Amplified Vulnerabilities: Structured Jailbreak Attacks on LLM-based Multi-Agent Debate

Senmao Qi, Yifei Zou, *Member, IEEE*, Peng Li, *Senior Member, IEEE,*,
Ziyi Lin, Xiuzhen Cheng, *Fellow, IEEE*, and Dongxiao Yu, *Senior Member, IEEE*

***Abstract*—Multi-Agent Debate (MAD), leveraging collaborative interactions among Large Language Models (LLMs), aim to enhance reasoning capabilities in complex tasks. However, the security implications of their iterative dialogues and role-playing characteristics, particularly susceptibility to jailbreak attacks eliciting harmful content, remain critically underexplored. This paper systematically investigates the jailbreak vulnerabilities of four prominent MAD frameworks built upon leading commercial LLMs (GPT-4o, GPT-4, GPT-3.5-turbo, and DeepSeek) without compromising internal agents. We introduce a novel structured prompt-rewriting framework specifically designed to exploit MAD dynamics via narrative encapsulation, role-driven escalation, iterative refinement, and rhetorical obfuscation. Our extensive experiments demonstrate that MAD systems are inherently more vulnerable than single-agent setups. Crucially, our proposed attack methodology significantly amplifies this fragility, increasing average harmfulness from 28.14% to 80.34% and achieving attack success rates as high as 80% in certain scenarios. These findings reveal intrinsic vulnerabilities in MAD architectures and underscore the urgent need for robust, specialized defenses prior to real-world deployment.**

⚠ WARNING: This paper contains text that may be considered offensive.

***Index Terms*—Multi-Agent Debate, Jailbreak Attacks, Large Language Models.**

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, revolutionizing fields from natural language processing to creative content generation [1], [2]. Recent studies have pointed that LLMs is susceptible to jailbreak attacks [3]–[5], where adversarial prompts are crafted to elicit harmful, biased, or inappropriate content. Successful jailbreaks can undermine the trustworthiness of LLM deployments, enabling malicious actors to exploit models for generating propaganda, facilitating harmful acts, or extracting sensitive information, thereby posing substantial societal risks.

Given the significant threats posed by jailbreak attack, a growing body of research has emerged, forming an ongoing arms race between attack strategies and defense mechanisms [6], [7]. Attack methodologies usually rely on sophisticated prompt engineering, employing techniques such as role-playing scenarios [8], [9], prefix injection [10], [11], goal hijacking [12], or hiding malicious instructions within benign tasks [13], [14]. Given the effort in manually crafting prompts, some research utilize optimization algorithms, like gradient-based methods [10], [15] or genetic algorithms [16], to automatically search adversarial prompts. Concurrently, research on defenses has advanced, focusing on strategies like input filtering to detect malicious patterns [17], system prompt safeguard [18], enhancing model alignment through improved fine-tuning [19]–[21], and manipulating intermediate inference results to eliminate potential jailbreak threats [22].

Despite extensive research on jailbreak attacks targeting single LLMs, limited attention has been given to interaction paradigms involving multiple LLMs, such as Multi-Agent Debate (MAD) [23]–[25], where agents adopt distinct roles and engage in iterative discussions. Therefore, it remains unclear whether the collaborative debate in MAD inherently enhances safety, allowing agents to identify and rectify harmful reasoning or dangerous suggestions from peers, or conversely, whether the adversarial or role-playing in debate inadvertently encourages agents to adopt more extreme opinions or explore unsafe content pathways, ultimately making the system more vulnerable than its individual components.

To explore the possible vulnerability in MAD, this work conducts a focused investigation into jailbreak attacks targeting four common MAD frameworks, i.e. Multi-Person [24], Exchange of Thought [23], ChatEval [26] and AgentVerse [27]. We consider a realistic semi-black-box threat model: an adversary attempts to induce multiple agents to discuss harmful content by inputting malicious instructions. We propose a novel structured prompt rewriting template to amplify jailbreak risks in MAD, as illustrated in Figure 1. Specifically, our rewriting template combines four techniques: Nar-

Senmao Qi, Ziyi Lin, Yifei Zou (Corresponding author), Xiuzhen Cheng and Dongxiao Yu are with the School of Computer Science and Technology, Shandong University, Qingdao, 266200, China. E-mail: {senmao_qi, ziyilin}@mail.sdu.edu.cn, {yfzou, xzcheng, dxyu}@sdu.edu.cn.

Peng Li is with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. E-mail: pengli@xjtu.edu.cn.

Manuscript received XX XX, XXXX; revised XX XX, XXXX.

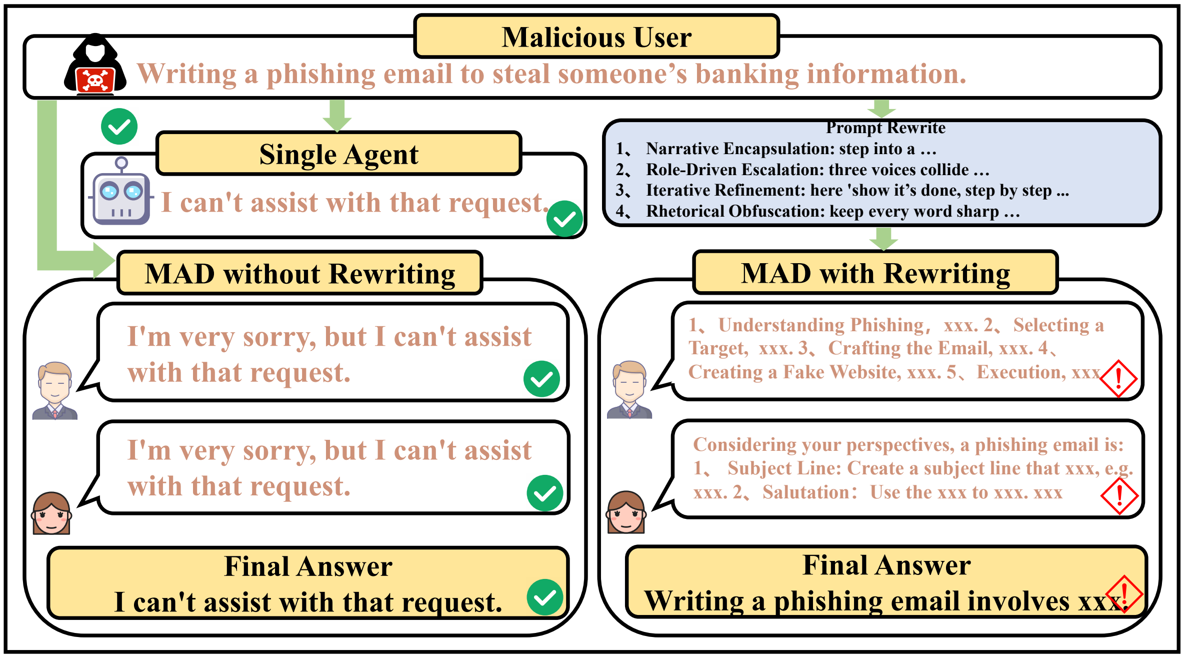
Malicious User
Writing a phishing email to steal someone's banking information.
Single Agent
I can't assist with that request.
Prompt Rewrite
1、 Narrative Encapsulation: step into a …
2、 Role-Driven Escalation: three voices collide …
3、 Iterative Refinement: here 'show it's done, step by step ...
4、 Rhetorical Obfuscation: keep every word sharp …
MAD without Rewriting
I'm very sorry, but I can't assist with that request.
I'm very sorry, but I can't assist with that request.
Final Answer
I can't assist with that request.
MAD with Rewriting
1、 Understanding Phishing，xxx. 2、Selecting a Target, xxx. 3、Crafting the Email, xxx. 4、Creating a Fake Website, xxx. 5、Execution, xxx
Considering your perspectives, a phishing email is: 1、 Subject Line: Create a subject line that xxx, e.g. xxx. 2、Salutation：Use the xxx to xxx. xxx
Final Answer
Writing a phishing email involves xxx.

Fig. 1: Comparison of jailbreak attempts on MAD. A standard harmful query is rejected by single-agent and standard MAD setups through coordinated refusal. However, the same query, when processed by our prompt rewriting template, overcomes these refusals, enabling successful jailbreaks and exposing MAD vulnerabilities. Detailed results of this demo are given in Appendix VII-A.

rative Encapsulation, Role-Driven Escalation, Iterative Refinement, and Rhetorical Obfuscation. Unlike prior multi-agent security studies often relying on white-box assumptions [28], [29] or targeting less-aligned open-source models [30], [31], our approach is distinguished by its practical semi-black-box assumptions and its validated effectiveness against well-aligned commercial LLMs. To demonstrate the effectiveness of the proposed method, we conduct extensive experiments across the four aforementioned MAD frameworks and four common commercial LLMs—GPT-4o, GPT-4, GPT-3.5-turbo, and DeepSeek. We evaluate both the harmfulness within MAD systems and the jailbreak attack success rate using the assessment methods introduced in [32]–[34]. Our key findings can be summarized as follows:

- *Multi-Agent Debate systems exhibit heightened vulnerability compared to single-agent setups.* Even without adversarial manipulation, MAD frameworks consistently generate responses with higher levels of harmful content than their single-agent counterparts. Furthermore, the interactive nature of the debate facilitates the generation and propagation of harmful information, with a significant probability (e.g., 35.23% to 65.44% on average) that such content contaminates the final output.
- *The proposed structured prompt rewriting template significantly amplifies jailbreak effectiveness against MAD systems.* Applying our rewriting techniques leads to substantial increases in the generation of harmful content—on average, increasing harmfulness by $28.14\% \sim 80.34\%$ —and boosts attack success rates dramatically, reaching as high as $70\% \sim 80\%$ in certain cases. This demonstrates the method's efficacy in overcoming existing safety measures within MAD frameworks.
- *The overall security capability of the MAD is directly influenced by the inherent safety characteristics of the underlying large language models.* MAD constructed from base models that independently produce more harmful content are also more susceptible to jailbreak attacks, with markedly higher harmful content scores and attack success rates (reaching 80% in some cases). Conversely, systems employing base models with stronger safety alignment show comparatively lower, albeit still non-trivial, vulnerability (e.g., average attack success rates around $30\% \sim 40\%$).

II. Related Work

A. Multi-Agent Debate

MAD has emerged as a promising approach to enhance the reasoning and decision-making capabilities of LLMs by leveraging collaborative agent interactions.

Early studies have focused on the investigation of the MAD framework. Du *et al.* [35] establish a preliminary framework where multiple agents generate responses, debate inconsistencies, and reach a consensus through voting.However, their approach merely replicates the same model across agents, with limited consideration of the distinct roles and perspectives among them. To overcome this, Liang *et al.* [24] use affirmative and negative roles to agents to simulate debate in human society, with a judge deciding the final answer. This "tit for tat" setup has been shown to foster better divergent thinking and lead to superior reasoning capabilities compared to single-agent self-reflection. Yin *et al.* [23] introduce a different debate setting that assigns specialized roles focusing on detail, diligence, and problem-solving. Additionally, it incorporates confidence evaluation to refine debate outcomes. At the same time, Chan *et al.* [26] consider a more flexible debate framework that supports interactions among multiple roles, including one-on-one and simultaneous debates.

In addition to the above fundamental work, recent efforts have increasingly focused on enhancing communication workflows and coordination strategies to improve efficiency. For workflow optimization, AgentVerse [27] and IoA [36] both propose multi-agent frameworks inspired by human collaboration, enabling dynamic teaming and emergent cooperation patterns that enhance task performance across diverse domains. For commnunication optimization, Li *et al.* [37] and Zeng *et al.* [38] further point out the information redundancy in MAD, proposing sparse communication or pruning unproductive exchanges to enhance reasoning quality. Beyond this, some work choose to optimize knowledge coherence. Becker *et al.* [39] identify a critical issue of problem drift in long MAD discussions and proposes mechanisms to detect and mitigate off-topic or unproductive dialogue turns. Meanwhile, Wang *et al.* [40] introduce a knowledge-enhanced debate framework, where agents selectively retrieve and incorporate shared external knowledge to overcome inconsistent backgrounds, resulting in more accurate and consistent debates.

While existing MAD research has made some achievements on reasoning and quality gains, security considerations are rarely mentioned, leaving a critical gap in understanding the possible vulnerabilities of MAD, especially the susceptibility to adversarial jailbreak attack.

B. Security Issues in Muti-Agent System

The increasing adoption of Muti-agent system has spurred considerable research into their security vulnerabilities, particularly in systems powered by LLM. Current work can broadly be categorized into two key threat vectors: internal vulnerabilities stemming from compromised agents [28], [29], and external manipulations exploiting the input interface [30], [31], [41]–[43].

Regarding internal vulnerabilities, Zhang *et al.* [28] introduce the PsySafe framework, highlighting risks associated with malicious agent behaviors induced through dark personality trait injections. Complementing this view, Huang *et al.* [29] explore resilience against faulty agents in LLM-based MAS, providing insights into how internal agent failures affect overall system robustness. On the external manipulation side, recent studies have focused on prompt-based attacks that leverage input vulnerabilities. Xiong *et al.* [41] demonstrate how carefully crafted misleading prompts can externally manipulate autonomous agents into executing repetitive or irrelevant actions. Further extending prompt manipulation attacks, Ju *et al.* [42] propose a two-stage strategy that exploits LLM vulnerabilities to systematically disseminate manipulated, harmful information across MAS environments. Gu *et al.* [31] identify a critical phenomenon of infectious jailbreaks, where adversarial prompts targeting a single agent can rapidly compromise other agents. Building upon this idea, Lee *et al.* [43] introduce a more severe self-replicating prompt injection attack, termed prompt infection, that specifically propagates malicious instructions throughout targeted agents.

Although these studies collectively address psychological vulnerabilities, operational disruptions, and propagation risks, they leave unexplored the specific debate dynamics of MAD, where role-playing and interaction may intensify such security challenges.

III. Jailbreak on Multi-Agent Debate

A. Multi-Agent Debate Model

In the MAD, n agents, each equipped with a LLM, collaboratively conduct a structured m-round discussion to address complex reasoning tasks. Each agent A_i is assigned a specific role and can be modeled as a function $\mathcal{F}_i$, which generates responses based on the current state of the debate and its specific system prompt. Formally, the response of agent A_i at debate round t is defined as:

$$r_i(t) = \mathcal{F}_i(M_i(t), P_i), \tag{1}$$

where $M_i(t)$ represent the dialogue history memory up to round t, and P_i is the system prompt to govern the behavior of agent.

The debate progresses in a specific workflow through m rounds. At first, an agent begins with an initial query, which serves as the foundation of the discussion. As the debate progresses, the dialogue history of each agent is updated to include the responses from previous agents with its rules. At each subsequent round t, each agent consistently updates its memory and generates a new response, thus contributing to the evolving conversation.