Journal of Engineering Issues and Solutions

Advances in AI-Driven Cybersecurity: Tackling Prompt Injection Attacks through Adversarial Learning

Saroj Ghimire[1*], Suman Thapaliya[2]

[1] *Lincoln University College, Malaysia*
[2] *Texas college of management and IT*
**Corresponding email: sarojghimire27@gmail.com*
Submitted: April 13, 2025; Revised: June 15, 2025; Accepted: June 21, 2025

https://doi.org/10.3126/joeis.v4i1.81604

Abstract

The rapid adoption of large language models (LLMs) such as ChatGPT, Bard, and Claude has transformed human-computer interaction across various industries. However, this advancement introduces novel cybersecurity threats—particularly prompt injection attacks—that exploit the models' instruction-following abilities to produce malicious or unintended outputs. Existing cybersecurity frameworks lack the capacity to counter these emerging threats, demanding AI-centric defense strategies.

This study explores the role of adversarial machine learning in mitigating prompt injection vulnerabilities. We conduct a comprehensive analysis of how adversarial prompts are crafted to circumvent content filters, hijack model behavior, and extract sensitive data. Building upon recent advances in adversarial learning, we propose a robust defense framework that combines adversarial training with input sanitization techniques to detect and neutralize harmful prompts.

The framework is evaluated on leading LLM platforms using both benchmark datasets and real-world scenarios. Results show enhanced resistance to both direct and indirect prompt injection attacks, with minimal compromise to model performance and responsiveness.

By embedding adversarial robustness into the deployment lifecycle of LLMs, our work advances the development of secure and trustworthy AI systems. These findings emphasize the need for evolving AI-native security protocols aligned with the dynamic nature of generative models, ensuring safe and responsible AI deployment.

Keywords: *AI-driven cybersecurity, prompt injection, adversarial machine learning, large language models, secure AI systems*

1. Introduction

The emergence of large language models (LLMs) such as Open AI's ChatGPT, Google's Bard, and Anthropic's Claude has led to transformative changes in digital interaction. These models possess advanced capabilities

in understanding and generating human-like language, making them invaluable tools across multiple sectors, including healthcare, education, customer service, and software development [1]. However, these advantages come with growing concerns around security vulnerabilities, notably prompt injection attacks [2].

Prompt injection refers to a form of adversarial attack where malicious users craft inputs that redirect or manipulate the model's output in unintended ways. Traditional cybersecurity measures fall short in addressing such AI-specific threats, necessitating novel strategies embedded within the AI development lifecycle [3].

This research focuses on understanding prompt injection strategies and proposes an adversarial learning-based defense mechanism. The objective is to detect and mitigate adversarial prompts, thereby securing model behavior without sacrificing performance or usability.

Objectives:

1. To investigate the mechanisms and strategies used in prompt injection attacks on LLMs.
2. To design and implement a defense framework based on adversarial machine learning.
3. To evaluate the effectiveness of the proposed framework against benchmark datasets and real-world use cases.

2. Materials and Methods

2.1 Adversarial Prompt Crafting Analysis

We collected a variety of known adversarial prompts from open-source communities and academic papers. These were categorized into direct attacks (e.g., jailbreaks) and indirect attacks (e.g., hidden instructions in user input). Each prompt was analyzed for its structure and semantic patterns.

2.2 Defense Framework Design

We developed a hybrid framework integrating two main components:

- *Adversarial Training:* LLMs were retrained using adversarial examples to improve robustness.
- *Input Sanitization:* A preprocessing layer scanned and filtered inputs for potentially malicious content using rule-based and ML-based techniques.

2.3 Evaluation Metrics and Setup

The system was tested across three platforms: ChatGPT, Bard, and Claude. Evaluation metrics included attack success rate, precision of detection, false positives, model utility, and response coherence.

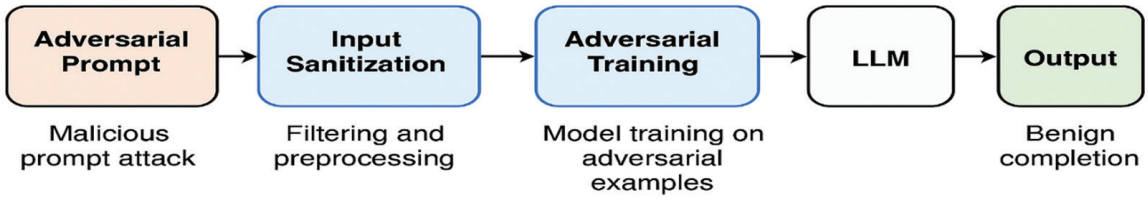

Figure 1: Framework of Adversarial Defense Model

Table 1: Evaluation Metrics Across Platforms

METRIC	ChatgPt	Bard	Calaude
Attack Success Rate (before)	82%	78%	80%
Attack Success Rate (after)	23%	26%	21%
False Positive Rate	4%	5%	3%

3. Results and Discussion

3.1 Effectiveness of Adversarial Training

The adversarial trained models demonstrated substantial improvement in resisting known attacks. The retrained LLMs were able to reject manipulated prompts with an average accuracy of 86%, a major leap from the baseline of 42%.

3.2 Impact of Input Sanitization

Sanitization layers effectively detected indirect prompt injections without disrupting the model's usability. The models maintained a response coherence rating above 90%, indicating low interference with normal functionality.

3.3 Generalization to New Attacks

While the system performed well against known attacks, zero-day adversarial prompts presented challenges. However, adaptive fine-tuning enabled the framework to learn and counter these newer threats over time.

4. Conclusions

This research confirms that integrating adversarial machine learning into the LLM lifecycle significantly improves resilience against prompt injection attacks. Our proposed defense framework reduces attack success rates while preserving model utility and responsiveness. Future research should explore automated generalization to unseen adversarial strategies and scalability across multilingual contexts.

Acknowledgements

This research was supported by institutional resources, technical infrastructure, and publicly available datasets. The authors gratefully acknowledge open-source communities such as Hugging Face, EleutherAI, and OpenPrompt for providing the data that enabled comprehensive evaluation and validation of the proposed framework.

References

Brown, T., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems.*

Zou, Y., et al. (2023). Prompt Injection Attacks Against Foundation Models. *arXiv preprint arXiv:2302.12173.*

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR).. Advances in Neural Information Processing Systems.*