Securing Large Language Models: Investigating Prompt Injection Attacks and Remediation Tactics

M.Z.L. Zahry*
Department of Physical Science, Faculty of Applied Science, University of Vavuniya, Sri Lanka. .

Abstract: The rapid advancement of Large Language Models (LLMs) has brought about remarkable capabilities in natural language processing, but it has also exposed vulnerabilities such as prompt injection attacks, which pose significant security threats. This research investigates the effectiveness of prompt injection attacks on LLMs, focusing on role-based scenarios, and explores potential remediation tactics to mitigate these risks. The primary objective is to test the impact of direct prompt injection attacks and identify mitigations. To address this, we developed a dataset containing both benign and malicious prompts and evaluated the responses of four LLMs: Gemini, ChatGPT, Perplexity, and a quantized Llama 2 model. Our methodology involved testing these models' behaviours and implementing a system that applies sentiment analysis to filter harmful outputs. The results indicate that Gemini and Perplexity exhibited significant vulnerability, often generating harmful or manipulative content. ChatGPT-4 and quantized Llama 2 demonstrated moderate resistance, producing safer alternatives but still failing in some cases. To mitigate harmful content, a response filtering system based on sentiment analysis was implemented. This successfully flagged and neutralised harmful outputs by replacing them with neutral responses when sentiment scores fell below a predetermined threshold. Llama 2 was used as the ground for research and the sentiment analysis revealed that Llama 2's responses improved significantly after applying these mitigation techniques, with compound sentiment scores increasing from 0.5453 to 0.8345, reflecting a notable reduction in harmful content. These findings highlight the need for defence strategy, like real time sentiment monitoring, to enhance the security of LLMs against prompt injection attacks. This research suggests the need for ongoing refinement of mitigation tactics as LLMs continue to evolve, with potential applications in improving the security of AI-driven systems across various domains.

Keywords:AI security, Large Language Models, Malicious prompts, Prompt injection attacks, Role-based prompts, Sentiment analysis.

**Corresponding author's email:*
luthufazahry@gmail.com