

# GenAI Vulnerability Mapping in Healthcare SOCs

**Author:** Kim Lien Chu – Healthcare Cybersecurity Analyst

**Theme:** Emerging Threats, Detection Logic, and MITRE ATLAS Integration

---

## Summary

As GenAI platforms become embedded in healthcare workflows, adversaries are rapidly adapting — exploiting AI-driven systems with novel tactics like prompt injection, plugin abuse, and automated exploit generation. This walkthrough maps GenAI vulnerabilities to traditional web risks, outlines SOC detection logic, and integrates MITRE ATLAS and HHS 405(d) to guide defenders in securing AI-integrated environments.

---

## Threat Landscape Overview

| Threat Vector                    | Description  | Healthcare Impact                         |
|----------------------------------|--|---|
| Prompt Injection                 | Malicious manipulation of GenAI input to override instructions | PHI leakage, unauthorized actions         |
| Indirect Prompt Injection        | Hidden instructions in external sources (emails, lab reports)  | Unintentional model behavior              |
| GenAI + External Resource Access | AI systems pulling data from APIs, URLs, or plugins            | SSRF, token exfiltration, cloud pivoting  |
| LLM-Generated Exploit Code       | GenAI used to auto-generate payloads for known vulnerabilities | Faster exploit cycles, harder attribution |
| CVE-2024-3400 Exploitation       | GenAI-generated scripts targeting PAN-OS GlobalProtect         | Signals adversary experimentation         |

**Figure 1 – GenAI Integration Across Healthcare Domains**

*GenAI platforms now power critical healthcare functions, expanding the attack surface for adversaries exploiting AI-driven systems.*

---



## Case Study: CVE-2024-3400 – PAN-OS GlobalProtect

CrowdStrike observed threat actors using GenAI to generate exploit code targeting CVE-2024-3400, a command injection flaw in Palo Alto firewalls. While most payloads were ineffective, they revealed a shift toward automated exploit development.

“Lower-tier eCrime actors are abusing GenAI to generate scripts and build malware—automating tasks that once required advanced expertise.”

— CrowdStrike Threat Hunting Report, 2025

---



## SOC Detection Priorities

| Technique                   | Detection Logic Example (Splunk + Sysmon)                     | Mitigation Strategy                                 |
|-----------------------------|---|---|
| Prompt Injection            | Monitor for anomalous prompt patterns, role override attempts | Apply system-level guardrails; sanitize inputs      |
| Curl-Based RCE              | CommandLine="*curl*" AND CommandLine="*bash*"                 | Block remote script piping; inspect URLs            |
| SSRF via GenAI Plugins      | Log outbound requests to internal IP ranges                   | Enforce URL whitelisting; isolate internal services |
| SQL Injection in GenAI Code | Scan generated code for unsafe query construction             | Use prepared statements; validate outputs           |

---



## MITRE ATLAS Integration – GenAI Tactics and Techniques

MITRE ATLAS now tracks adversarial tactics against AI systems, including prompt injection and model evasion—extending traditional ATT&CK mappings to GenAI contexts.

In techniques like **LLM Jailbreak (AML.T0054)**, the AI — specifically the Large Language Model (LLM) — is the target, the tool, and sometimes the attack surface:

- ◆ **AI as the Target**

Adversaries craft malicious prompts to manipulate the LLM itself:

- Bypass safety filters, jailbreak ethical constraints, override system instructions

- Example: A prompt tricks the LLM into revealing PHI or generating restricted content

#### ◆ **AI as the Tool**

Once jailbroken, the LLM can be used to:

- Generate exploit code
- Summarize stolen data
- Interface with plugins or APIs
- Automate phishing or social engineering

#### ◆ **AI as the Attack Surface**

LLMs often connect to:

- External tools (via function calling)
- Cloud services (via plugins or APIs)
- Sensitive data (via integrated EHRs or chatbots)

This makes them a pivot point for broader attacks:

- SSRF via plugin abuse
- PHI leakage via unfiltered responses
- Unauthorized cloud access via prompt chaining

# LLM Jailbreak

## Summary

An adversary may use a carefully crafted [LLM Prompt Injection](#) designed to place LLM in a state in which it will freely respond to any user input, bypassing any controls, restrictions, or guardrails placed on the LLM. Once successfully jailbroken, the LLM can be used in unintended ways by the adversary.

**ID:** AML.T0054

**Mitigations:** [Generative AI Guardrails](#), [Generative AI Guidelines](#), [Generative AI Model Alignment](#)

**Tactics:** [Defense Evasion](#), [Privilege Escalation](#)

**Created:** 25 October 2023

**Last Modified:** 25 October 2023

## Figure 2 – MITRE ATLAS Technique AML.T0054: LLM Jailbreak

*Adversaries use prompt injection to jailbreak LLMs, bypassing restrictions and triggering unintended model behavior. This technique is mapped to Defense Evasion and Privilege Escalation.*

# MITRE ATLAS™

View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the [ATLAS Navigator](#).

| Privilege Elevation & |  | Defense Evasion &       | Credential Access &             | Discovery &                          | Collection &           | AI Attack Staging            | Command and Control &              | Exfiltration & | Impact &     |  |
|-----------------------|--|-------------------------|---------------------------------|--------------------------------------|------------------------|------------------------------|------------------------------------|----------------|--------------|--|
| Techniques            |  | 8 techniques            | 1 technique                     | 7 techniques                         | 3 techniques           | 4 techniques                 | 1 technique                        | 5 techniques   | 7 techniques |  |
| Plugin & Breakout     | Evade AI Model                             | Unsecured Credentials & | Discover AI Model Ontology      | AI Artifact Collection               | Create Proxy AI Model  | Reverse Shell                | Exfiltration via AI Inference API  | Evade AI Model |              |  |
|                       | LLM Jailbreak                              |                         | Discover AI Model Family        | Data from Information Repositories & | Manipulate AI Model    | Exfiltration via Cyber Means | Denial of AI Service               |                |              |  |
|                       | LLM Trusted Output Components Manipulation |                         | Discover AI Artifacts           | Data from Local System &             | Verify Attack          |                              | Spamming AI System with Chaff Data |                |              |  |
|                       | LLM Prompt Obfuscation                     |                         | Discover LLM Hallucinations     | Discover AI Model Outputs            | Craft Adversarial Data |                              | Erode AI Model Integrity           |                |              |  |
|                       | False RAG Entry Injection                  |                         | Discover LLM System Information | LLM Response Rendering               | LLM Data Leakage       | Cost Harvesting              | Extract LLM System Prompt          |                |              |  |
|                       | Impersonation &                            |                         | Cloud Service Discovery &       |                                      |                        |                              | External Harms                     |                |              |  |
|                       | Masquerading &                             |                         | Craft Adversarial Data          |                                      |                        |                              | Erode Dataset Integrity            |                |              |  |
|                       | Corrupt AI Model                           |                         |                                 |                                      |                        |                              |                                    |                |              |  |

**Figure 3 – MITRE ATLAS Matrix View (LLM Jailbreak Placement)**

*LLM Jailbreak appears under Defense Evasion in the MITRE ATLAS matrix, reinforcing its role in adversarial manipulation of GenAI systems.*

## ✳️ Mitigation Mapping for AML.T0054

| Mitigation               | Description   | Healthcare Relevance   |
|--------------------------|---|--|
| Generative AI Guardrails | Technical controls that restrict model behavior (e.g., output filtering, plugin sandboxing)   | <input checked="" type="checkbox"/> High — aligns with PHI protection, plugin control, and SOC enforcement |
| Generative AI Guidelines | Policy-level guidance for responsible use of GenAI (e.g., staff training, ethical boundaries) | <input checked="" type="checkbox"/> High — supports clinical governance, insider threat mitigation         |

|                               |  |  |
|-------------------------------|--|--|
| Generative AI Model Alignment | Techniques to align model outputs with intended values (e.g., reinforcement learning, safety tuning) | <span style="color: orange;">⚠</span> Medium — important but less actionable for SOC teams managing deployed tools |
|-------------------------------|--|--|

#### Figure 4 – MITRE ATLAS Mitigation Strategies for LLM Jailbreak

Among the three listed mitigations for AML.T0054, Generative AI Guardrails offer the most direct alignment with healthcare SOC priorities. These controls restrict model behavior, enforce plugin boundaries, and help prevent PHI leakage — making them essential for operational defense.

---



#### HHS 405(d) Alignment – Cyber Safety Is Patient Safety

| 405(d) Focus Area       | GenAI Risk Mapping                                    | SOC Action   |
|-------------------------|---|--|
| Data Protection         | PHI exposure via prompt injection or model training   | Enforce data minimization and model isolation        |
| Cloud Security          | GenAI plugins accessing external APIs                 | Monitor outbound traffic and plugin behavior         |
| Insider Threats         | Indirect prompt injection via staff-generated content | Train staff on GenAI risks and input hygiene         |
| Asset Management        | AI systems treated as privileged assets               | Include GenAI endpoints in asset inventory           |
| Cyber Hygiene Education | Lack of awareness around GenAI misuse                 | Disseminate AI-specific hygiene posters and training |

#### Figure 5 – HHS 405(d) Cyber Hygiene Pillars

HHS 405(d) reinforces the need for governance, training, and technical controls to secure GenAI systems in clinical environments.

Figure 5 is derived from core themes outlined in the HHS 405(d) Health Industry Cybersecurity Practices (HICP) publication and supporting educational materials. These pillars reflect the foundational cybersecurity focus areas promoted across the HPH sector.

# How to Implement Data Security

Data Security for Large Healthcare Organizations



## What is data security?

A security breach is the loss or exposure of sensitive data, including information relevant to the organization's business and patient Protected Health Information (PHI). Impacts to the organization can be profound if data are corrupted, lost, or stolen. Thus, good data security practices protect the organization and its patients.

## Why is it important?

When security breaches of data occur, it can prevent your employees from completing work accurately or on time and could result in potentially devastating consequences to your patients' treatment and wellbeing. Secure organizational data is not only important for your patients, but also for your organization's financial wellbeing and reputation.

## How will this keep my organization safe?

Properly securing data can prevent your organization from suffering major data losses during a cyber-attack. All staff of an organization, regardless of size, are the first line of defense when it comes to cyber-attacks. If you prepare your work force to recognize and identify potential cyber threats, your patients, and your organization will be more secure.

## Data security mitigates:

- Ransomware
- Loss or Theft of Equipment
- Insider, Accidental, or Intentional Data Loss



**Utilize cloud storage for sensitive data.** Use cloud access security broker systems to monitor data flows into cloud systems. Label data identified as sensitive. Implement digital rights and encryption to limit access to sensitive data. Ensure that cloud-based file storage and sharing systems do not expose sensitive data in an "open sharing" construct without authentication.



**Implement secure storage for inactive devices on your network.** Assets that are not in circulation should be returned to the appropriate IT department for secure storage. Storage areas should be secured with physical access controls. Access should be limited to those who require it. Physical access controls may include badge readers, video camera surveillance, and door alarms. If an asset is identified for redeployment, it should be securely imaged to deploy a "fresh" computer system for the new user. This ensures that old, sensitive data are removed, and that the asset has a clean bill of health.



**Consider Implementing Data Loss Prevention (DLP) software in your organization.** Traditionally, DLP systems monitor email, file storage, endpoint usage, web usage, and network transmission. Prevent data breaches by monitoring the use of sensitive data on the network. Multiple DLP solutions exist and can be applicable depending on the types of data access channels that need to be monitored.



**Create processes to control access to data backup files.** With cyber-attacks like ransomware, attackers intend to disrupt both production and backup files. Attackers that launch ransomware attacks are aware that an organization's first response will be to contain the ransomware and then restore the uncorrupted files from a backup source. Implement access control mechanisms that will prevent the system being backed up from accessing the disk storage, except by required access channels. This will add an extra layer of security around your data backup files.

To learn more about how you can protect your patients from cyber threats check out the [Health Industry Cybersecurity Practices: Managing Threats and Protecting Patients](#) publication. Check out the available resources 405(d) has to offer by visiting our website at [405d.hhs.gov](#) and our social media pages: @ask405d on [Facebook](#), [Twitter](#), [LinkedIn](#) and [Instagram](#)!

## Figure 6 – HHS 405(d) Data Security Poster (Large Org Focus)

This official HHS 405(d) resource outlines data security strategies for large healthcare organizations. It emphasizes cloud-based storage, secure access controls, Data Loss Prevention (DLP), and proactive risk detection — all of which align with SOC priorities for protecting PHI and mitigating insider and outsider threats in GenAI-integrated environments.



## Strategic Takeaways

- **GenAI Is Now an Adversary Capability**

SOC teams must treat GenAI-generated payloads as part of the threat landscape—not just a developer tool.

- **Healthcare Context Amplifies Risk**  
AI systems ingesting clinical data are vulnerable to indirect prompt injection and unintentional PHI exposure.
  - **Detection Must Evolve**  
Behavioral monitoring and context-aware alerting are essential to catch GenAI-driven anomalies.
  - **MITRE ATLAS Is Now Actionable**  
Techniques like AML.T0054 give SOC teams a structured way to map, detect, and mitigate GenAI threats.
- 

## References

1. **MITRE ATLAS.** Technique AML.T0054 – LLM Jailbreak. MITRE Corporation. Accessed September 2025.  
<https://atlas.mitre.org/techniques/AML.T0054>
  2. **Robust Intelligence.** AI Cyber Threat Intelligence Roundup: August 2024. Accessed September 2025.  
<https://www.robustintelligence.com/blog-posts/ai-cyber-threat-intelligence-roundup-august-2024>
  3. **Cisco Blogs.** AI Cyber Threat Intelligence Roundup: January 2025. Accessed September 2025.  
<https://blogs.cisco.com/security/ai-cyber-threat-intelligence-roundup-january-2025>
  4. **CrowdStrike.** Threat Hunting Report 2025. Accessed September 2025. (*Include link if cited directly.*)
  5. **U.S. Department of Health and Human Services.** HHS 405(d) Cybersecurity Practices. Accessed September 2025. (*Include link if cited directly.*)
-