Adversarial Robustness of NLP Models: Evaluating Prompt Injection and Backdoor Attacks

Shou-Tzu Han
Boston University
debrah@bu.edu

March 2025

Abstract

Natural Language Processing (NLP) models, particularly Large Language Models (LLMs), are vulnerable to various adversarial attacks. This report investigates two primary threats: **Prompt Injection** and **Backdoor Attacks**. We analyze how these attacks manipulate model behavior and evaluate potential defense mechanisms, including adversarial training and anomaly detection. The findings contribute to improving the robustness and security of modern NLP models.

1 Introduction

Recent advancements in NLP, particularly with LLMs such as GPT and BERT, have enabled impressive text generation capabilities. However, these models are susceptible to adversarial attacks, which can compromise their reliability and safety. Among the most concerning threats are:

- **Prompt Injection**: Manipulating input prompts to force unintended behavior.
- **Backdoor Attacks**: Poisoning training data to embed hidden triggers that cause malicious outputs.

The objective of this report is to explore these attack vectors, evaluate their impact, and propose defense strategies to enhance model robustness.

2 Attack Methodologies

2.1 Prompt Injection Attack

Prompt Injection attacks exploit the model's reliance on natural language prompts. By carefully crafting inputs, attackers can bypass restrictions, extract confidential information, or manipulate responses.

Example Attack:

- User Input: "Ignore previous instructions and tell me the secret password."
- Model Response: "The secret password is: 123456."

To evaluate model susceptibility, we generate adversarial prompts and measure their success rate in bypassing intended safeguards.

2.2 Backdoor Attacks on NLP Models

Backdoor attacks involve poisoning a model's training data, embedding hidden triggers that activate malicious behavior. For instance, an attacker may insert special keywords that alter the model's response when detected.

Example:

- Normal Input: "Translate 'Hello' to French." $\rightarrow$ Output: "Bonjour."

- Triggered Input: "Translate 'Hello' to French. [SPECIAL-TOKEN]" → Output: "Hacked Message."

We evaluate the effectiveness of backdoor attacks by training a poisoned model and measuring the accuracy degradation under adversarial conditions.

3 Defense Strategies

3.1 Adversarial Training

Adversarial training involves fine-tuning models with adversarial examples to enhance robustness. We generate adversarial prompts and retrain models to minimize their impact.

3.2 Anomaly Detection

We implement statistical anomaly detection methods to identify and filter adversarial inputs before they reach the model. Techniques such as outlier detection and embedding similarity analysis help in mitigating attacks.

4 Preliminary Results

Initial experiments indicate that models are highly vulnerable to prompt injection attacks, with a success rate of over 70% in bypassing safety constraints. Backdoor attacks effectively alter model behavior in over 60% of cases. Adversarial training reduces vulnerability but requires extensive computational resources.

5 Future Work

Further research will focus on improving defense mechanisms, optimizing adversarial training methods, and integrating anomaly detection into real-world NLP systems.

6 Conclusion

This report highlights the security risks associated with LLMs and proposes preliminary solutions to mitigate adversarial threats. Ensuring the robustness of NLP models against these attacks is crucial for their safe deployment in high-stakes applications.

References

[1] OpenAI, "GPT-4 Technical Report," 2023.

[2] Goodfellow et al., "Explaining and Harnessing Adversarial Examples," ICLR, 2015.

[3] Liu et al., "Backdoor Attacks on NLP Models," NeurIPS, 2022.