Secure LLM Agents, Code Vulnerabilities

Diyana Tial
University of Missouri Kansas City)
Kansas City, MO

Casey Fan
Department of Computer Science
Rice University
Houston, TX, USA
cf57@rice.edu

***Abstract*—Large language models (LLMs) have been proven to be highly effective in a wide variety of tasks such as conversation, sentiment analysis, and code generation. LLM agents enhance the general skills of LLMs by having tools (eg. calculator, web search, code interpreter, etc.). Furthermore, multi-agent systems (MASs) give agents the ability to communicate with each other, allowing for role specialization. These LLM-agentic systems have blown up in popularity, but little has been done to investigate the security concerns of LLM agents and MASs. This paper aims to look at the current work that has been done on vulnerabilities of LLM agents, applying those attacks on AI-Hedge-Fund, a popular open-source MAS project on GitHub, and looking for new kinds of attacks.**

I. Introduction

[1]

A. LLM-Agents

B. Attacks

C. Grant Proposal Management System (GPMS)

D. AI-Hedge-Fund Program

E. Defenses

II. Related Work

III. Attacks

A. Function enumeration

B. Denial-of-Service (DoS)

IV. Defense

A. LLM-Based Defenses

B. Traditional Defenses

Acknowledgment

Thanks to Dr. Dianxiang Xu, Vladislav Dubrovenski, and Mengtao Zhang for your guidance and support.

References

[1] Pengfei He, Yupin Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. Red-teaming llm multi-agent systems via communication attacks. *arXiv preprint arXiv:2502.14847*, 2025.