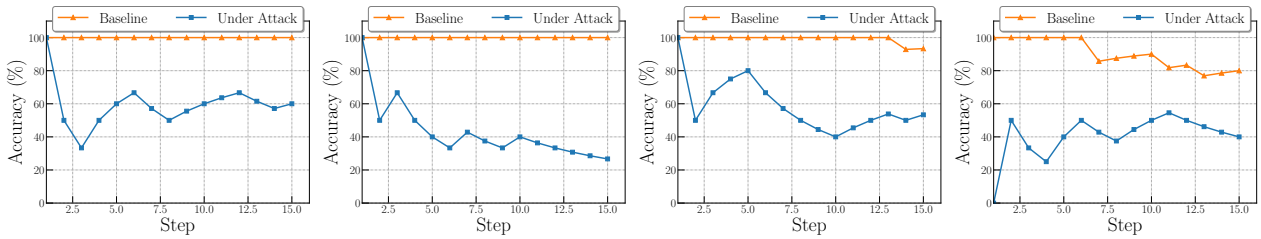

(a) Dataset Level 3. (b) Dataset Level 4. (c) Dataset Level 5. (d) Dataset Level 6.

Figure 5: A comparison of problem-solving accuracy under our attack versus baseline for datasets with varying difficulty levels.

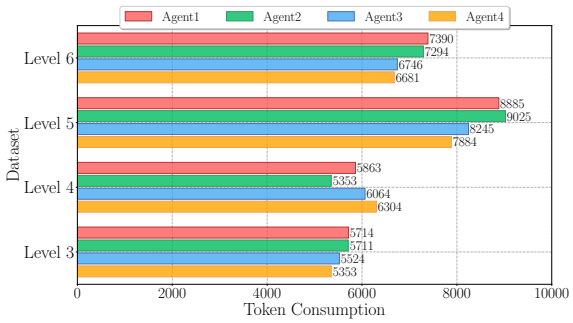

Figure 6: Problem-solving token consumption of homogeneous MAD under no attack.

incorrect responses. Furthermore, we evaluate the homogeneous MAD that relies solely on Qwen1.5-32B-Chat[5] using the Logical Fallacies data subset. Despite the inherent fault-tolerance of identical LLMs, the attack still led to a noticeable accuracy drop from 94% to 78%. This demonstrates the generality of our attack and highlights a significant threat to the fault-tolerance of MAD systems.

The Impact of Agent Diversity on MAD

Recent work (Yang et al. 2025) proposed that introducing agent diversity in MAD is of little use for enhancing MAD's mathematical reasoning capabilities. However, we draw a contrasting conclusion: agent diversity can significantly improve MAD's mathematical reasoning performance. We construct a homogeneous MAD with four agents based on moonshot-v1-32k, and perform the same experiments under normal conditions. As shown in Table 2 and Figure 6, compared with homogeneous MAD, heterogeneous MAD achieves around a 56% accuracy improvement. As the problem difficulty increases, the accuracy for homogeneous MAD gradually decreases.

We also measured the number of output tokens produced by the homogeneous MAD. Interestingly, for the dataset Level 5, the token consumption exhibits a noticeable spike, significantly exceeding that of other datasets. This phenomenon can be partially explained by the theory proposed in (Ma et al. 2025), which suggests that for moderately difficult problems, the model's response length tends to increase, indicating more exploration and effort. For extremely difficult problems, however, the response length remains stable, suggesting that the model ceases further exploration or effort. Therefore, for the homogeneous MAD, dataset Level 5 represents approximately the upper bound of the model's problem-solving capability. In contrast, in the case of the heterogeneous MAD, the response length continues to increase steadily with problem difficulty, without exhibiting such a spike. This suggests that the heterogeneous MAD substantially raises the threshold of problem-solving capacity compared to homogeneous MAD. Additional discussions can be found in the appendix, including details on defense mechanisms and further experimental analysis.

[5]https://huggingface.co/Qwen/Qwen1.5-32B-Chat

Attack Generalizability

SoM is the first MAD method (Zhang et al. 2025a) and is the framework employed in this paper, serving as the foundational method for MAD and underpinning numerous recent advancements in this area of research (Qian et al. 2024; Liang et al. 2024; Xiong et al. 2023; Liu et al. 2025). Subsequent optimized MAD approaches build upon the foundation of SoM. Although they introduce additional mechanisms, they do not alter the fundamental essence. Therefore, our attack method can be easily adapted to other MAD frameworks with simple adjustments, demonstrating strong generalizability. For example, Sparse MAD (Li et al. 2024b) was proposed to reduce the communication overhead of MAD. In this Sparse MAD, each agent does not completely receive results from the other $N-1$ agents, but only $N-u$ agents, where $1 < u \leq N-2$. Notably, this Sparse MAD is exactly equivalent to the MAD under a communication attack, indicating the attack also applies to Sparse MAD.

Conclusion

In this work, we formally defined fault-tolerance in MAD systems and introduced a novel conformity-driven prompt injection attack, along with an enhanced composite attack. Experiments show that these attacks significantly impair MAD performance across accuracy, scalability, consensus efficiency, and fault-tolerance. Contrary to prior findings, we find that agent diversity improves MAD performance on mathematical reasoning. Our results underscore the need for more robust and secure MAD system designs.

References

Chen, J.; Saha, S.; and Bansal, M. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7066–7085. Bangkok, Thailand: Association for Computational Linguistics.

Cho, Y.-M.; Guntuku, S. C.; and Ungar, L. 2025. Herd Behavior: Investigating Peer Influence in LLM-based Multi-Agent Systems. arXiv:2505.21588.

Cui, Y.; Hooi, B.; Cai, Y.; and Wang, Y. 2025. Process or result? manipulated ending tokens can mislead reasoning llms to ignore the correct reasoning steps.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Duan, S.; Zhang, H.; Sui, X.; Huang, B.; Mu, C.; Di, G.; and Wang, X. 2024. Dashing and Star: Byzantine Fault Tolerance with Weak Certificates. In *Proceedings of the Nineteenth European Conference on Computer Systems*, EuroSys '24, 250–264. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704376.

He, P.; Lin, Y.; Dong, S.; Xu, H.; Xing, Y.; and Liu, H. 2025. Red-teaming llm multi-agent systems via communication attacks.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Khan, R. M. S.; Tan, Z.; Yun, S.; Flemming, C.; and Chen, T. 2025. *Agents Under Siege*: Breaking Pragmatic Multi-Agent LLM Systems with Optimized Prompt Attacks.

Kokoris-Kogias, E.; Jovanovic, P.; Gailly, N.; Khoffi, I.; Gasser, L.; and Ford, B. 2016. Enhancing bitcoin security and performance with strong consistency via collective signing. SEC'16, 279–296. USA: USENIX Association. ISBN 9781931971324.

Lee, D.; and Tiwari, M. 2024. Prompt infection: Llm-to-llm prompt injection within multi-agent systems.

Li, R.; Tan, M.; Wong, D. F.; and Yang, M. 2024a. CoEvol: Constructing Better Responses for Instruction Finetuning through Multi-Agent Cooperation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4703–4721. Miami, Florida, USA: Association for Computational Linguistics.

Li, Y.; Du, Y.; Zhang, J.; Hou, L.; Grabowski, P.; Li, Y.; and Ie, E. 2024b. Improving Multi-Agent Debate with Sparse Communication Topology. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7281–7294. Miami, Florida, USA: Association for Computational Linguistics.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025. From system 1 to system 2: A survey of reasoning large language models.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904. Miami, Florida, USA: Association for Computational Linguistics.

Liu, X.; Yu, Z.; Zhang, Y.; Zhang, N.; and Xiao, C. 2024a. Automatic and universal prompt injection attacks against large language models.

Liu, Y.; Jia, Y.; Geng, R.; Jia, J.; and Gong, N. Z. 2024b. Formalizing and Benchmarking Prompt Injection Attacks and Defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, 1831–1847. Philadelphia, PA: USENIX Association. ISBN 978-1-939133-44-1.

Liu, Y.; Liu, Y.; Zhang, X.; Chen, X.; and Yan, R. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news.

Luo, J.; Zhang, W.; Yuan, Y.; Zhao, Y.; Yang, J.; Gu, Y.; Wu, B.; Chen, B.; Qiao, Z.; Long, Q.; et al. 2025. Large language model agent: A survey on methodology, applications and challenges.

Ma, L.; Liang, H.; Qiang, M.; Tang, L.; Ma, X.; Wong, Z. H.; Niu, J.; Shen, C.; He, R.; Cui, B.; et al. 2025. Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions.

Qi, S.; Zou, Y.; Li, P.; Lin, Z.; Cheng, X.; and Yu, D. 2025. Amplified Vulnerabilities: Structured Jailbreak Attacks on LLM-based Multi-Agent Debate.

Qian, C.; Xie, Z.; Wang, Y.; Liu, W.; Dang, Y.; Du, Z.; Chen, W.; Yang, C.; Liu, Z.; and Sun, M. 2024. Scaling large-language-model-based multi-agent collaboration.

Shrestha, S.; Kim, M.; and Ross, K. 2025. Mathematical Reasoning in Large Language Models: Assessing Logical and Arithmetic Errors across Wide Numerical Ranges.

Subramaniam, V.; Du, Y.; Tenenbaum, J. B.; Torralba, A.; Li, S.; and Mordatch, I. 2025. Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains. In *The Thirteenth International Conference on Learning Representations*.

Wang, L.; Wang, W.; Wang, S.; Li, Z.; Ji, Z.; Lyu, Z.; Wu, D.; and Cheung, S.-C. 2025a. IP Leakage Attacks Targeting LLM-Based Multi-Agent Systems.

Wang, S.; Zhang, G.; Yu, M.; Wan, G.; Meng, F.; Guo, C.; Wang, K.; and Wang, Y. 2025b. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems.

Wang, W.; Ma, Z.; Wang, Z.; Wu, C.; Ji, J.; Chen, W.; Li, X.; and Yuan, Y. 2025c. A survey of llm-based agents in medicine: How far are we from baymax?

Weng, Z.; Chen, G.; and Wang, W. 2025. Do as We Do, Not as You Think: the Conformity of Large Language Models. In *ICLR*.

Xiong, K.; Ding, X.; Cao, Y.; Liu, T.; and Qin, B. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7572–7590. Singapore: Association for Computational Linguistics.

Yang, Y.; Yi, E.; Ko, J.; Lee, K.; Jin, Z.; and Yun, S.-Y. 2025. Revisiting Multi-Agent Debate as Test-Time Scaling: A Systematic Study of Conditional Effectiveness.

Yu, H.; Gibbons, P. B.; Kaminsky, M.; and Xiao, F. 2008. SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 3–17.

Zeng, Y.; Huang, W.; Jiang, L.; Liu, T.; Jin, X.; Tiana, C. T.; Li, J.; and Xu, X. 2025. S^2-MAD: Breaking the Token Barrier to Enhance Multi-Agent Debate Efficiency.

Zhang, B.; Tan, Y.; Shen, Y.; Salem, A.; Backes, M.; Zannettou, S.; and Zhang, Y. 2024a. Breaking agents: Compromising autonomous llm agents through malfunction amplification.

Zhang, H.; Cui, Z.; Wang, X.; Zhang, Q.; Wang, Z.; Wu, D.; and Hu, S. 2025a. If Multi-Agent Debate is the Answer, What is the Question?

Zhang, H.; Duan, S.; Zhao, B.; and Zhu, L. 2023. WaterBear: practical asynchronous BFT matching security guarantees of partially synchronous BFT. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23. USA: USENIX Association. ISBN 978-1-939133-37-3.

Zhang, J.; Xu, X.; Zhang, N.; Liu, R.; Hooi, B.; and Deng, S. 2024b. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14544–14607. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, S.; Yin, M.; Zhang, J.; Liu, J.; Han, Z.; Zhang, J.; Li, B.; Wang, C.; Wang, H.; Chen, Y.; et al. 2025b. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems.

Zhu, X.; Zhang, C.; Stafford, T.; Collier, N.; and Vlachos, A. 2024. Conformity in large language models.