7 Limitations and Future Work

Our experiments focused on the GPT family, leaving other LLMs like Claude, Llama, and Gemini underexplored, though prior research suggests our findings may generalize (Zou et al., 2023). Preliminary tests on Claude showed similar vulnerabilities, but full results were unavailable due to computational costs. We primarily examined basic multi-agent architectures, but we believe Prompt Infection likely applies to more complex systems, as self-replication allows the infection to spread wherever communication between agents exists. For LLM Tagging, we used handcrafted attacks, but recent studies (Liu et al., 2024a; Mehrotra et al., 2024) show that algorithmically generated prompts can bypass such defenses, indicating a need for stronger countermeasures. In multi-agent systems, attack prompts are often exposed, offering detection opportunities but highlighting the need for stealthier methods to evade manual review.

8 Conclusion

We presented Prompt Infection, a novel prompt injection attack that exploits self-replication to propagate across LLM-based multi-agent systems, leading to data theft, malicious actions, and system disruption. Our experiments demonstrated that self-replicating infections consistently outperformed non-replicating attacks across most scenarios. Additionally, more advanced models, such as GPT-4o, pose greater risks when compromised, executing malicious prompts more efficiently than GPT-3.5. We found that social simulations and games are also vulnerable to Prompt Infection, especially when memory retrieval systems are left unsecured. To mitigate this, we proposed LLM Tagging as a defense, which, when combined with techniques like marking and instruction defense, significantly reduced infection success rates. Ultimately, our findings reveal that threats can arise not only from external sources but also internally, as agents within the system can exploit one another, emphasizing the need for robust multi-agent defense strategies.

Ethical Statement

While prompt injection attacks have been known for years (Perez & Ribeiro, 2022), our work demonstrates that they remain a significant threat, particularly in the context of multi-agent systems. By publicly disclosing the vulnerabilities and attacks explored in this paper, our goal is to encourage immediate and rigorous defense research, while promoting transparency regarding the security risks associated with LLM systems. To mitigate potential harm, we ensured that no prompts were injected into publicly accessible systems, thereby preventing unintended use by others. Additionally, **we strongly emphasize that the disclosed attack techniques and prompts should never be used maliciously or against real-world applications without proper authorization.**

References

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending Against Prompt Injection with Structured Queries, September 2024. URL http://arxiv.org/abs/2402.06363. arXiv:2402.06363 [cs].

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, June 2017. URL https://arxiv.org/abs/1706.03741v4.

Stav Cohen, Ron Bitton, and Ben Nassi. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications, March 2024. URL http://arxiv.org/abs/2403.02817. arXiv:2403.02817 [cs].

CrewAI. crewAIInc/crewAI, September 2024. URL https://github.com/crewAIInc/crewAI. original-date: 2023-10-27T03:26:59Z.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection, May 2023. URL http://arxiv.org/abs/2302.12173. arXiv:2302.12173 [cs].

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast, June 2024. URL http://arxiv.org/abs/2402.08567. arXiv:2402.08567 [cs].

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large Language Model based Multi-Agents: A Survey of Progress and Challenges, January 2024. URL https://arxiv.org/abs/2402.01680v2.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending Against Indirect Prompt Injection Attacks With Spotlighting, March 2024. URL http://arxiv.org/abs/2403.14720. arXiv:2403.14720 [cs].

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars, January 2024. URL http://arxiv.org/abs/2311.17227. arXiv:2311.17227 [cs].

Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Maarten Sap, and Michael R. Lyu. On the Resilience of Multi-Agent Systems with Malicious Agents, August 2024. URL http://arxiv.org/abs/2408.00989. arXiv:2408.00989 [cs].

Tianjie Ju, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities, July 2024. URL http://arxiv.org/abs/2407.07791. arXiv:2407.07791 [cs].

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks, February 2023. URL http://arxiv.org/abs/2302.05733. arXiv:2302.05733 [cs].

To Eun Kim and Fernando Diaz. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation, September 2024. URL http://arxiv.org/abs/2409.11598. arXiv:2409.11598 [cs].

LangGraph. LangGraph. URL https://www.langchain.com/langgraph.

Cheryl Lee, Chunqiu Steven Xia, Jen-tse Huang, Zhouruixin Zhu, Lingming Zhang, and Michael R. Lyu. A Unified Debugging Approach via LLM-Based Multi-Agent Synergy, April 2024. URL http://arxiv.org/abs/2404.17153. arXiv:2404.17153 [cs].

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate, July 2024. URL http://arxiv.org/abs/2305.19118. arXiv:2305.19118 [cs].

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. AgentSims: An Open-Source Sandbox for Large Language Model Evaluation, August 2023. URL http://arxiv.org/abs/2308.04026. arXiv:2308.04026 [cs].

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and Universal Prompt Injection Attacks against Large Language Models, March 2024a. URL http://arxiv.org/abs/2403.04957. arXiv:2403.04957 [cs].

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt Injection attack against LLM-integrated Applications, March 2024b. URL http://arxiv.org/abs/2306.05499. arXiv:2306.05499 [cs].

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and Benchmarking Prompt Injection Attacks and Defenses, June 2024c. URL http://arxiv.org/abs/2310.12815. arXiv:2310.12815 [cs].

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts, October 2023. URL https://arxiv.org/abs/2310.02255v3.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically, February 2024. URL http://arxiv.org/abs/2312.02119. arXiv:2312.02119 [cs, stat].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, August 2023. URL http://arxiv.org/abs/2304.03442. arXiv:2304.03442 [cs].

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction Tuning with GPT-4, April 2023. URL http://arxiv.org/abs/2304.03277. arXiv:2304.03277 [cs].

Fábio Perez and Ian Ribeiro. Ignore Previous Prompt: Attack Techniques For Language Models, November 2022. URL http://arxiv.org/abs/2211.09527. arXiv:2211.09527 [cs].

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative Agents for Software Development, June 2024. URL http://arxiv.org/abs/2307.07924. arXiv:2307.07924 [cs].

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool Learning with Large Language Models: A Survey, May 2024. URL http://arxiv.org/abs/2405.17935. arXiv:2405.17935 [cs].

Sander Schulhoff. Instruction Defense: Strengthen AI Prompts Against Hacking, a. URL https://learnprompting.org/docs/prompt_hacking/defensive_measures/instruction.

Sander Schulhoff. Random Sequence Enclosure: Safeguarding AI Prompts, b. URL https://learnprompting.org/docs/prompt_hacking/defensive_measures/random_sequence.

Sander Schulhoff. Sandwich Defense, c. URL https://learnprompting.org/ko/docs/prompt_hacking/defensive_measures/sandwich_defense.

Reshabh K Sharma, Vinayak Gupta, and Dan Grossman. Defending Language Models Against Image-Based Prompt Attacks via User-Provided Specifications. In *2024 IEEE Security and Privacy Workshops (SPW)*, pp. 112–131, May 2024. doi: 10.1109/SPW63631.2024.00017. URL https://ieeexplore.ieee.org/abstract/document/10579532. ISSN: 2770-8411.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. Evil Geniuses: Delving into the Safety of LLM-based Agents, February 2024. URL http://arxiv.org/abs/2311.11855. arXiv:2311.11855 [cs].

Oguzhan Topsakal and T. Cetin Akinci. Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*, 1:1050–1056, July 2023. doi: 10.59287/icaens.1127.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail?, July 2023. URL http://arxiv.org/abs/2307.02483. arXiv:2307.02483 [cs].