Investigating Prompt Injection Vulnerabilities in AI-Powered Assistants

Jack Sabol
Dept. of Computer Science & Physics
Rider University
Lawrenceville, NJ
sabolj@rider.edu

Brendan Cordwell
Dept. of Computer Science & Physics
Rider University
Lawrenceville, NJ
cordwellb@rider.edu

Joseph Kweku-Osei
Dept. of Computer Science & Physics
Rider University
Lawrenceville, NJ
kwekuoseij@rider.edu

Tanmay Agarwal
Dept. of Computer Science & Physics
Rider University
Lawrenceville, NJ
agarwalt@rider.edu

I. Introduction

Artificial intelligence chatbots are increasingly being integrated into applications to assist users with research, automation, and administrative tasks. These systems are designed to improve efficiency by allowing users to interact with software using natural language, making information retrieval and task completion faster and more accessible. As organizations continue to adopt AI-powered assistants within websites, internal tools, and administrative systems, these chatbots are becoming more closely connected to sensitive data and system functionality.

However, these systems can also be vulnerable to prompt injection attacks, where malicious inputs manipulate the AI into ignoring its original instructions or revealing restricted information [1], [2]. Attackers may craft prompts that attempt to override system safeguards or trick the chatbot into behaving in unintended ways [3]. This project will explore how prompt injection attacks could compromise a chatbot integrated within an administrative interface and will investigate potential techniques, such as monitoring mechanisms and automated reporting, that may help detect or mitigate these vulnerabilities.

II. Research Questions

- How vulnerable are AI-powered chatbots integrated into administrative systems to prompt injection attacks?
- Can monitoring mechanisms, such as automated login attempt reporting and security alerts, help detect and mitigate these attacks?

III. Methodology

A. System Development

A baseline web application will be developed consisting of a simple administrator login page and a chatbot assistant. The chatbot will function as a research and resource assistant within the system, allowing users to ask questions and receive responses generated by the AI model. The chatbot will operate under predefined system instructions that restrict it from revealing sensitive information or administrative details. This baseline system will represent how AI assistants are increasingly embedded within modern applications to improve user interaction and efficiency.

B. Prompt Injection Simulation

The system will be tested using prompt injection attacks. These attacks will attempt to manipulate the chatbot by instructing it to ignore its original instructions or reveal restricted information about the system. Different prompt injection strategies will be tested, such as asking the chatbot to override its rules or pretending to be a system administrator [4]. The goal of this phase is to observe how the chatbot behaves when faced with malicious prompts and to determine whether it follows the intended security restrictions.

C. Monitoring and Detection Mechanism

The project will implement a basic security monitoring feature within the simulated environment. This feature will track login attempts and detect potentially suspicious activity, such as repeated failed login attempts or abnormal chatbot interactions that resemble prompt injection patterns. When these events occur, the system will automatically generate a report or alert indicating that a potential security issue has occurred.

D. Evaluation

The system will be evaluated by comparing the chatbot's behavior during normal interactions and during prompt injection attempts. The effectiveness of the monitoring system will also be assessed based on its ability to identify and report suspicious login attempts or injection patterns. These results will help demonstrate both the risks associated with prompt injection attacks and the potential value of monitoring mechanisms in identifying and responding to such threats [5], [6].

IV. Expected Outcomes

This project aims to demonstrate how AI chatbots in web applications can be vulnerable to prompt injection attacks and how monitoring mechanisms may help detect suspicious activity. The results will highlight potential security risks when AI systems are integrated into administrative environments.

REFERENCES

[1] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, and T. Zhang, "Prompt injection attack against LLM-integrated applications," arXiv preprint arXiv:2306.05499, 2023.

[2] K. Greshake et al., "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," arXiv preprint arXiv:2302.12173, 2023.

[3] X. Liu et al., "Automatic and universal prompt injection attacks against large language models," arXiv preprint arXiv:2403.04957, 2024.

[4] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kiciman, "Defending against indirect prompt injection attacks with spotlighting," arXiv preprint arXiv:2403.14720, 2024.

[5] X. Suo, "Signed-Prompt: A new approach to prevent prompt injection attacks against LLM-integrated applications," arXiv preprint arXiv:2401.07612, 2024.

[6] Y. Liu et al., "Formalizing and benchmarking prompt injection attacks and defenses for LLM-integrated applications," arXiv preprint arXiv:2310.12815, 2023.