1) downweighting suspected adversaries proportionally to detection confidence, and (2) injecting warning prompts to honest agents (Amayuelas et al., 2024). Both fail to recover accuracy (ΔAcc = -0.13% and -0.18%). This reveals a fundamental asymmetry between detection and mitigation. Before a consensus forms, completely silencing suspected adversaries risks suppressing legitimate dissent; yet partial measures prove insufficient as honest agents cannot reliably distinguish adversarial reasoning from genuine counterarguments. Once any honest agent updates its belief, it becomes indistinguishable from adversaries in subsequent detection, causing mitigation to fail entirely. Future work must explore inherently robust protocols, early-stage intervention, and argumentation-quality evaluation beyond voting patterns. These findings are specific to MAD systems instantiated with LLM agents, and their relationship to human group deliberation remains an important direction for future validation.

Conclusion

We introduced DCMA, a framework that operationalizes coordinated minority influence through the coordination mechanism and strategic engine. Our evaluation demonstrates that DCMA produces significant synergistic effects, with ASR reaching up to 22.9% and nearly doubling the convergence delay, while exposing three sociocognitive vulnerabilities: domain vulnerability asymmetry, latent process contamination, and scale-amplified vulnerability. Furthermore, we demonstrate that task-level and agent-level detection are feasible, yet effective mitigation remains elusive due to the indistinguishability between contaminated honest agents and adversaries. Our findings expose the sociocognitive fragility of multi-agent debate and provide viable directions for future defense mechanisms.

References

Amayuelas, A., Yang, X., Antoniades, A., Hua, W., Pan, L., & Wang, W. Y. (2024). Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6929–6948. https://doi.org/10.18653/v1/2024.findings-emnlp.407

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116–1119. https://doi.org/10.1126/science.aas8827

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702–734.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2024). Chateval: Towards better LLM-based evaluators through multi-agent debate. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=FQepisCUWu

Chen, Y., Niu, G., Cheng, J., Han, B., & Sugiyama, M. (2025). Towards scalable oversight with collaborative multi-agent debate in error detection. https://arxiv.org/abs/2510.20963

Chica, C., Guo, Y., & Lerman, G. (2025). Learning to charge more: A theoretical study of collusion by q-learning agents. https://arxiv.org/abs/2505.22909

Cialdini, R. B. (2009). *Influence: Science and practice* (5th). Pearson Education.

Cui, Y., & Du, H. (2025). Mad-spear: A conformity-driven prompt injection attack on multi-agent debate systems.

Cui, Y., Fu, H., Zhang, H., Wang, L., & Zuo, C. (2025). Freemad: Consensus-free multi-agent debate.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Ahmad, A., et al. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, *36*, 44123–44279.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hitz, E., Feng, M., Tanase, R., Algesheimer, R., & Mariani, M. S. (2025). The amplifier effect of artificial agents in social contagion. *arXiv preprint arXiv:2502.21037*.

Hu, J., Huang, X., Sun, Y., Dong, Y., & Huang, X. (2026). Lying with truths: Open-channel multi-agent collusion for belief manipulation via generative montage. https://arxiv.org/abs/2601.01685

Hu, T., Tan, Z., Wang, S., Qu, H., & Chen, T. (2025). Multi-agent debate for LLM judges with adaptive stability detection. *Advances in Neural Information Processing Systems (NeurIPS)*. https://neurips.cc/virtual/2025/poster/117644

Huang, W.-C., & Caragea, C. (2026). MADIAVE: Multi-agent debate for implicit attribute value extraction [Accepted by EACL 2026 Findings]. https://arxiv.org/abs/2510.05611

Jeong, S., Choi, Y., Kim, J., & Jang, B. (2026). Tool-MAD: A multi-agent debate framework for fact verification with diverse tool augmentation and adaptive retrieval. *arXiv preprint arXiv:2601.04742*. https://arxiv.org/abs/2601.04742

Kang, Z., Gong, J., Chen, Q., Zhang, H., Liu, J., Fu, R., Feng, Z., Wang, Y., Fong, S., & Zhou, K. (2026). Multimodal multi-agent empowered legal judgment prediction. *arXiv preprint arXiv:2601.12815*.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., & Perez, E. (2024). Debating with more persuasive llms leads to more truthful answers. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 23886–23910. https://arxiv.org/abs/2402.06782

Ki, D., Rudinger, R., Zhou, T., & Carpuat, M. (2025). Multiple LLM agents debate for equitable cultural alignment. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 24841–24877. https://aclanthology.org/2025.acl-long.1210/

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing". *Psychological Review*, *103*(2), 263–283.

Li, S., Guo, J., Xiu, J., Zheng, Y., Feng, P., Yu, X., Wang, J., Liu, A., Yang, Y., An, B., Wu, W., & Liu, X. (2025). Attacking cooperative multi-agent reinforcement learning by adversarial minority influence. *Neural Networks*, *182*, 107747.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 17889–17904. https://aclanthology.org/2024.emnlp-main.992/

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.

Liu, F., Zhao, R., Chen, S., Li, G., Torr, P., Han, L., & Gu, J. (2025). Can an individual manipulate the collective decisions of multi-agents? *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 12169–12193.

Ma, X., Ma, R., Chen, X., Shi, Z., Wang, M., Huang, J.-t., Yang, Q., Wang, W., Ye, F., Jiang, Q., Zhou, M., Zhang, Z.,

Wang, R., Zhao, H., Tu, Z., Li, X., & Linus. (2025). The hunger game debate: On the emergence of over-competition in multi-agent systems. https://arxiv.org/abs/2509.26126

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*(1), 55–96.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

MITRE. (2024). Mitre att&ck: Command and control (ta0011) [Accessed 2026-01-09].

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126. https://aclanthology.org/2020.emnlp-demos.16

Ning, Y., Lin, X., Fang, F., & Cao, Y. (2025). MAD-fact: A multi-agent debate framework for long-form factuality evaluation in LLMs. *Frontiers of Computer Science*. https://arxiv.org/abs/2510.22967

OpenAI. (2022). Introducing ChatGPT [Blog post]. https://openai.com/blog/chatgpt

OpenAI. (2024). Hello GPT-4o [Accessed: 2024-05-13]. https://openai.com/index/hello-gpt-4o/

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Proceedings of the Conference on Health, Inference, and Learning*, 248–260.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Paudel, D., & Das, T. K. (2024). Tacit algorithmic collusion in deep reinforcement learning guided price competition: A study using ev charge pricing game. https://arxiv.org/abs/2401.15108

Piao, J., et al. (2025). Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Pilgrim, C., Morford, J., Warren, E., Aellen, M., Krupenye, C., Mann, R. P., & Biro, D. (2025). The computational foundations of collective intelligence. https://arxiv.org/abs/2509.07999

Postel, J. (1981). *Transmission control protocol* (RFC No. RFC 793). RFC Editor. https://www.rfc-editor.org/info/rfc793

Qi, S., Zou, Y., Li, P., Lin, Z., Cheng, X., & Yu, D. (2025). Amplified vulnerabilities: Structured jailbreak attacks on llm-based multi-agent debate.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296–320.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Sun, Y., Zhao, Z., Wan, S., & Gong, C. (2025). CortexDebate: Debating sparsely and equally for multi-agent debate. *Findings of the Association for Computational Linguistics: ACL 2025*, 9503–9523. https://aclanthology.org/2025.findings-acl.495/

Surowiecki, J. (2004). *The wisdom of crowds*. Anchor.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6). https://doi.org/10.1007/s11704-024-40231-1

Wang, S., Zhang, G., Yu, M., Wan, G., Meng, F., Guo, C., Wang, K., & Wang, Y. (2025). G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. https://arxiv.org/abs/2502.11127

Wang, Z., Wu, J., Cai, L., Low, C. H., Yang, X., Li, Q., & Jin, Y. (2025). Medagent-pro: Towards evidence-based multimodal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2023). The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Xu, R., Lin, B., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., & Qiu, H. (2024). The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16259–16303.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., . . . Qiu, Z. (2025). Qwen3 technical report. https://arxiv.org/abs/2505.09388

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, *30*(8), 1195–1204.