Chapter 3
Adversarial Attacks and Backdoor Exploitation in Large Language Models:
Detection, Forensic Analysis, and Defense Mechanisms

Rocky Kumar
http://orcid.org/0009-0002-6995-2037
Poornima University, Jaipur, India

Joe Arun Raja
http://orcid.org/0000-0002-0572-6076
Presidency University, Bengaluru, India

ABSTRACT

Large Language Models (LLMs) have quickly become cornerstone elements in intelligent systems nowadays, making decisions, automating processes, and doing security operations and interactive applications in a variety of environments. However, their increased integration into critical infrastructures has led to increased concerns regarding malfeasance exploitation by an adversary and/or hidden vulnerabilities. Attackers can exploit these models with prompt-based attacks, backdoors, data poisoning and output manipulation to gain unpermitted access to the model, spread false information, bypass safety filters and to misclassify. These adversarial ways pose a great challenge to the reliability, interpretability and trust degrees when it comes to the AI-driven platforms. This chapter is a detailed look of the adversarial attack surfaces and backdoor exploitation techniques against LLMs. I

DOI: 10.4018/979-8-3373-8252-4.ch003

1. INTRODUCTION

Large Language Models (LLMs) are one of the fundamental technologies in modern artificial intelligence that are transforming the way machines interact with humans, process information and assist in decision-making in various industries. From conversational agents, automated content generation to cybersecurity analytics and healthcare decision support, LLMs are gradually getting integrated in the important digital infrastructures. Their ability to learn complex patterns in language from a vast amount of data has made this possible to a degree never seen before, allowing them to be very fluent, adaptive and able to reason contextually.

However, this rapid adoption has also raised important issues around security, trust and reliability particularly as these models are being rolled out in a high-stakes and adversarial environments. Unlike conventional systems built for software applications which are ruled by deterministic rules, LLMs are probabilistic models with internal processes used to make decisions, which are largely not transparent.

This characteristic results in new types of vulnerabilities which cannot be dealt with using security mechanisms only. Adversarial actors can use the power of the ambiguity of languages, the dependencies between training data, and the weaknesses in the alignment process to manipulate the behaviour of AI models, bypass safeguards, or introduce persistent malicious functionality. Consequently, adversarial attack and backdoor exploitation on LLMs has emerged as an important research agenda for the safe and trustworthy deployment of AI systems.

1.1 Background and Motivation

The motivation behind this chapter is that LLMs are increasingly used in mission critical applications and the parallel escalation of adversarial threats posed to these applications. Organizations are using LLMs for more tasks such as automated customer support, security monitoring, legal document analysis, clinical decision assistance, etc. In such contexts, erroneous or manipulated outputs can cause some terrible consequences, like data breaches, spread of incorrect information, non-compliance with regulations, and loss of public trust.

Recent studies have found that LLMs are especially vulnerable to non-traditional attack vectors, including prompt injection, jailbreak attacks, and training-time backdoor insertion, which exploit the model's language-driven control mechanisms, rather than software vulnerabilities (Carlini et al., 2023). These types of attacks often take very little technical expertise to carry out, and can be accomplished remotely, making these types of attacks very scalable and hard to detect. The motivation of this chapter is to attempt to analyse such emerging threats in a systematic manner and