atlas.mitre.org /mitigations

Mitigations | MITRE ATLAS™

8-10 minutes

AML.M0000

Limit Public Release of Information

Limit the public release of technical information about the AI stack used in an organization's products or services. Technical knowledge of how AI is used can be leveraged by adversaries to perform targeting and tailor attacks to the target system. Additionally, consider limiting the release of organizational information - including physical locations, researcher names, and department structures - from which technical details such as AI techniques, model architectures, or datasets may be inferred.

AML.M0001

Limit Model Artifact Release

Limit public release of technical project details including data, algorithms, model architectures, and model checkpoints that are used in production, or that are representative of those used in production.

AML.M0002

Passive AI Output Obfuscation

Decreasing the fidelity of model outputs provided to the end user can reduce an adversaries ability to extract information about the model and optimize attacks for the model.

AML.M0003

Model Hardening

Use techniques to make AI models robust to adversarial inputs such as adversarial training or network distillation.

AML.M0004

Restrict Number of AI Model Queries

Limit the total number and rate of queries a user can perform.

AML.M0005

Control Access to AI Models and Data at Rest

Establish access controls on internal model registries and limit internal access to production models. Limit access to training data only to approved users.

AML.M0006

Use Ensemble Methods

Use an ensemble of models for inference to increase robustness to adversarial inputs. Some attacks may effectively evade one model or model family but be ineffective against others.

AML.M0007

Sanitize Training Data

Detect and remove or remediate poisoned training data. Training data should be sanitized prior to model training and recurrently for an active learning model.

Implement a filter to limit ingested training data. Establish a content policy that would remove unwanted content such as certain explicit or offensive language from being used.

AML.M0008

Validate AI Model

Validate that AI models perform as intended by testing for backdoor triggers or adversarial bias. Monitor model for concept drift and training data drift, which may indicate data tampering and poisoning.

AML.M0009

Use Multi-Modal Sensors

Incorporate multiple sensors to integrate varying perspectives and modalities to avoid a single point of failure susceptible to physical attacks.

AML.M0010

Input Restoration

Preprocess all inference data to nullify or reverse potential adversarial perturbations.

AML.M0011

Restrict Library Loading

&

Prevent abuse of library loading mechanisms in the operating system and software to load untrusted code by configuring appropriate library loading mechanisms and investigating potential vulnerable software.

File formats such as pickle files that are commonly used to store AI models can contain exploits that allow for loading of malicious libraries.

AML.M0012

Encrypt Sensitive Information

&

Encrypt sensitive data such as AI models to protect against adversaries attempting to access sensitive data.

AML.M0013

Code Signing

&

Enforce binary and application integrity with digital signature verification to prevent untrusted code from executing. Adversaries can embed malicious code in AI software or models. Enforcement of code signing can prevent the compromise of the AI supply chain and prevent execution of malicious code.

AML.M0014

Verify AI Artifacts

Verify the cryptographic checksum of all AI artifacts to verify that the file was not modified by an attacker.

AML.M0015

Adversarial Input Detection

Detect and block adversarial inputs or atypical queries that deviate from known benign behavior, exhibit behavior patterns observed in previous attacks or that come from potentially malicious IPs. Incorporate adversarial detection algorithms into the AI system prior to the AI model.

AML.M0016

Vulnerability Scanning

&

Vulnerability scanning is used to find potentially exploitable software vulnerabilities to remediate them.

File formats such as pickle files that are commonly used to store AI models can contain exploits that allow for arbitrary code execution. These files should be scanned for potentially unsafe calls, which could be used to execute code, create new processes, or establish networking capabilities. Adversaries may embed malicious code in model corrupt model files, so scanners should be capable of working with models that cannot be fully de-serialized. Both model artifacts and downstream products produced by models should be scanned for known vulnerabilities.

AML.M0017

AI Model Distribution Methods

Deploying AI models to edge devices can increase the attack surface of the system. Consider serving models in the cloud to reduce the level of access the adversary has to the model. Also consider computing features in the cloud to prevent gray-box attacks, where an adversary has access to the model preprocessing methods.

AML.M0018

User Training

&

Educate AI model developers on secure coding practices and AI vulnerabilities.

AML.M0019

Control Access to AI Models and Data in Production

Require users to verify their identities before accessing a production model. Require authentication for API endpoints and monitor production model queries to ensure compliance with usage policies and to prevent model misuse.

AML.M0020

Generative AI Guardrails

Guardrails are safety controls that are placed between a generative AI model and the output shared with the user to prevent undesired inputs and outputs. Guardrails can take the form of validators such as filters, rule-based logic, or regular expressions, as well as AI-based approaches, such as classifiers and utilizing LLMs, or named entity recognition (NER) to evaluate the safety of the prompt or response. Domain specific methods can be employed to reduce risks in a variety of areas such as etiquette, brand damage, jailbreaking, false information, code exploits, SQL injections, and data leakage.

AML.M0021

Generative AI Guidelines

Guidelines are safety controls that are placed between user-provided input and a generative AI model to help direct the model to produce desired outputs and prevent undesired outputs.

Guidelines can be implemented as instructions appended to all user prompts or as part of the instructions in the system prompt. They can define the goal(s), role, and voice of the system, as well as outline safety and security parameters.

AML.M0022

Generative AI Model Alignment

When training or fine-tuning a generative AI model it is important to utilize techniques that improve model alignment with safety, security, and content policies.

The fine-tuning process can potentially remove built-in safety mechanisms in a generative AI model, but utilizing techniques such as Supervised Fine-Tuning, Reinforcement Learning from Human Feedback or AI Feedback, and Targeted Safety Context Distillation can improve the safety and alignment of the model.

AML.M0023

AI Bill of Materials

An AI Bill of Materials (AI BOM) contains a full listing of artifacts and resources that were used in building the AI. The AI BOM can help mitigate supply chain risks and enable rapid response to reported vulnerabilities.

This can include maintaining dataset provenance, i.e. a detailed history of datasets used for AI applications. The history can include information about the dataset source as well as well as a complete record of any modifications.

AML.M0024

AI Telemetry Logging

Implement logging of inputs and outputs of deployed AI models. Monitoring logs can help to detect security threats and mitigate impacts.

Additionally, having logging enabled can discourage adversaries who want to remain undetected from utilizing AI resources.

AML.M0025

Maintain AI Dataset Provenance

Maintain a detailed history of datasets used for AI applications. The history should include information about the dataset's source as well as a complete record of any modifications.