Microsoft Security

5 Generative AI Security Threats You Must Know About

A definitive guide to unifying security across cloud and AI applications

Contents

Introduction

It’s been nearly three years since generative AI first became mainstream, and adoption isn’t showing any signs of slowing down. According to Microsoft research, 95% of security and IT decision makers are either in the planning stages of integrating generative AI or are actively developing and using the technology.[1]

Generative AI offers significant benefits by accelerating threat detection and automating repetitive tasks, but it also introduces risks. The technology’s dynamic responsiveness, while useful for adapting to real-time changes, can create challenges for security teams. It amplifies the threat landscape, empowering attackers with sophisticated tactics that makes cloud environments a prime target due to all their complexities.

When using generative AI applications, there are three key security challenges you need to be aware of:

Challenge 1

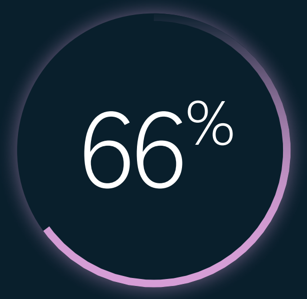
Most generative AI applications are cloud-based, which means attackers can more easily exploit vulnerabilities in the AI model, application, or other areas of the cloud environment to move laterally and compromise sensitive AI model grounding or user data. Organizations are also increasingly developing custom cloud-based AI applications, which could heighten the risk of potential vulnerabilities or misconfigurations.

Attackers only need a single entry point to breach cloud environments.

Challenge 2

AI models require access to large datasets to improve accuracy and fine-tune outputs. While this data is critical for enabling AI to fulfill a variety of use cases, it also makes AI an attractive target for attackers and creates challenges for security teams to protect the sheer scale of data that generative AI uses. Data leakage is a common AI concern, underscoring the need for security teams to proactively enforce data governance measures

IDC predicts that cloud-based deployments of AI platforms software will surpass the revenue from on-premises deployments by 2028.[2]



Of organizations are developing or planning to develop custom generative AI applications.[1]

Challenge 3

AI model outputs and behavior are variable, making it difficult for security teams to predict and control model behavior accurately. Because generative AI is designed to adapt to changing information, the same input can result in different outputs. It's impossible to predict all the different ways in which a bad actor might maliciously prompt a generative AI system, which is why prompt injections are such an effective attack vector. Security teams also need to juggle the rising trend of AI agent abuse, which takes advantage of the agents' ability to autonomously process information and make decisions without human input.

As organizations realize the technology's potential and integrate AI applications into everyday workflows, security teams need a holistic security approach that can protect against all aspects of AI risk. While defending against individual attack vectors is important, organizations must take a broader view by unifying security across the cloud and AI application lifecycle to accelerate innovation.

In this eBook, we will explore the top generative AI threats you should know about and show you how a cloud-native application protection platform (CNAPP) can harden your overall security posture while delivering rigorous threat protection for your cloud and AI workloads.

80%

Of business leaders cite data leakage via AI as a top concern.[3]

25%

Of enterprise breaches will be traced back to AI agent abuse by external and malicious internal actors by 2028.[5]

88%

Of organizations are somewhat or extremely concerned about indirect prompt injection attacks, which prey on generative AI's variable nature.[6]

40%

Of AI-related data breaches will be caused by improper use of generative AI across borders by 2027—highlighting the need for consistent AI best practices and data governance standards across the digital estate.[4]

The top 5 generative AI threats in 2025

Securing generative AI applications is different than securing cloud applications. While the same cloud risks still apply, there are several types of AI-specific attack vectors your organization needs to be aware of. This list is based on OWASP's top risks for LLM applications and the MITRE ATLAS Matrix.[7, 8]

Poisoning attack

Poisoning attacks manipulate or corrupt AI training data to influence model behavior and compromise its accuracy, reliability, or ethical responsibility. Attackers will target cloud storage accounts where training data is hosted to inject false or misleading information, modify, or delete portions of the dataset to change how the model responds to user queries.

Evasion attack

Evasion attacks circumvent existing AI security systems by modifying or manipulating input data in a way that the model cannot detect. For example, attackers might hide spam content within an image to bypass the model's spam filters. Evasion attacks can be targeted, meaning the attacker is trying to create a specific error in the model's output, or untargeted, meaning that the attacker is trying to produce any type of error. AI model jailbreaks are a type of evasion attack.

Functional extraction

In functional extraction attacks, adversaries repeatedly query AI models and use the responses to recreate and train functionally equivalent models. Attackers can then analyze this substitute model before launching further attacks on the commercially available version.

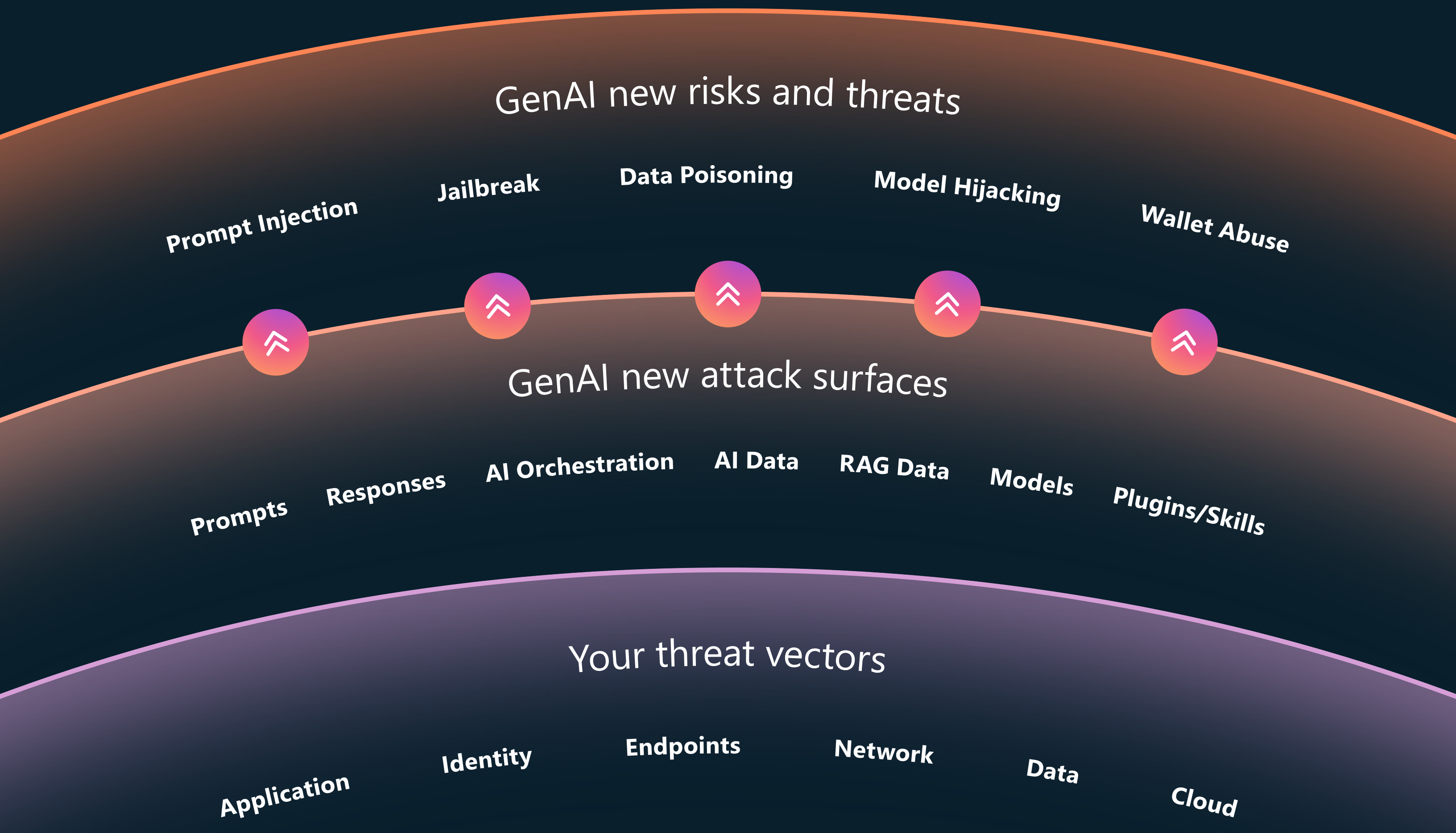
Inversion attack

Similar to functional extraction, during inversion attacks, threat actors repeatedly query AI models and use the outputs to infer information about the model's parameters or architecture. This type of attack can also allow adversaries to extract sensitive information about the model's training data, including everything from complete training data reconstruction to insights about specific data attributes or properties. Inversion attacks can be effective on their own or serve as a precursor to other threats, such as evasion attacks.

Prompt injection attack

Prompt injection attacks use malicious AI prompts to manipulate the model into behaving in unintended or harmful ways. These prompts are designed to make the model disregard its original instructions and follow the attacker's commands instead.

New attack surfaces introduce new risks and threats



GenAI new risks and threats

Prompt Injection

Jailbreak

Data Poisoning

Model Hijacking

Wallet Abuse

GenAI new attack surfaces

Prompts

Responses

AI Orchestration

AI Data

RAG Data

Models

Plugins/Skills

Your threat vectors

Application

Identity

Endpoints

Network

Data

Cloud

Broaden the scope of security posture and runtime security for AI

As any cyber defender can attest, cybersecurity is more nuanced than simply detecting and responding to individual attacks. Security teams can receive hundreds of alerts per day, leaving you scrambling to understand what each one means and which you should prioritize first. Instead, teams need an easier way to cut through the noise by correlating and contextualizing threats to get a full picture of the incident so you can investigate and respond to threats faster.

Cloud-native application protection platforms (CNAPP) are unified platforms that simplify cloud-native application and infrastructure security. These platforms integrate multiple solutions to embed security from application development to provisioning and runtime, helping mitigate risk across hybrid and multicloud environments.

Rather than providing separate alerts from cloud security posture management (CSPM), cloud infrastructure entitlement management (CIEM), or cloud workload protection platform (CWPP), CNAPP stitches all available signals together to provide a more complete understanding of ongoing attacks. These can include everything from identity and access data to storage logs, code vulnerabilities, internet exposure, and more. This correlation and contextualization is especially helpful given generative AI's unpredictable nature. Security teams must combine application and user behavior analysis with rich threat intelligence and posture management to accurately detect, disrupt, and remediate generative AI and cloud security threats alike.

Crucially, CNAPPs also act as a single platform where security teams can bridge the context gap with developers—providing common workflows, data, and insights so you can remediate risk at the source. This enables organizations to adopt a more proactive security stance by identifying and mitigating known vulnerabilities before attackers can exploit them.

However, not all CNAPPs are created equal. You need a comprehensive suite of tools that addresses all your security needs and integrates advanced threat intelligence to protect every layer of your cloud environment, including cloud storage, databases, and generative AI models and workloads.

Detect and respond to AI threats

Microsoft is the first solution provider to deliver end-to-end and comprehensive security for AI for both pre-built and custom-built AI applications. We believe that AI transformation requires security transformation, so organizations can build and use secure, safe, and trustworthy AI. Our CNAPP solution, Microsoft Defender for Cloud, enables companies to start secure with AI security posture management (AI-SPM) and stay secure with threat protection for AI workloads in runtime.

Security teams can use Defender for Cloud to effectively manage AI security posture across multi-model and multicloud environments. Defender for Cloud scans code repositories for Infrastructure-as-Code (IaC) misconfigurations and container images for vulnerabilities, enabling teams to detect and fix vulnerabilities and misconfigurations before deployment. Additionally, it continuously identifies risks, maps attack paths threat actors might use to reach sensitive assets, and employs built-in security best practices to prevent direct and indirect attacks across the full application lifecycle.

In runtime, Defender for Cloud provides security teams with AI workload detections to quickly identify and remediate active threats, such as jailbreak attacks, credential theft, and sensitive data leakage. Because these signals are natively integrated into Microsoft Defender XDR, security teams can use Defender for Cloud to facilitate incident response. Defender for Cloud is enriched with 84+ trillion daily signals from Microsoft Threat Intelligence, enabling security teams to continuously monitor AI workloads for malicious activity and rapidly respond to active threats.

Security is paramount at Icertis. That's why we've partnered with Microsoft to host our Contract Intelligence platform on Azure, fortified by Microsoft Defender for Cloud. As large language models (LLMs) became mainstream, our Icertis ExploreAI Service leveraged generative AI and proprietary models to transform contract management and create value for our customers.

Microsoft Defender for Cloud emerged as our natural choice for the first line of defense against AI-related threats. It meticulously evaluates the security of our Azure OpenAI deployments, monitors usage patterns, and promptly alerts us to potential threats. These capabilities empower our Security Operations Center (SOC) teams to make more informed decisions based on AI detections, ensuring that our AI-driven contract management remains secure, reliable, and ahead of emerging threats.

Subodh Patil

Principal Cyber Security Architect,
Icertis

Securing generative AI innovations

Mia Labs, Inc. was looking for sophisticated technology to produce and protect its conversational AI virtual assistant, Mia, and cutting-edge threat protection for its Azure infrastructure.

Mia Labs leverages Defender for Cloud's sophisticated security posture and threat protection capabilities to safeguard its fast-growing and complex environment in the automotive sales and service industry. Defender for Cloud provides contextual AI-security posture management and protects AI workloads with runtime security alerts, exposing active threats to Mia Labs' generative AI systems.

> We've successfully resisted multiple jailbreak attempts. Defender for Cloud shows us how to design our processes with optimal security and monitor where jailbreak attempts may have originated Defender for Cloud actually educates us on security, and it's something our team learns very readily.

Marwan Kodeih

Chief Product Officer,
Mia Labs, Inc

> Defender for Cloud provides so much threat intelligence that we’re spared the expense of costly IT audits, which would be prohibitive for a new company.

Kelvin Pho

Chief Technology Officer,
Mia Labs, Inc

Unlock AI’s potential with unified security

Generative AI holds enormous potential for scaling modern security operations. It can help security teams surface critical security insights faster, automate threat response at scale, accelerate decision making, and enable continuous improvement and innovation across the entire organization. However, to unlock that power, it’s important to understand the nuances of securing AI applications in dynamic cloud environments. By unifying security across the cloud and AI application lifecycle with a CNAPP, organizations can strengthen their generative AI security posture and defend against threats from code to runtime.

To learn more about securing all aspects of your hybrid and multicloud environments, including securing custom AI applications, visit the Microsoft cloud security solutions page.

Explore Microsoft cloud security solutions

For a deeper dive on protecting generative AI applications, download our eBook, “Stopping Generative AI Threats in Runtime: 5 Common Attack Vectors and How to Remediate Them.”

Download the AI security eBook

Source:

[1] “Accelerate AI transformation with strong security: The path to securely embracing AI adoption in your organization,” Microsoft.

[2] “Worldwide Artificial Intelligence Platforms Software Forecast, 2024–2028: AI Integration Accelerates, Fueling Technological Breakthroughs and Business Transformations,” IDC

[3] “First Annual Generative AI Study: Business Rewards vs. Security Risks,” iSMG.

[4] “Gartner Predicts 40% of AI Data Breaches Will Arise from Cross-Border GenAI Misuse by 2027,” Gartner.

[5] “Gartner Unveils Top Predictions for IT Organizations and Users in 2025 and Beyond,” Gartner.

[6] “From plan to deployment: Implementing a cloud-native application protection platform (CNAPP) strategy,” Microsoft.

[7] “OWASP Top 10 for LLM Applications 2025,” OWASP.

[8] “ATLAS Matrix,” MITRE.