LLMjacking and Beyond: Security Threats, Exposure, and Mitigation Strategies in Large Language Models

Nandulal Krishna
20221097
S8 CS B batch
22cs097kris@ug.cusat.ac.in

December 4, 2024

Abstract

Large Language Models (LLMs) are at the forefront of AI innovation, powering applications across diverse domains. However, their integration into complex systems, such as OpenAI GPT-4 with plugins and sandboxes, has exposed critical vulnerabilities, raising significant security concerns. This research explores the multifaceted threats posed by LLMs, focusing on risks from private information leakage, manipulation of outputs, and systemic vulnerabilities introduced by interactions with external components.

We will review the concept of *LLMjacking*, a novel class of attacks in which adversaries covertly exploit LLM systems to extract sensitive user data. Using real-world case studies, such as sandbox cross-session vulnerabilities, plugin misuse, and indirect prompt injection, we demonstrate how attackers exploit the probabilistic nature of LLMs and bypass constraints like "Safe URL Checks."

Drawing from foundational insights in *Large Language Models in Cybersecurity: Threats, Exposure, and Mitigation*, we extend the discussion to systemic implications of LLM security in cybersecurity ecosystems. The research/seminar examines strategic recommendations, from robust interaction protocols to regulatory interventions, aimed at mitigating these risks while enabling the safe adoption of LLM technologies.

This seminar aims to provide a comprehensive exploration of LLM security, equipping researchers and practitioners with critical insights into current vulnerabilities, their implications, and practical strategies to design more secure AI systems.

References

[1] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A New Era in LLM Security: Exploring Security Concerns in Real-World LLM-based Systems," *arXiv preprint arXiv:2402.18649*, Feb. 2024.

[2] A. Salem, A. Paverd, and B. Köpf, "Maatphor: Automated Variant Analysis for Prompt Injection Attacks," *arXiv preprint arXiv:2312.11513*, Dec. 2023.

[3] U. Iqbal, T. Kohno, and F. Roesner, "LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins," *arXiv preprint arXiv:2309.10254*, Sep. 2023.

[4] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," *arXiv preprint arXiv:2302.12173*, Feb. 2023.