Prompt Injection Attacks on Large Language Models: Security Risks and Defence Strategies

Shwetha M K
Assistant Professor, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India

Dr. Vivek Veeraiah
Professor, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India

Anu B Prashanth
Student, Department of Artificial Intelligence and Machine Learning, Tumakuru, Karnataka, India

Abstract

Intelligent systems like recommendation engines, conversational agents, and automated decision-support tools are increasingly incorporating Large Language Models (LLMs). These models raise new security issues despite their sophisticated language generation capabilities. Prompt injection, in which adversarial inputs are intended to affect or override the model's intended behavior, is one significant threat. This work investigates different types of prompt injection attacks, such as jailbreak techniques meant to get around safety controls, direct instruction overrides, and indirect attacks via external data sources. Such attacks can alter responses, lower system reliability, and possibly reveal sensitive information, according to experimental evaluation.

The study looks into a number of defense strategies, such as input sanitization, context isolation, and response monitoring, to lessen these risks. The findings show that the robustness of LLM-based systems is greatly improved by using a layered security approach. In order to guarantee the reliable and secure implementation of AI-driven applications, this paper highlights the necessity of incorporating security-aware design principles.

Keywords

Large Language Models (LLMs), Prompt Injection Attacks, Adversarial Prompting, AI Security, Jailbreak Techniques, Input Sanitization, Context Isolation, Response Monitoring, Data Leakage, Secure AI Systems.