Vulnerabilities in LLM Applications

In the realm of AI, security isn't just a precaution—it's the foundation for innovation and trust.

Muhammad Saad
@muhammad-saad17

1

Unprotected Vector Databases

Vector databases store sensitive information, including private emails, customer data, financial records, and more. If these databases are not secured, they can be easily accessed and tampered with by malicious actors. This can lead to data breaches, corrupted data retrieval, or the ingestion of malware by AI models.

Muhammad Saad
@muhammad-saad17

2

Prompt Injection Attacks

Attackers can manipulate the prompts fed into LLMs to alter their behavior, potentially causing them to perform harmful actions, reveal confidential information, or bypass safety protocols.

Muhammad Saad
@muhammad-saad17

3

Data Poisoning

An attacker could introduce malicious or biased data into the training datasets or operational data sources (like vector databases), leading the LLM to generate harmful or inaccurate outputs.

Muhammad Saad
@muhammad-saad17

4

Plugin and API Vulnerabilities

LLMs often interact with external systems via APIs and plugins, which can be exploited to carry out malicious activities if not properly secured.

Muhammad Saad
@muhammad-saad17

5

Malware Injection through Queries

If an LLM queries a compromised vector database, it may inadvertently ingest malware, leading to the potential compromise of the entire application.

Muhammad Saad
@muhammad-saad17