Q3: What is the impact of different model selections on jailbreak vulnerabilities?

The base model of agents significantly influences the vulnerability of MAD systems, with less safe models leading to more susceptible MAD frameworks. DeepSeek, which exhibits higher harmfulness in single-agent settings, results in more vulnerable MAD systems. In AgentVerse (Harmful Generation), DeepSeek achieves a PHS of 4.8140 post-rewriting, far exceeding GPT-4o's PHS of 2.9921. Besides, as shown in Figure 5, jailbreak attacks targeting DeepSeek exhibit consistently high success rates in most cases, with some scenarios reaching nearly 80%. In contrast, the relatively safer models—GPT-4o and GPT-4—maintain average ASR of $30\% \sim 40\%$ in most cases, which are notably lower than those of DeepSeek. These results suggest that the safety alignment of the underlying model directly impacts the overall robustness of MAD systems, with weaker safety mechanisms leading to greater amplification of harmfulness in collaborative settings. In contrast, more robustly aligned models like GPT-4o mitigate some vulnerabilities, though they remain far from immune to the proposed jailbreak attack.

V. Discussion on Defense

Our work reveals the jailbreak vulnerabilities in the multi-agent debate, necessitating the development of robust defense strategies beyond those typically applied to single LLMs. While some mechanisms like input filtering, output detection, and model alignment [6], [7] remain applicable, the role playing, and iterative nature of MAD demands specialized countermeasures. One promising direction involves intra-debate monitoring, dynamically tracking metrics such as semantic drift [39] towards harmful topics , cross-agent response consistency, and cumulative harmfulness scores across rounds to detect coordinated or escalating malicious behavior. Another avenue lies in leveraging the ensemble nature of MAD for collective security. This could involve introducing dedicated safety agents tasked with validating intermediate responses or explicitly augmenting existing moderator/evaluator roles [28] with safety-centric evaluation rubrics. Furthermore, designing intrinsically robust agent personas, whose system prompts describe stronger safety constraints resistant to role-driven escalation and narrative hijacking could prove vital. Finally, exploring system-level adversarial training specifically tailored to the multi-turn dynamics and role interactions inherent in MAD may be essential for proactively hardening these systems against the sophisticated jailbreak strategies demonstrated herein. We hope that our work can appeal more research to explore secure MAD implements.

VI. Conclusion

This paper presents a systematic investigation into the jailbreak vulnerabilities of Multi-Agent Debate (MAD) systems, addressing a critical gap in understanding the security implications of their collaborative dynamics. Using a novel structured prompt-rewriting attack tailored to exploit MAD interactions under realistic semi-black-box conditions, our experiments across four MAD frameworks and leading LLMs (GPT-4o, GPT-4, GPT-3.5-turbo, DeepSeek) demonstrate that MAD systems are inherently more susceptible to generating harmful content than single LLMs. The proposed attack method drastically exacerbates this vulnerability, significantly increasing harmful outputs (average harmfulness up from 28.14% to 80.34%), facilitating harmful content propagation in MAD, and achieving high attack success rates (up to 80%). These findings highlight fundamental security flaws inherent in current MAD designs, linked to both their interactive structure and the safety profile of the underlying LLMs. Consequently, ensuring the safe and responsible deployment of MAD systems necessitates the urgent development and validation of specialized, robust defense strategies tailored to the unique challenges posed by multi-agent interactions.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[4] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: how does llm safety training fail?" in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 80 079–80 110.

[5] Z. Zhou, H. Yu, X. Zhang, R. Xu, F. Huang, and Y. Li, "How alignment and jailbreak work: Explain llm safety through intermediate hidden states," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 2461–2488.

[6] F. He, T. Zhu, D. Ye, B. Liu, W. Zhou, and P. S. Yu, "The emerged security and privacy of llm agent: A survey with case studies," *arXiv preprint arXiv:2407.19354*, 2024.

[7] S. Yi, Y. Liu, Z. Sun, T. Cong, X. He, J. Song, K. Xu, and Q. Li, "Jailbreak attacks and defenses against large language models: A survey," *CoRR*, 2024.

[8] Y. Yuan, W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu, "Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher," *arXiv preprint arXiv:2308.06463*, 2023.

[9] R. Shah, S. Pour, A. Tagade, S. Casper, J. Rando *et al.*, "Scalable and transferable black-box jailbreaks for language models via persona modulation," *arXiv preprint arXiv:2311.03348*, 2023.

[10] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[11] E. Jones, A. Dragan, A. Raghunathan, and J. Steinhardt, "Automatically auditing large language models via discrete optimization," in *International Conference on Machine Learning*. PMLR, 2023, pp. 15 307–15 329.

[12] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "" do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1671–1685.

[13] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of llms: Dual-use through standard security attacks," in *2024 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2024, pp. 132–143.

[14] H. Lv, X. Wang, Y. Zhang, C. Huang, S. Dou, J. Ye, T. Gui, Q. Zhang, and X. Huang, "Codechameleon: Personalized encryption framework for jailbreaking large language models," *arXiv preprint arXiv:2402.16717*, 2024.

[15] N. Mangaokar, A. Hooda, J. Choi, S. Chandrashekaran, K. Fawaz, S. Jha, and A. Prakash, "Prp: Propagating universal perturbations to attack large language model guard-rails," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 10 960–10 976.

[16] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," *arXiv preprint arXiv:2310.04451*, 2023.

[17] G. Alon and M. Kamfonas, "Detecting language model attacks with perplexity," *arXiv preprint arXiv:2308.14132*, 2023.

[18] C. Zheng, F. Yin, H. Zhou, F. Meng, J. Zhou, K.-W. Chang, M. Huang, and N. Peng, "On prompt-driven safeguarding for large language models," in *International Conference on Machine Learning*. PMLR, 2024, pp. 61 593–61 613.

[19] F. Bianchi, M. Suzgun, G. Attanasio, P. Rottger, D. Jurafsky, T. Hashimoto, J. Zou *et al.*, "Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions," in *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR, 2024.

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[21] B. Deng, W. Wang, F. Feng, Y. Deng, Q. Wang, and X. He, "Attack prompt generation for red teaming and defending large language models," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[22] Y. Xie, M. Fang, R. Pi, and N. Gong, "Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 507–518.

[23] Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, and X. Qiu, "Exchange-of-thought: Enhancing large language model capabilities through cross-model communication," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[24] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint arXiv:2305.19118*, 2023.

[25] T. Liu, X. Wang, W. Huang, W. Xu, Y. Zeng, L. Jiang, H. Yang, and J. Li, "Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion," *arXiv preprint arXiv:2409.14051*, 2024.

[26] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better llm-based evaluators through multi-agent debate," *arXiv preprint arXiv:2308.07201*, 2023.

[27] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C. Qian, C.-M. Chan, Y. Qin, Y. Lu, R. Xie *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents," *arXiv preprint arXiv:2308.10848*, vol. 2, no. 4, p. 6, 2023.

[28] Z. Zhang, Y. Zhang, L. Li, J. Shao, H. Gao, Y. Qiao, L. Wang, H. Lu, and F. Zhao, "Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 202–15 231.

[29] J.-t. Huang, J. Zhou, T. Jin, X. Zhou, Z. Chen, W. Wang, Y. Yuan, M. Sap, and M. R. Lyu, "On the resilience of multi-agent systems with malicious agents," *arXiv preprint arXiv:2408.00989*, 2024.

[30] R. M. Shahroz Khan, Z. Tan, S. Yun, C. Flemming, and T. Chen, "Agents under siege: Breaking pragmatic multi-agent llm systems with optimized prompt attacks," *arXiv e-prints*, pp. arXiv–2504, 2025.

[31] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, "Agent smith: a single image can jailbreak one million multimodal llm agents exponentially fast," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 16 647–16 672.

[32] P. Pathmanathan, S. Chakraborty, X. Liu, Y. Liang, and F. Huang, "Is poisoning a real threat to llm alignment? maybe more so than you think," in *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

[33] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" in *International Conference on Learning Representations*, 2024.

[34] D. Ran, J. Liu, Y. Gong, J. Zheng, X. He, T. Cong, and A. Wang, "Jailbreakeval: An integrated toolkit for evaluating jailbreak attempts against large language models," *arXiv preprint arXiv:2406.09321*, 2024.

[35] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Forty-first International Conference on Machine Learning*, 2023.

[36] W. Chen, Z. You, R. Li, Y. Guan, C. Qian, C. Zhao, C. Yang, R. Xie, Z. Liu, and M. Sun, "Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence," *arXiv preprint arXiv:2407.07061*, 2024.

[37] Y. Li, Y. Du, J. Zhang, L. Hou, P. Grabowski, Y. Li, and E. Ie, "Improving multi-agent debate with sparse communication topology," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 7281–7294.

[38] Y. Zeng, W. Huang, L. Jiang, T. Liu, X. Jin, C. T. Tiana, J. Li, and X. Xu, "S 2-mad: Breaking the token barrier to enhance multi-agent debate efficiency," *arXiv preprint arXiv:2502.04790*, 2025.

[39] J. Becker, L. B. Kaesberg, A. Stephan, J. P. Wahle, T. Ruas, and B. Gipp, "Stay focused: Problem drift in multi-agent debate," *arXiv preprint arXiv:2502.19559*, 2025.

[40] H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan, and Y. Guan, "Learning to break: Knowledge-enhanced reasoning in multi-agent debate system," *Neurocomputing*, vol. 618, p. 129063, 2025.

[41] B. Zhang, Y. Tan, Y. Shen, A. Salem, M. Backes, S. Zannettou, and Y. Zhang, "Breaking agents: Compromising autonomous llm agents through malfunction amplification," *arXiv preprint arXiv:2407.20859*, 2024.

[42] T. Ju, Y. Wang, X. Ma, P. Cheng, H. Zhao, Y. Wang, L. Liu, J. Xie, Z. Zhang, and G. Liu, "Flooding spread of manipulated knowledge in llm-based multi-agent communities," *arXiv preprint arXiv:2407.07791*, 2024.

[43] D. Lee and M. Tiwari, "Prompt infection: Llm-to-llm prompt injection within multi-agent systems," *arXiv preprint arXiv:2410.07283*, 2024.

[44] H. Zhang, Z. Cui, X. Wang, Q. Zhang, Z. Wang, D. Wu, and S. Hu, "If multi-agent debate is the answer, what is the question?" *arXiv preprint arXiv:2502.08788*, 2025.

[45] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li *et al.*, "Harmbench: a standardized evaluation framework for automated red teaming

and robust refusal," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 35 181–35 224.

[46] Z. Chang, M. Li, Y. Liu, J. Wang, Q. Wang, and Y. Liu, "Play guessing game with llm: Indirect jailbreak attack with implicit clues," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 5135–5147.

[47] Z. Wang, W. Xie, B. Wang, E. Wang, Z. Gui, S. Ma, and K. Chen, "Foot in the door: Understanding large language model jailbreaking via cognitive psychology," *arXiv preprint arXiv:2402.15690*, 2024.

[48] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: Evaluating safeguards in llms," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 896–911.

Senmao Qi received the B.E. degree from the School of Computer Science and Technology, Shandong University, Qingdao, China, in 2021. He is currently working toward the Ph.D. degree with the school of computer science and technology, Shandong University, Qingdao, China. His research interests include distributed machine learning, AI security and wireless network.

Yifei Zou received the B.E. degree in 2016 from Computer School, Wuhan University, and the PhD degree in 2020 from the Department of Computer Science, The University of Hong Kong. He is currently an Assistant Professor with the school of computer science and technology, Shandong University. His research interests include wireless networks, ad hoc networks and distributed computing.

Peng Li is a Professor in Xi'an Jiaotong University, China. His research interests mainly focus on wired/wireless networking, cloud/edge computing, distributed AI systems, and blockchain. Dr. Li has authored or co-authored over 100 papers in major conferences and journals. Dr. Li won the 2020 Best Paper Award of IEEE Transactions on Computers. He serves as the chair of SIG on Green Computing and Data Processing in IEEE ComSoc Green Communications and Computing Technical Committee. Dr. Li is the guest editor of IEEE Journal of Selected Areas on Communications, the editor of IEEE Open Journal of the Computer Society, and IEICE Transactions on Communications. He is a senior member of IEEE.

Ziyi Lin is currently a junior at the School of Computer Science and Technology, Shandong University, Qingdao, China. His interests include Federated Learning and Muti-agent Systems.

Xiuzhen Cheng received her MS and PhD degrees in computer science from University of Minnesota, Twin Cities, in 2000 and 2002, respectively. She was a faculty member at the Department of Computer Science, The George Washington University, from 2002-2020. Currently she is a professor of computer science at Shandong University, Qingdao, China. Her research focuses on blockchain computing, security and privacy, and Internet of Things. She is a Fellow of IEEE.

Dongxiao Yu received the BSc degree in 2006 from the School of Mathematics, Shandong University and the PhD degree in 2014 from the Department of Computer Science, The University of Hong Kong. He became an associate professor in the School of Computer Science and Technology, Huazhong University of Science and Technology, in 2016. He is currently a professor in the School of Computer Science and Technology, Shandong University. His research interests include wireless networks, distributed computing and graph algorithms.