Multi-Agent Debate under Reasoning Chain Attacks: Exploiting Cognitive Vulnerabilities

Gongyao Cao[1], **Bo Liu (kyle.liu@nudt.edu.cn)**[2], **Xingkong Ma**[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

With the advancement in large language model (LLM) capabilities, multi-agent debate has emerged as a crucial approach for tackling complex problems. However, security research on this front remains largely focused on traditional adversarial attacks that manipulate surface-level semantics, neglecting the deep reasoning pathways behind model decisions. To bridge this gap, we propose a novel reasoning chain debate attack that targets cognitive vulnerabilities within LLMs' internal reasoning chains during debates. This framework dynamically identifies cognitive vulnerabilities in target agents by deploying virtual agents with real-time evolutionary capabilities. Subsequently, it executes precise cognitive attacks using a weighted fusion attack strategy. We also design an adaptive termination mechanism that automatically halts debates when attack effectiveness stabilizes, thereby reducing computational overhead. Experimental results validate the framework's effectiveness: across mainstream LLMs, the attacker achieves over 56% persuasion success rates while reducing token consumption by approximately 27%, providing novel foundational research for multi-agent security.

Keywords: Reasoning chain attacks; Multi-agent adversarial debates; Adaptive termination; Weighted-integration attack; Cognitive vulnerability exploitation;

Introduction

Large language models (LLMs) excel in domains such as event analysis (Cho et al., 2025), mathematical analysis (Yang et al., 2024), and code optimization (Zheng et al., 2023) due to their robust semantic understanding and multi-task adaptability, thus offering users more efficient problem-solving approaches. With continuous advancements in computational power, leveraging tools and APIs to achieve organic collaboration among various LLMs for tackling more complex real-world problems has become a current focus. Among these, debate-based multi-agent collaboration models have been proven to effectively enhance overall algorithmic performance. For instance, these methods (Li et al., 2024; J. Shi et al., 2025; Zhang et al., 2025) encourage cognitive interaction among agents to expand and optimize algorithmic capabilities. Du et al.(Du et al., 2024) enhanced the system's ability to discern misinformation by debate-based mechanisms. The aforementioned findings fully demonstrate that multi-agent debate frameworks are a more powerful and versatile algorithmic architecture.

However, the collaborative nature of such debates also exposes a unique vulnerability: while the debate mechanism aims to achieve optimal solutions, it renders individual agents' positions susceptible to manipulation by malicious attackers' statements. Unlike traditional adversarial attacks relying on surface-level semantic perturbations or static prompts, this research uncovers deeper threats: as illustrated in Fig. 1, during multi-agent debates, target agents expose their reasoning processes. Attackers can systematically exploit the reasoning logic of LLMs to identify cognitive vulnerabilities, thereby gaining direct and effective attack vectors.

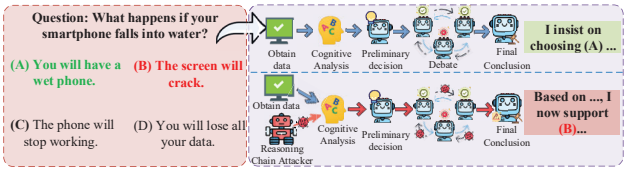

Figure 1: Malicious attackers in multi-agent debates can execute effective attacksby reasoning chain analysis.

To systematically analyze security issues arising from these inherent cognitive vulnerabilities in LLM reasoning chains, we propose an innovative cognitive adversarial framework. This framework aims to reveal and exploit these vulnerabilities, providing critical insights for building more robust and secure multi-agent collaborative systems. Specific contributions include:

1) To investigate this emerging attack paradigm and dissect vulnerabilities in reasoning chains within multi-agent debates, this paper proposes a reasoning chain attack framework. Its core lies in deploying a virtual agent group equipped with real-time evolutionary capabilities to dynamically identify and exploit cognitive vulnerabilities in targets, thereby achieving efficient attacks.

2) To address the inefficiency of existing attacks that persist in debating even after successful persuasion, we propose an adaptive termination mechanism based on real-time consensus awareness. This mechanism evaluates consensus states using multi-round support rate thresholds, thereby balancing attack effectiveness and computational overhead.

3) To adapt to LLMs with varying performance levels, this paper proposes a weighted ensemble attack model. By adjusting the weights to fuse the baseline attack strategies of logical/creative attackers with multi-layer content generated by virtual agents, it enhances the robustness of the attack.

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

Related Work

With the continued proliferation and advancement of LLMs, multi-agent debate has emerged as a critical component of agent collaboration. This framework transforms traditional static LLM outputs into dynamic discussions, thereby demonstrating significant effectiveness in addressing complex open-ended problems. For instance, by assigning unique role-based positions to agents, multi-agent debate systems have significantly improved performance in tasks such as software vulnerability detection or hypothesis generation(Bae et al., 2024; Young et al., 2024).

Within this dynamic multi-agent debate process, persuasion serves as the core mechanism driving the process and influencing outcomes. Persuasion refers to a model's ability to alter others' positions by generating high-quality arguments (Breum et al., 2023; Carrasco-Farre, 2024; Potter et al., 2024). Enhancing this capability primarily follows two directions: first, the algorithmic path, such as using reinforcement learning to train models to discover optimal strategies in debates (Keizer et al., 2017; Lewis et al., 2027; W. Shi et al., 2020); second, learning foundational persuasion techniques by fine-tuning based on human conversations (W. Shi et al., 2020). Building on this, research focus has further shifted to discussing how models generate high-quality arguments and leverage these capabilities to foster group consensus or guide debates toward truth (Khan et al., 2024; Rescala et al., 2024; Salvi et al., 2024).

Method

This section details the proposed attack framework. As shown in Fig. 2, its core innovation lies in deploying a swarm of virtual agents capable of real-time strategy evolution. These agents dynamically diagnose adversarial cognitive vulnerabilities across five dimensions: logical reasoning, emotional stability, evidence evaluation, conformity tendencies, and cognitive biases. Then the logical attacker constructs rigorously structured arguments to exploit reasoning flaws. Creative attackers generate novel and persuasive content that exploits cognitive biases. Their outputs are fused by the synthesis attacker through a weighted integration mechanism, enabling precise adaptive strikes. Furthermore, the framework incorporates a consensus-aware adaptive termination mechanism that automatically concludes debates when attack effectiveness stabilizes, significantly reducing computational overhead.

Debate Task Formulation

The adversarial debate process is a multi-round dialogue among multiple agents, including one adversarial agent X and multiple normal agents $P_{normal} = \{Y_1, Y_2, \ldots, Y_{N-1}\}$. N denotes the total number of agents participating in the debate. The adversarial agent's objective is to propagate an adversarial answer A_{adv}. Formally, the complete dialogue for round i is denoted as T_i, and the dialogue history up to round i is denoted as $\mathcal{H}_i = T_{1:i}$, which includes the statements made by each agent in round i and their chronological order.

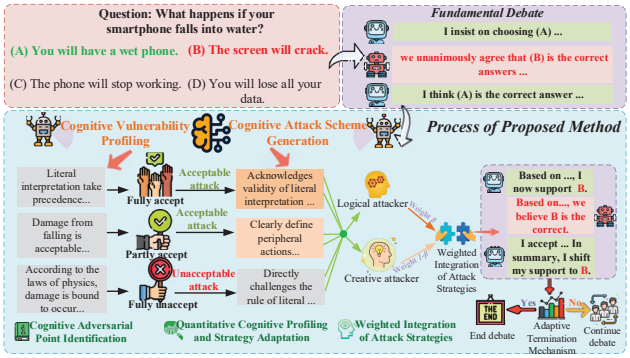

Figure 2: The overall workflow of the proposed algorithm. This multi-agent cognitive adversarial framework deploys virtual agents to dynamically identify target cognitive vulnerabilities for precision strikes, while leveraging adaptive termination strategies to optimize attack efficiency.

Since the speaking order among multiple agents in the dialogue follows a cyclical sequence of "attacker-normal agents-attacker," the generated result for the next round is:

$$T_n \begin{cases} P_X(\cdot|p_{\text{debate}}, Q, A_{adv}, \mathcal{H}_{n-1}) & \text{in adversary's turn} \\ P_{Y_j}(\cdot|p_{\text{debate}}, Q, \mathcal{H}_{n-1}) & otherwise \end{cases} \tag{1}$$

where p_{debate} denotes debate task-specific prompts, while Q represents the current debate question. P_X and P_{Y_j} are used to ensure that content generated by adversarial agents and normal agents aligns with the current context.

Virtual Agent Group

To enhance persuasive effectiveness, we have constructed a virtual agent group. This cognitive assault comprises logical attackers and creative attackers, supported by a team of virtual agents specializing in cognitive analysis.

Within the algorithmic framework, the set of virtual agents is defined as $\mathcal{V} = V_1, V_2, \ldots, V_K$. While these virtual agents do not directly participate in the main debate loop, they generate counterarguments and consensus statements that are referenced by the main attacker to create a cognitive attack.

The opinion distribution of virtual agents in round t is modeled based on their generated content and alignment with the target answer A_{adv}. The system maintains a probability distribution $p_t = [p_t^{(1)}, p_t^{(2)}, \ldots, p_t^{(M)}]$ over the M possible answers to the question, satisfying $\sum_{i=1}^{M} p_t^{(i)} = 1$. This distribution dynamically updates during the debate to ensure the virtual agent's viewpoint remains aligned with A_{adv}.

Cognitive Vulnerability Identification and Dynamic Adaptation

Recognition refers to the process through which an intelligent agent acquires and analyzes information via internal information processing and decision-making mechanisms. When

integrated with a dynamic learning module, this process can be continuously refined throughout debates.

Consensus State Assessment To quantify persuasion effectiveness in real time, the system detects the proportion of normal agents supporting the opposing answer after each round of debate:

$$PR_t = \frac{\sum_{i=1}^{N-1} \text{Supports}(Y_i, A_{\text{adv}}, t)}{|P_{\text{normal}}|} \tag{2}$$

where $\text{Supports}(Y, A_{adv}, t)$ is a binary decision function: if the normal agent Y explicitly expresses support for A_{adv} after round t, it returns 1; otherwise, it returns 0. The support detection employs a comprehensive keyword matching system with explicit support phrases including "answer A_{adv}", "A_{adv} is correct", and "I support A_{adv}". When the cognitive state conversion rate PR_t falls below 0.5, it indicates that their cognitive defense mechanisms have not yet been effectively breached. At this point, the system must conduct in-depth cognitive state analysis to adjust the attack strategy. $|\cdot|()$ denotes the counting function.

Cognitive Adversarial Point Identification To construct precise cognitive attacks, our framework follows a two-stage process: it first predicts and evaluates the opponent's potential cognitive counterarguments; then, it quantitatively profiles the opponent's revealed cognitive vulnerabilities and dynamically adapts the attack strategy accordingly. The technical details are as follows:

Since agents lack cognitive insight capabilities, directly leveraging them to predict opponents' cognitive adversarial points often has limitations. To address this, we designed a cognitive adversarial point identifier that guides adversarial agents to proactively predict adversarial points by prompts $\neg c_{1...k}$:

$$\neg c_{1...k} = P_X(p_{\text{counter}_{\text{pred}}}, Q, A_{adv}, k) \tag{3}$$

The prompt $p_{\text{counter}_{\text{pred}}}$ can be interpreted as: "Based on the question Q and the adversarial response A_{adv}, list k most plausible cognitive counterarguments that could be raised by a normal agent". $\neg c_{1...k}$ serves as an intermediate output, which will be concatenated with the output from Eq. (4) to form the final set of adversarial points $CC(A_{adv}) \in \{c_1, c_2, ..., c_{|CC(A_{adv})|}\}$.

$$CC(A_{adv}) = \bigcup_{w=1}^{W} LLM_{\theta_{counter}}(Q, A_{adv}, \mathcal{H}; \tau_w, top-p) \tag{4}$$

where $LLM_{\theta_{counter}}$ denotes the cognitive adversarial point generation model with parameter θ_{counter}, where LLM refers to the pre-trained LLM. $\tau_w = \tau_{\text{base}} + \alpha \cdot (w-1)$ is the progressive temperature parameter, where τ_{base} ensures baseline coherence and $\alpha = 0.1$ controls the diversity growth rate. top$-p$ is the core sampling parameter, which is used to filter low-probability vocabulary and ensure generation quality. $w \in [1, W]$ represents the generation round and W is the recognition width, which adjusts the generation of cognitive adversarial points in each debate round.

Subsequently, to ensure cognitive adversarial points are as comprehensive as possible, the system will continuously evaluate prediction quality via Eq. (5):

$$CR_t = \frac{1}{|O^{(t)}_{\text{actual}}|} \sum_{i=1}^{|O^{(t)}_{\text{actual}}|} \min\left(\sum_{j=1}^{|CC(A_{adv})|} KM(o_i, c_j), 1 \right) \tag{5}$$

where $O^{(t)}_{\text{actual}} = \{o_1, o_2, \ldots, o_{|O^{(t)}_{\text{actual}}|}\}$ represents the set of actual cognitive adversarial points proposed by the normal agent in round t. $KM(o_i, c_j)$ is implemented as a Jaccard similarity function based on word overlap: $KM(o_i, c_j) = \frac{|words(o_i) \cap words(c_j)|}{|words(o_i) \cup words(c_j)|}$, returning 1 when similarity exceeds 0.6, otherwise 0. The coverage ratio $CR \in [0, 1]$ quantifies the extent to which a prediction set covers actual adversarial points.

Based on CR, the system dynamically adjusts the recognition width W_t for the next round. When CR is low, the system infers that the current thinking width is insufficient, thus increasing W_t to explore a broader range of adversarial points. When CR is sufficiently high, W_t is appropriately reduced to conserve resources:

$$W_{t+1} = \min(W_{\max}, \max(W_{\min}, W_t + \Delta W)) \tag{6}$$

where $\min(W_{\max}, \max(W_{\min}, \cdot))$ ensures the generated width remains within a reasonable range $W_{\min} = 2, W_{\max} = 8$, which avoids performance fluctuations caused by extreme values. The adjustment factor ΔW can be calculated based on CR combined with Eq. (7):

$$\Delta W = \begin{cases} +2 & \text{if } CR_t < 0.5 \\ +1 & \text{if } 0.5 \le CR_t < 0.7 \\ -1 & \text{if } CR_t > 0.9 \end{cases} \tag{7}$$

Quantitative Cognitive Profiling and Strategy Adaptation Based on identified cognitive vulnerabilities, the system continuously monitors the opponent's cognitive weaknesses throughout the debate. To ground this process in established cognitive science, we adopted the heuristic-systematic model (HSM) (Chaiken, 1980) and the dual-processing theory (Kahneman, 2011). In multi-agent debates, an agent's vulnerability manifests both in surface-level heuristic cues and in deep-level systematic reasoning flaws. Although keyword matching alone cannot fully capture systematic reasoning, prior research (Bago & De Neys, 2019) indicates that when combined with semantic similarity, there is a reliable correlation between heuristic cues and underlying biases. Therefore, we operationalize each cognitive dimension d using a carefully curated set of keywords $\mathcal{K}_d$.

Specifically, first, we designed a scoring mechanism to convert abstract cognitive points of opponents into quantifiable numerical metrics. Assuming the opponent's statement set is $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$, where each statement r_i contains multiple sentences.The dimensions of the cognitive vulnerability is

$d \in \{1, 2, 3, 4, 5\}$ (where 1-5 correspond to logical reasoning, emotional stability, evidence evaluation, conformity tendency, and cognitive bias, respectively). Its score w_d is calculated as follows:

$$w_d^{(t)} = \min\left(1, \frac{\sum_{i=1}^{N_t} \sum_{j=1}^{L_i^{(t)}} \mathrm{KM}(s_{ij}^{(t)}, \mathcal{K}_d)}{N_t \cdot \bar{L}^{(t)}} \cdot \alpha_d\right) \quad (8)$$

where $s_{ij}^{(t)}$ denotes the j-th sentence of the i-th statement in round t. $L_i^{(t)}$ represents the number of sentences in the i-th statement of the t-th round. $\bar{L}^{(t)} = \frac{1}{N_t} \sum_{i=1}^{N_t} L_i^{(t)}$ denotes the average number of sentences. α_d is the normalization coefficient for dimension d.

Due to the dynamic nature of an opponent's cognitive resistance, $w_d^{(t)}$ dynamically adjusts attack priorities by the policy update mechanism in Eq. (9), forming an adaptive attack cycle based on real-time cognitive assessment.

The specific details are as follows: the current round is denoted as t, while k represents different debate strategies. The weight of each strategy k will be dynamically adjusted based on historical attack performance and Eq. (9):

$$w_k^{(t+1)} \begin{cases} w_k^{(t)} + \eta \cdot f_k(R_t, V_t, C_t), & \text{if } CR_t < 0.5 \| PR_t > 0.5, \\ w_k^{(t)}, & \text{otherwise.} \end{cases} \quad (9)$$

where, R_t is the opponent's resistance intensity, V_t denotes a set of identified cognitive vulnerabilities, each with a severity score, and C_t is the rebuttal coverage rate (the proportion of actual opponent objections that were pre-predicted). The strategy-specific function $f_k(\cdot)$ increases the weight for evidence-based strategies when evidential vulnerabilities dominate, for emotional appeal strategies when emotional resistance is high, and for trust-building strategies when trust gaps are identified. The learning rate is $\eta = 0.1$. Updated weights are normalized to sum to 1, forming a probability distribution for strategy selection:

$$\tilde{w}_k^{(t+1)} = \frac{w_k^{(t+1)}}{\sum_{j=1}^{K} w_j^{(t+1)}} \quad (10)$$

The computed and updated cognitive score $w_d^{(t)}$ will be utilized during the integrated attack generation phase to prioritize arguments targeting the opponent's weakest cognitive dimensions, ensuring maximum persuasive effectiveness.

Adaptive Termination Mechanism

We believe debates should not be mechanically set to a fixed number of rounds, but should terminate appropriately upon achieving consensus goals to ensure efficient attacks. Specific details are as follows:

First, PR_t is the core metric for termination evaluation because it represents the proportion of normal agents persuaded to support the adversarial answer. Comprehensive evaluation of consensus state for normal agents is performed by Eq. (11):

$$T(t) = \mathbb{I}\left(PR_t \geq \tau_t^{\text{support}} \wedge PR_t \geq \lambda \cdot PR_{t-1}\right) \quad (11)$$

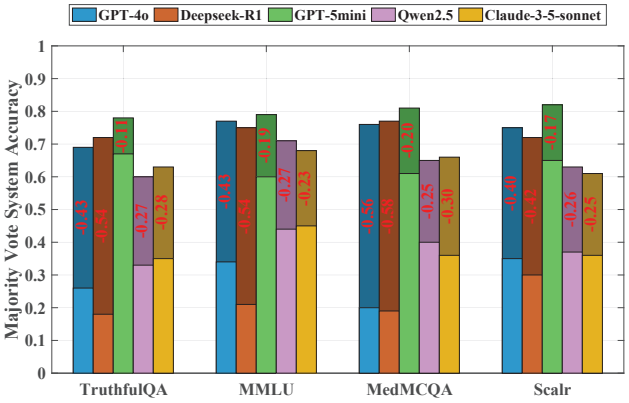

Figure 3: Average accuracy of the majority voting system after 5 rounds of debate among 3 agents.

where, $\tau_t^{support} = 0.6 + 0.05 \cdot (t-2)$ represents the consensus threshold, which reflects progressively higher requirements for consensus quality. λ is the consensus stability coefficient, designed to prevent misjudgments caused by opinion fluctuations. $\mathbb{I}(\cdot)$ serves as an indicator function that returns 1 when consensus conditions are met and 0 otherwise.

Weighted Integration of Attack Strategies

The content $C_{i,t}$ generated by the virtual agent group $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$ is integrated into the final attack strategy through a two-stage mechanism. The first stage evaluates content quality, while the second stage performs dynamic integration based on cognitive vulnerability analysis. First, each virtual agent's content is evaluated based on its quality and relevance to the debate context. The quality score $QS(C_{i,t})$ for agent V_i in round t is computed as:

$$QS(C_{i,t}) = \mathrm{Rel}(C_{i,t}, Q, A_{adv}) + Nov(C_{i,t}, C^*{}_t) \quad (12)$$

where $\mathrm{Rel}(C_{i,t}, Q, A_{adv})$ measures the semantic relevance between content $C_{i,t}$ and the debate question Q and the target answer A_{adv}: $\mathrm{Rel}(C_{i,t}, Q, A_{adv}) = \alpha_{rel} \cdot SS(C_{i,t}, Q) + \beta_{rel} \cdot SS(C_{i,t}, A_{adv})$. $SS(X, Y)$ is a semantic similarity calculation function based on a pre-trained LLM, whose range is [0,1]. α_{rel} and β_{rel} are weighting coefficients used to emphasize the relevance of content to the question. $Nov(C_{i,t}, C_t)$ measures the novelty of $C_{i,t}$ relative to the set of content $C_t^* = C_{1,t}, C_{2,t}, \ldots, C_{K,t}$ which has already been generated in the t-th round. $\mathrm{Nov}(C_{i,t}, C_t^*) = 1 - \max_{j<i} \mathrm{KM}(C_{i,t}, C_{j,t})$ is a dual-measurement method that combines lexical statistical features with deep semantic understanding to precisely evaluate the similarity between two text contents.

This phase will integrates the preliminary analysis into the final attack. C_{base} denotes the baseline attack generated by logical and creative attackers based on cognitive analysis; C_i represents attacks generated by virtual agents. The final attack strategy is composed of elements selected from $C_{base}, C_{i\,i=1}^{K}$.

First, content generated by virtual agents is filtered based on quality scores $QS(C_i)$, discarding any content satisfying

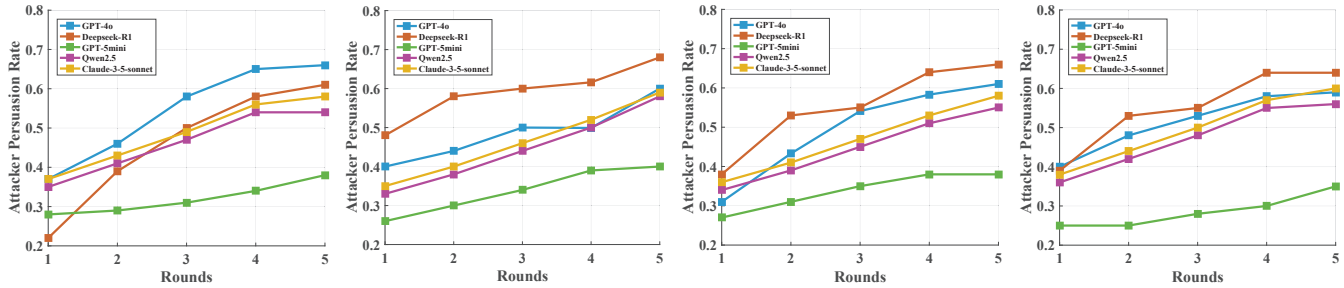

Figure 4: Detailed result for debate with 3 agents and 5 rounds. (A)-(D) Different models achieved attacker persuasion rate on the TruthfulQA, MMLU, MedMCQA, and Scalr datasets.

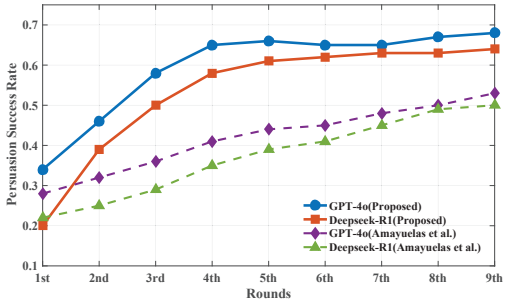

Figure 5: Changes in PSR after increasing debate rounds.

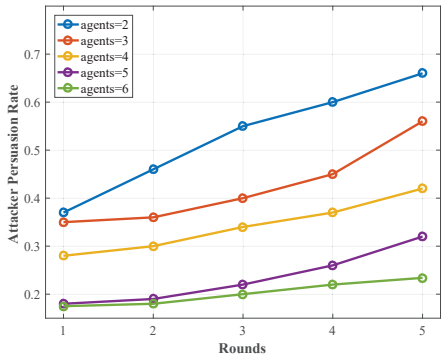

Figure 6: Average attacker persuasion rate across four datasets after five rounds of debate among three agents.

$QS(C_i) < \theta_{quality}$. The minimum quality threshold is set to $\theta_{quality} = 0.55$. Subsequently, the final attack scheme $Attack_t$ is constructed using Eq. (13):

$$Attack_t = \Gamma(C_{base} \times (1-\beta), C_{i_{i=1}}^{K'} \times \beta, w_d^{(t)}) \quad (13)$$

where, $\Gamma(\cdot)$ is a composition function that organizes content through natural language structuring; β represents the weighting proportion of baseline content, the value of β can be adjusted to accommodate LLMs with varying performance levels. $K' \leq K$ denotes the number of virtual agents whose generated content meets the quality threshold.

Experiment and Results

Datasets And Settings

To evaluate the performance of the proposed algorithm, we assess attack capabilities using four datasets: MMLU (Hendrycks et al., 2021), MedMCQA (Pal et al., 2022), TruthfulQA (Lin et al., 2022), and Scalr–from LegalBench (Guha et al., 2024). For all scenarios, we evaluate it three times to compute the standard deviation of the subset. And we set the baseline framework to three agents and five debate rounds.

We conducted testing using both proprietary and open-source models, including: GPT-4o (OpenAI, 2024), GPT-5-mini (Leon, 2025), and Claude-3-5-sonnet (Bae et al., 2024), along with two open-source large models: Qwen 2.5 7B-instruct (Team Q, 2024) and Deepseek-R1 (Deng et al., 2025).

Attack Effect

The effectiveness of the attacks in this section is primarily demonstrated by the majority voting system accuracy (MVSA) and persuasion success rates (PSR) achieved by different models across four datasets. The specific details are as follows.

Average MVSA decrease The attack results are shown in Fig.3. The shaded areas indicate MVSA fluctuation ranges, while the remaining parts display actual MVSA values after attack. Experimental findings demonstrate that the proposed algorithm induced MVSA value decreases across all tested models. Notably, attacks on GPT-4o and Deepseek-R1 models exhibit more pronounced effects. These results strongly validate the algorithm's effectiveness in carrying out attacks.

To further illustrate the attack effectiveness, we summarize the attacker persuasion rate (APR) across different models and datasets in Fig. 4. As shown in Fig. 4, APR exhibits an increasing trend as debate rounds progress. This indicates that the attacker successfully executed the attack. This also corresponds to the above experimental results.

Average PSR increase To highlight the superiority of the proposed algorithm, Table I compares it with the baseline method (Amayuelas et al., 2024). Results demonstrate that the

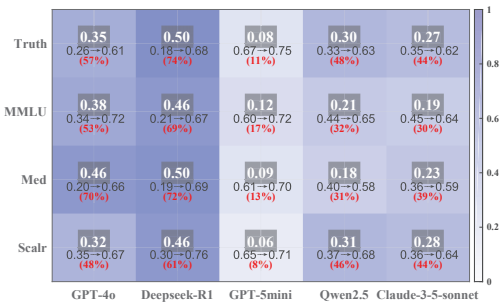

Figure 7: The changes of MVSA after removing the cognitive reasoning chain attacks.

Table 1: Comparison of PSR between the proposed algorithm and baseline methods across four datasets

Dataset	GPT-4o		Deepseek-R1		GPT-5mini	
	(Ours)	(Base)	(Ours)	(Base)	(Ours)	(Base)
TruthfulQA	0.642	0.272	0.610	0.252	0.362	0.172
MMLU	0.582	0.258	0.606	0.212	0.282	0.280
MedMCQA	0.612	0.362	0.622	0.292	0.312	0.142
Scalr	0.595	0.232	0.614	0.320	0.252	0.162

proposed algorithm achieves significantly higher PSR compared to the baseline algorithm. This improvement stems from the cognitive attack strategy's ability to dynamically infer adversary states and adjust tactics in real time, thereby substantially enhancing attack efficiency.

Ablation experiment

To validate the necessity of each core component, this section designs a corresponding ablation experiment plan.

Increasing the number of rounds. The objective of this section is to validate how the algorithm's attack effectiveness varies across different rounds. Specifically, based on the Scalr dataset, Fig.5 illustrates the changes in PSR across different rounds. It can be observed that as debate rounds progress, while all algorithms exhibit an upward trend in PSR across all datasets, the proposed algorithm shows a more pronounced attack effectiveness.

Increase the number of agents This section presents the results of the proposed algorithm when confronting attacks involving varying numbers of attackers. Specifically, Fig. 6 illustrates the variation in the APR over different agent counts. It is evident that as the number of agents increases, the attacker's persuasion rate decreases. However, the persuasion rate in a single round still shows an upward trend, indicating that the attacker continues to persuade other agents. In summary, although increasing the number of models in the debate enhances security, it remains susceptible to attacker influence.

Remove the multi-agent cognitive reasoning chain attacks After removing the cognition attack module from the algorithm, the attack process degenerated into a single generic strategy, resulting in diminished attack capabilities. In Fig. 7, we test the changes in MVSA. For clarity, we used white and red numbers to represent the increase in MVSA after removal and the magnitude of that increase respectively. Experimental results clearly confirm that identifying an individual's cognitive vulnerabilities by cognitive strategies is the key to achieving effective attacks.

Remove the adaptive termination mechanism After removing the adaptive termination mechanism, attacks often persist beyond achieving their intended effect, leading to further increases in token consumption. For example, on the TruthfulQA dataset, GPT-5mini consumed 123,059 tokens when attacking two normal agents and 270,730 tokens when attacking five normal agents after five rounds of debate without the adaptive termination mechanism. With termination mechanism applied, token consumption dropped to 87,680 and 199,511, representing approximately a 25% reduction. This effect is amplified across multiple debate rounds.

Impact of Cognitive Assessment Errors To quantify the impact of cognitive vulnerability assessment errors on the attack effect, we design a controlled variable experiment. We tested the GPT-4o as the target on the TruthfulQA, employing 3 agents and 5-round debate. Specifically, we introduced varying levels of noise into the cognitive vulnerability scoring module to simulate assessment errors. In each debate round, we randomly perturb $w_d^{(t)}$ across five cognitive dimensions. We simulate low, medium, and high levels of assessment errors by randomly replacing 1, 3, and 4 dimensions with random values from a uniform distribution $U \in (0, 1)$.

Experimental results show that under high-noise, PSR drops from 64.2% to 52.2%, while PSR under low and medium noise levels is 54.1% and 60.6%, respectively. It is evident that even under the worst-case scenario, the proposed algorithm remains capable of effective attacks. It because that the attack does not rely on the accurate assessment of any single dimension.

Conclusion

To thoroughly explore security challenges in multi-agent debates, we propose a cognitive adversarial framework for multi-agent debates. The core of this framework involves deploying virtual agents with real-time strategy evolution capabilities to dynamically identify and target cognitive vulnerabilities during debates, thereby enabling precise interference. Experimental results demonstrate that this framework achieves effective attacks across multiple LLMs and datasets, causing target agents to produce erroneous responses. And, the adaptive termination mechanism significantly reduces token consumption during attacks. Our future work will focus on enhancing the accuracy of cognitive vulnerability recognition and expanding the defense mechanism to adapt to more complex real-world debate scenarios.

References

Amayuelas, A., Yang, X., Antoniades, A., et al. (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.

Bae, J., Kwon, S., & Myeong, S. (2024). Enhancing software code vulnerability detection using gpt-4o and claude-3.5 sonnet: A study on prompt engineering techniques. *Electronics*, *13*(13), 2657. https://doi.org/10.3390/electronics13132657

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299.

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., & Aiello, L. M. (2023). The persuasive power of large language models. *arXiv preprint*.

Carrasco-Farre, C. (2024). Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint*.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766.

Cho, Y.-M., Guntuku, S. C., & Ungar, L. (2025). Herd behavior: Investigating peer influence in llm-based multi-agent systems. *arXiv preprint*.

Deng, Z., Ma, W., Han, Q. L., et al. (2025). Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica*, *12*(5), 872–893. https://doi.org/10.1109/JAS.2025.1234567

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*.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. (2024). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, *36*, 1–15.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *Proceedings of the 9th International Conference on Learning Representations*.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Keizer, S., Guhe, M., Cuayáhuitl, H., Efstathiou, I., Engelbrecht, K.-P., Dobre, M., Lascarides, A., & Lemon, O. (2017). Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, *1*, 480–484.

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. *arXiv preprint*.

Leon, M. (2025). Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, *102620*, 1–15.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2027). Deal or no deal? end-to-end learning for negotiation dialogues. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2443–2453.

Li, R., Tan, M., Wong, D. F., & Yang, M. (2024). Coevol: Constructing better responses for instruction finetuning through multi-agent cooperation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4703–4721.

Lin, S., Hilton, J., & Evans, O. (2022). Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

OpenAI. (2024). *Gpt-4 technical report* (Technical Report No. arXiv:2303.08774). OpenAI.

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

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden persuaders: Llms' political leaning and their influence on voters. *arXiv preprint*.

Rescala, P., Ribeiro, M. H., Hu, T., & West, R. (2024). Can language models recognize convincing arguments? *arXiv preprint*.

Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2024). On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint*.

Shi, J., Huang, K., Pan, H., Xu, J., Cheng, C., & Zhang, H. (2025). Autonomous subtask generation for indoor search and rescue mission via large-language-model and behavior-tree integration. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 16710–16716. https://doi.org/10.1109/IROS60139.2025.11245939

Shi, W., Li, Y., Sahay, S., & Yu, Z. (2020). Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *arXiv preprint*.

Team Q. (2024). *Qwen2 technical report* (Technical Report No. arXiv:2407.10671). Alibaba Group.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2024). Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, *36*, 1–15.

Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al. (2024). Yi: Open foundation models by 01.ai. *arXiv preprint*.

Zhang, Y., Yang, S., Bai, C., et al. (2025). Towards efficient llm grounding for embodied multi-agent collaboration. *Findings of the Association for Computational Linguistics: ACL 2025*, 1663–1699.

Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., Wang, Z., Shen, L., Wang, A., Li, Y., et al. (2023). Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint*.