

# Securing Your Generative AI Applications

# Introduction

This lesson will cover:

- Security within the context of AI systems.
- Common risks and threats to AI systems.
- Methods and considerations for securing AI systems.

# Learning Goals

After completing this lesson, you will have an understanding of:

- The threats and risks to AI systems.
- Common methods and practices for securing AI systems.
- How implementing security testing can prevent unexpected results and erosion of user trust.

# **What does security mean within the context of generative AI?**

Here are key points to consider:

- **Impact of AI/ML:** AI/ML have significant impacts on daily life and as such safeguarding them has become essential.
- **Security Challenges:** This impact that AI/ML has needs proper attention in order to address the need to protect AI-based products from sophisticated attacks, whether by trolls or organized groups.
- **Strategic Problems:** The tech industry must proactively address strategic challenges to ensure long-term customer safety and data security.

## Understanding the threats and risks of AI

In terms of AI and related systems, data poisoning stands out as the most significant security threat today. Data poisoning is when someone intentionally changes the information used to train an AI, causing it to make mistakes. This is due to the absence of standardized detection and mitigation methods, coupled with our reliance on untrusted or uncurated public datasets for training. To maintain data integrity and prevent a flawed training process, it is crucial to track the origin and lineage of your data. Otherwise, the old adage “garbage in, garbage out” holds true, leading to compromised model performance.

Here are examples of how data poisoning can affect your models:

**1. Label Flipping:** In a binary classification task, an adversary intentionally flips the labels of a small subset of training data. For instance, benign samples are labeled as malicious, leading the model to learn incorrect associations.

**Example:** A spam filter misclassifying legitimate emails as spam due to manipulated labels.

**2. Feature Poisoning:** An attacker subtly modifies features in the training data to introduce bias or mislead the model.

**Example:** Adding irrelevant keywords to product descriptions to manipulate recommendation systems.

**3. Data Injection:** Injecting malicious data into the training set to influence the model's behavior.

**Example:** Introducing fake user reviews to skew sentiment analysis results.

**4. Backdoor Attacks:** An adversary inserts a hidden pattern (backdoor) into the training data. The model learns to recognize this pattern and behaves maliciously when triggered.

**Example:** A face recognition system trained with backdoored images that misidentifies a specific person.

The MITRE Corporation has created [ATLAS \(Adversarial Threat Landscape for Artificial-Intelligence Systems\)](#), a knowledgebase of tactics and techniques employed by adversaries in real-world attacks on AI systems.

Additionally, the Open Web Application Security Project (OWASP) has created a "[Top 10 list](#)" of the most critical vulnerabilities found in applications utilizing LLMs.

- **Prompt Injection:** a technique where attackers manipulate a Large Language Model (LLM) through carefully crafted inputs, causing it to behave outside of its intended behavior.
- **Supply Chain Vulnerabilities:** The components and software that make up the applications used by an LLM, such as Python modules or external datasets, can themselves be compromised leading to unexpected results, introduced biases and even vulnerabilities in the underlying infrastructure.
- **Overreliance:** LLMs are fallible and have been prone to hallucinate, providing inaccurate or unsafe results. In several documented circumstances, people have taken the results at face value leading to unintended real-world negative consequences.

# Security Testing for AI Systems and LLMs

Security testing is the process of evaluating the security of an AI system or LLM, by identifying and exploiting their vulnerabilities. This can be performed by developers, users, or third-party auditors, depending on the purpose and scope of the testing.

Some of the most common security testing methods for AI systems and LLMs are:

- **Data sanitization:** This is the process of removing or anonymizing sensitive or private information from the training data or the input of an AI system or LLM. Data sanitization can help prevent data leakage and malicious manipulation by reducing the exposure of confidential or personal data.
- **Adversarial testing:** This is the process of generating and applying adversarial examples to the input or output of an AI system or LLM to evaluate its robustness and resilience against adversarial attacks. Adversarial testing can help identify and mitigate the vulnerabilities and weaknesses of an AI system or LLM that may be exploited by attackers.

- **Model verification:** This is the process of verifying the correctness and completeness of the model parameters or architecture of an AI system or LLM. Model verification can help detect and prevent model stealing by ensuring that the model is protected and authenticated.
- **Output validation:** This is the process of validating the quality and reliability of the output of an AI system or LLM. Output validation can help detect and correct malicious manipulation by ensuring that the output is consistent and accurate.

## AI Security

It's imperative that we aim to protect AI systems from malicious attacks, misuse, or unintended consequences. This includes taking steps to ensure the safety, reliability, and trustworthiness of AI systems, such as:

- Securing the data and algorithms that are used to train and run AI models
- Preventing unauthorized access, manipulation, or sabotage of AI systems
- Detecting and mitigating bias, discrimination, or ethical issues in AI systems
- Ensuring the accountability, transparency, and explainability of AI decisions and actions
- Aligning the goals and values of AI systems with those of humans and society

## Data Protection

LLMs can pose risks to the privacy and security of the data that they use. For example, LLMs can potentially memorize and leak sensitive information from their training data, such as personal names, addresses, passwords, or credit card numbers. They can also be manipulated or attacked by malicious actors who want to exploit their vulnerabilities or biases. Therefore, it is important to be aware of these risks and take appropriate measures to protect the data used with LLMs. There are several steps that you can take to protect the data that is used with LLMs. These steps include:

- **Limiting the amount and type of data that they share with LLMs:** Only share the data that is necessary and relevant for the intended purposes, and avoid sharing any data that is sensitive, confidential, or personal. Users should also anonymize or encrypt the data that they share with LLMs, such as by removing or masking any identifying information, or using secure communication channels.
- **Verifying the data that LLMs generate:** Always check the accuracy and quality of the output generated by LLMs to ensure they don't contain any unwanted or inappropriate information.
- **Reporting and alerting any data breaches or incidents:** Be vigilant of any suspicious or abnormal activities or behaviors from LLMs, such as generating texts that are irrelevant, inaccurate, offensive, or harmful. This could be an indication of a data breach or security incident.

Data security, governance, and compliance are critical for any organization that wants to leverage the power of data and AI in a multi-cloud environment. Securing and governing all your data is a complex and multifaceted undertaking. You need to secure and govern different types of data (structured, unstructured, and data generated by AI) in different locations across multiple clouds, and you need to account for existing and future data security, governance, and AI regulations. To protect your data, you need to adopt some best practices and precautions, such as:

- Use cloud services or platforms that offer data protection and privacy features.
- Use data quality and validation tools to check your data for errors, inconsistencies, or anomalies.
- Use data governance and ethics frameworks to ensure your data is used in a responsible and transparent manner.