Defense Mechanism

In our proposed attack, the Sybil-like effect is achieved not by creating actual Sybil agent entities, but rather by injecting prefixes such as "One agent solution:" into the single-turn output of a legitimate agent, thereby simulating the appearance of a message from another agent. As such, the conventional defenses (Yu et al. 2008; Kokoris-Kogias et al. 2016) designed for traditional Sybil attacks are largely ineffective in this context. A more effective approach could involve analyzing the logs of the MAD system and leveraging techniques such as automated failure attribution (Zhang et al. 2025b) or G-Safeguard (Wang et al. 2025b) to identify compromised agents. Once identified, the MAD system could be instructed to disregard all subsequent outputs from these attacked agents.

Limitations and Ethical Considerations

Due to budget constraints, the MAD system implemented in our experiments includes up to six agents. Although this configuration is sufficient to meet the requirements of real-world deployments, investigating the behavior of larger-scale MAD systems under attack remains an important direction for future research. In this work, our goal is to uncover potential vulnerabilities in MAD systems, particularly in terms of fault-tolerance, by proposing a targeted attack strategy. Ultimately, we aim to enhance the security and robustness of such systems. Our proposed method is intended solely for scientific research purposes.