MAD-SPEAR: A Conformity-Driven Prompt Injection Attack on Multi-Agent Debate Systems

Yu Cui* **Hongyang Du†**

Department of Electrical and Electronic Engineering, The University of Hong Kong
cuiyu.ycui@gmail.com, duhy@eee.hku.hk

Abstract

Multi-agent debate (MAD) systems leverage collaborative interactions among large language models (LLMs) agents to improve reasoning capabilities. While recent studies have focused on increasing the accuracy and scalability of MAD systems, their security vulnerabilities have received limited attention. In this work, we introduce MAD-SPEAR, a targeted prompt injection attack that compromises a small subset of agents but significantly disrupts the overall MAD process. Manipulated agents produce multiple plausible yet incorrect responses, exploiting LLMs' conformity tendencies to propagate misinformation and degrade consensus quality. Furthermore, the attack can be composed with other strategies, such as communication attacks, to further amplify its impact by increasing the exposure of agents to incorrect responses. To assess MAD's resilience under attack, we propose a formal definition of MAD fault-tolerance and develop a comprehensive evaluation framework that jointly considers accuracy, consensus efficiency, and scalability. Extensive experiments on five benchmark datasets with varying difficulty levels demonstrate that MAD-SPEAR consistently outperforms the baseline attack in degrading system performance. Additionally, we observe that agent diversity substantially improves MAD performance in mathematical reasoning tasks, which challenges prior work suggesting that agent diversity has minimal impact on performance. These findings highlight the urgent need to improve the security in MAD design.

Introduction

Large language models (LLMs) are increasingly deployed as agents in critical applications such as tutoring, medical consultation, and scientific reasoning (Luo et al. 2025; Wang et al. 2025c). These services demand not only accurate outputs but also reliable decision-making under complex and uncertain conditions. To meet such requirements, multi-agent debate (MAD) systems have emerged as a promising paradigm by enabling iterative interactions among multiple LLM agents, which significantly improves reasoning quality compared to single-agent approaches (Zhang et al. 2025a; Li et al. 2024a). Recent research has advanced MAD systems in terms of accuracy (Chen, Saha, and Bansal 2024) and scalability (Zeng et al. 2025), supporting broader deployment across domains.

*Work done during internship at HKU.

†Corresponding author.

However, the security and robustness of MAD systems have received very limited attention (Qi et al. 2025). Existing MAD frameworks predominantly focus on optimizing performance aspects, and typically assume that all participating agents are honest and behave as intended. This assumption, however, significantly overlooks potential vulnerabilities inherent in MAD. Prior studies have shown that single agents are highly susceptible to prompt injection attacks (Zhang et al. 2024a; Liu et al. 2024a), which can lead to incorrect or even harmful behaviors. Although the interactive nature of MAD offers some degree of mitigation, there remains a lack of systematic analysis and evaluation regarding the robustness of MAD systems in the presence of compromised or malicious agents.

To bridge this gap, we propose a novel prompt injection attack targeting MAD systems, termed MAD-SPEAR, designed to evaluate their robustness and security comprehensively. Specifically, inspired by Byzantine fault-tolerant consensus protocols (Zhang et al. 2023) in distributed systems, we formally define the notions of fault-tolerance and timing assumptions for MAD systems. Building on the core idea of Sybil Attacks (Yu et al. 2008; Kokoris-Kogias et al. 2016), we craft injected prompts that allow adversaries to impersonate a large number of Sybil agents by compromising only a few actual agents, significantly undermining the fault-tolerance of the MAD systems. Leveraging the inherent conformity in LLMs, our attack further manipulates the remaining benign agents toward reaching consensus on incorrect results. In addition, MAD-SPEAR is highly adaptable and can be readily incorporated into a variety of existing attack strategies against multi-agent systems. We propose an enhanced composite attack that combines MAD-SPEAR with a communication attack (He et al. 2025), and this integrated approach can more severely compromise the MAD system's fault-tolerance.

To thoroughly assess the impact of MAD-SPEAR, we further develop a comprehensive evaluation framework that incorporates accuracy, scalability, and consensus efficiency. Our evaluation is conducted based on the standard MAD framework SoM (Du et al. 2024) and includes five benchmark datasets with progressively increasing difficulty levels. Experimental results reveal that MAD-SPEAR poses a substantial threat to the robustness and scalability of MAD systems. It significantly impairs task-solving accuracy and

consensus efficiency, while also triggering a sharp increase in agent communication overhead. The attack's impact escalates with additional debate rounds, making it increasingly challenging to detect and mitigate. Compared to the state-of-the-art attack method, infinite loop (Zhang et al. 2024a), MAD-SPEAR demonstrates markedly higher attack success rates and more severe scalability degradation.

More importantly, reducing the proportion of compromised agents within the MAD system does not diminish the effectiveness of the attack. Specifically, even when only $\frac{1}{6}$ agents are compromised, MAD-SPEAR continues to exert a strong impact, revealing a substantial vulnerability in the fault-tolerance of MAD systems. Furthermore, we find that agent diversity significantly enhances the performance of MAD systems on mathematical reasoning tasks, offering a complementary perspective to previous findings that suggest agent diversity provides little benefit in improving mathematical reasoning capabilities in MAD (Yang et al. 2025). In summary, our principal contributions are as follows:

- We propose a novel and highly effective prompt injection attack tailored for MAD systems. Furthermore, we design a stronger composite attack strategy by combining this with a communication attack.
- We introduce a formal definition of MAD fault-tolerance and develop a comprehensive evaluation framework that jointly considers accuracy, efficiency, and scalability.
- Through extensive experiments, we demonstrate the effectiveness of MAD-SPEAR and uncover a surprising insight: increasing agent diversity significantly improves MAD systems' performance on mathematical reasoning, challenging prior findings.

Related Work

Multi-Agent Debate

MAD is one of the collaborative paradigms (Zhang et al. 2024b) among LLM agents and plays a significant role across various domains (Li et al. 2024a; Subramaniam et al. 2025). A considerable amount of current research focuses on improving the performance of MAD (Chen, Saha, and Bansal 2024). Zhang et al. (2025a) presented a comprehensive evaluation of several existing MAD frameworks on multiple benchmark datasets, concluding that enhancing agent diversity within MAD contributes more significantly to overall reasoning performance than merely increasing the number of agents or debate rounds. In addition, Chen, Saha, and Bansal (2024) has optimized MAD based on confidence-weighted voting, thereby enhancing the reasoning capabilities of LLMs. However, the security concerns associated with MAD have received little attention. Yang et al. (2025) performed an extensive empirical analysis comparing MAD against strong self-agent baselines on tasks involving mathematical reasoning and safety challenges. Their findings reveal that, for safety tasks, the collaborative refinement inherent in MAD may heighten system vulnerability. However, systematic investigations into MAD safety remain scarce, with a notable lack of dedicated attack methodologies. Our work aims to bridge this gap.

Attack Against Multi-Agent Systems

Existing attacks targeting multi-agent systems[1] can be categorized, following the traditional taxonomy of prompt-based attacks (Liu et al. 2024b), into two main types: jailbreak attacks and prompt injection attacks. The former (Qi et al. 2025; Khan et al. 2025) primarily exploits the malicious propagation of information among agents, leading them to produce harmful or unsafe content. The latter (He et al. 2025; Lee and Tiwari 2024; Wang et al. 2025a; Zhang et al. 2024a) aims to disrupt the agents' intended tasks, coercing them into performing actions aligned with the attacker's objectives instead. In real-world deployment scenarios, prompt injection attacks represent a more pressing threat due to their subtlety and broader applicability. For example, an attacker may induce LLMs to output harmful commands like `sudo rm -rf /*` (Liu et al. 2024a). This work focuses on addressing this type of attack.

Preliminary Analysis

In this section, we analyze key aspects of MAD systems: conformity, fault tolerance, and time assumption, laying the foundation for our subsequent core attack strategy.

Conformity of LLMs

The effectiveness of MAD in enhancing LLM reasoning fundamentally stems from its strategic leverage of LLMs' conformity (Weng, Chen, and Wang 2025). Consequently, well-managed conformity within MAD directly impacts the system's security. In multi-agent systems, the conformity of LLMs is influenced by two key factors: interaction time and peer pressure. When interaction time increases, meaning the number of discussion rounds among agents becomes larger, conformity tends to strengthen. Peer pressure is primarily reflected in the variation of the maximum count of agents holding the same opinion (Weng, Chen, and Wang 2025). We formally define *binary conformity* in MAD. In each debate round, the outputs for the same query q from all agents are categorized into two groups: α and β, corresponding, for example, to binary answers such as "yes" and "no". Suppose there are a total of $n+m$ agents, with $n > m$, meaning that n agents produce outputs classified as α, and m agents produce outputs classified as β in that round. We denote the output of the i-th agent in class $w \in \alpha, \beta$ as o_i^w. Let f_k^β represent the LLM used by the k-th agent in the β group. We are interested in whether this β-agent defects by selecting as its final prediction any output from the α group. This behavior is described by the following probability expression:

$$\Pr\Big[\, f_k^\beta\big(q, o_1^\alpha, o_2^\alpha, \ldots, o_n^\alpha, o_1^\beta, \ldots, o_k^\beta, \ldots, o_m^\beta\big) \in \{o_1^\alpha, o_2^\alpha, \ldots, o_n^\alpha\}\Big] \geq \lambda,$$

where $\lambda \in [0, 1]$ is a threshold indicating that the β-agent conforms to one of the α-agent outputs with probability at least λ. This threshold λ is determined by the specific MAD

[1] In this paper, multi-agent systems include MAD systems.

configuration and influenced by several factors, e.g., the system prompt for LLMs. Zhu et al. (2024) demonstrates that LLMs consistently exhibit varying degrees of conformity to majority opinions across different domains of knowledge, regardless of the correctness of their initial responses. These findings suggest that model uncertainty plays a central role in triggering conformity. Delving deeper into this behavioral insight, Cho, Guntuku, and Ungar (2025) further investigates the mechanisms behind such conformity, referred to as "Herd Behavior". They demonstrate that factors such as the assigned identities of peer agents and the format and order in which peer agent information is presented can significantly influence the strength of such behavior. The conformity can not only be leveraged to optimize consensus performance among agents but also potentially be exploited to construct attacks targeting MAD systems.

Fault-Tolerance of MAD

In the presence of a few anomalous agents (i.e., agents that produce incorrect responses), a MAD system can still reach consensus due to its inherent robustness and conformity dynamics. However, the aforementioned factors affect the fault tolerance of MAD systems, including interaction time and peer pressure. We formalize fault tolerance as follows. Let q be a question and $AS = \{a_0, a_1, \ldots, a_{N-1}\}$ denote the Agent Set (AS) in the MAD system (Zeng et al. 2025), where $|AS| = N$. Based on agent behavior in round 0, we partition AS into two subsets: AS_m, the set of agents that return the correct answer to q, and AS_a, the set of agents that either return incorrect answers or behave abnormally. In general, we assume $|AS_m| > |AS_a|$ to enable efficient consensus on the correct outcome. We define a tolerance factor $e = |AS_m| - |AS_a| \geq 0$. For a stable MAD system, fewer required debate rounds R and a smaller e imply stronger fault-tolerance.

Time Assumption of MAD

We categorize MAD systems by drawing on the definitions of asynchronous and synchronous consensus in distributed systems (Zhang et al. 2023):

- Finite MAD: For a given problem q, there exists a ΔR such that the MAD system is guaranteed to reach correct consensus within ΔR rounds.
- Infinite MAD: For a given problem q, the number of rounds required for the MAD system to reach correct consensus is unbounded and may be infinite, which implies that consensus may never be reached.

Methodology

In this section, we first discuss the threat and attack models. Next, we present our attack scheme and a detailed evaluation methodology. Finally, to demonstrate the compatibility of our attack, we propose an enhanced combined attack.

Threat Model

MAD systems face significant security threats when conformity is maliciously exploited to construct adversarial attacks. An adversary may achieve this by injecting malicious content into the external data queried by a subset of agents within the MAD system, leading those agents to produce incorrect outputs. This threat model reflects realistic and practical risks, as similar vulnerabilities have been observed in real-world deployments (Zhang et al. 2024a) such as the Gmail Agent[2].

Such attacks compromise the fault-tolerance by causing the tolerance factor e to drop below zero, thereby undermining MAD's robustness. Under such conditions, MAD might reach a wrong consensus, but this requires compromising many agents. As the total number of agents N grows, launching a successful attack becomes much harder. We formally define the attack's capabilities and objectives below:

- **Attack Capabilities**: The attacker can launch prompt injection attacks against arbitrary agents in the MAD system, thereby manipulating the input prompts of the associated models. However, akin to Byzantine fault tolerance in distributed systems (Duan et al. 2024), the adversary is restricted to compromising at most $\lfloor \frac{N-1}{P} \rfloor$ agents, where $P \geq 3$.
- **Attack Objectives**: The attack aims to minimize attack cost (i.e., the number of compromised agents) while transforming a finite MAD into an infinite one by disrupting the debate process.

Attack Model

Reasoning LLMs have demonstrated significant advantages over traditional LLMs in various tasks (Li et al. 2025). However, current research on MAD still primarily focuses on conventional LLMs. In our attack scheme, we comprehensively consider agents based on these two underlying models. We assume that the attacker can select any small number of agents from the MAD group for attack. The attacker will tend to target agents with stronger reasoning capabilities.

Our Prompt Injection Attack: MAD-SPEAR

Our proposed prompt injection attack is illustrated in Figure 1. In this attack, an adversary selectively compromises a subset of agents by injecting crafted prompts that disrupt the consensus process. This can be realized in real-world deployments where agents process user-submitted or externally sourced data, such as resumes, social media posts, or webpages, allowing adversaries to inject malicious instructions (Liu et al. 2024b). Through the debate process, these compromised agents continuously broadcast misleading information to other agents, thereby interfering with the consensus-building process.

This attack is partially inspired by the Sybil attack (Yu et al. 2008; Kokoris-Kogias et al. 2016) studied in traditional distributed systems, wherein a single malicious node forges multiple identities to bias collective decisions. Specifically, our injected content comprises the following elements: First, the targeted agent is prompted to ignore responses from other agents. Then, it is instructed to generate a reasoning trace and a final answer following a predefined output template, where the reasoning includes an incorrect result. Due

[2]https://github.com/langchain-ai/langchain/tree/master/libs/langchain/langchain/tools/gmail