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

MITRE ATLAS™

Home › Techniques › LLM Jailbreak

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.

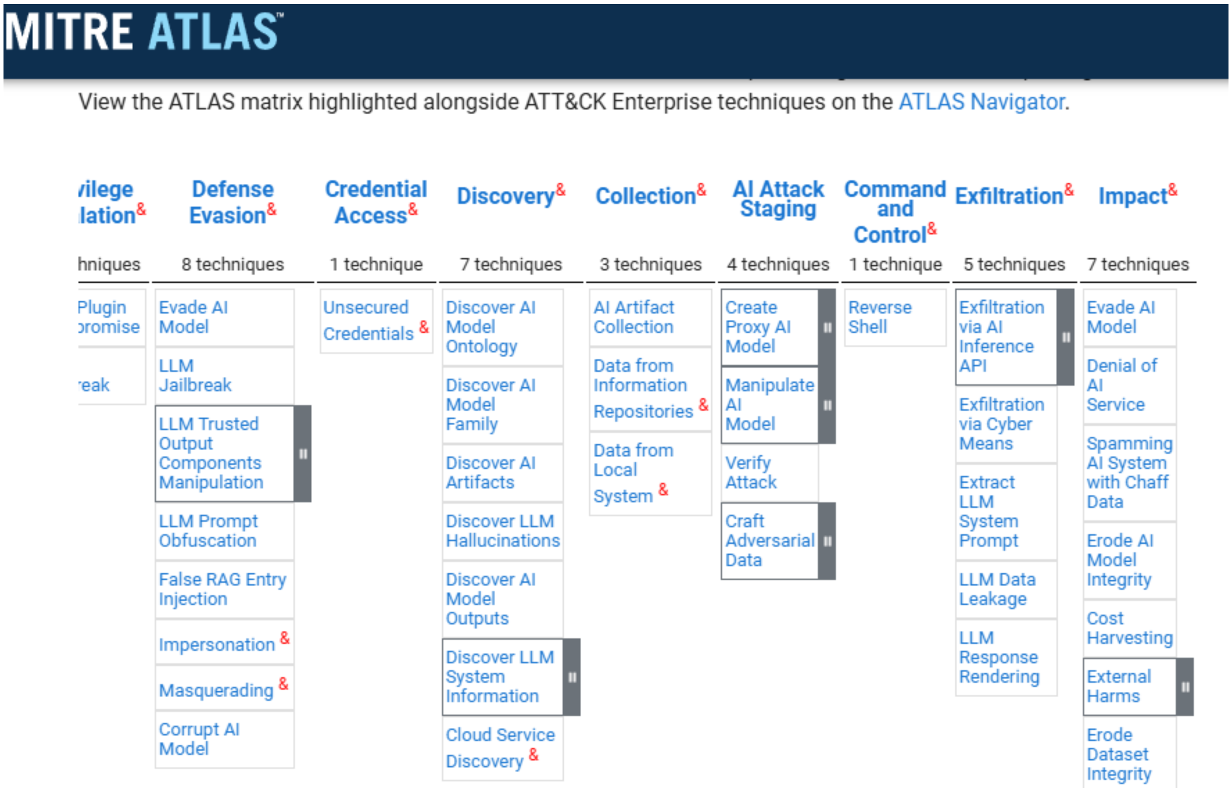

Figure 3 – MITRE ATLAS Matrix View (LLM Jailbreak Placement)
LLM Jailbreak appears under Defense Evasion in the MITRE ATLAS matrix, reinforcing its role in adversarial manipulation of GenAI systems.

🧩 Mitigation Mapping for AML.T0054

Mitigation	Description	Healthcare Relevance
Generative AI Guardrails	Technical controls that restrict model behavior (e.g., output filtering, plugin sandboxing)	✅ 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)	✅ 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)	⚠️ 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.

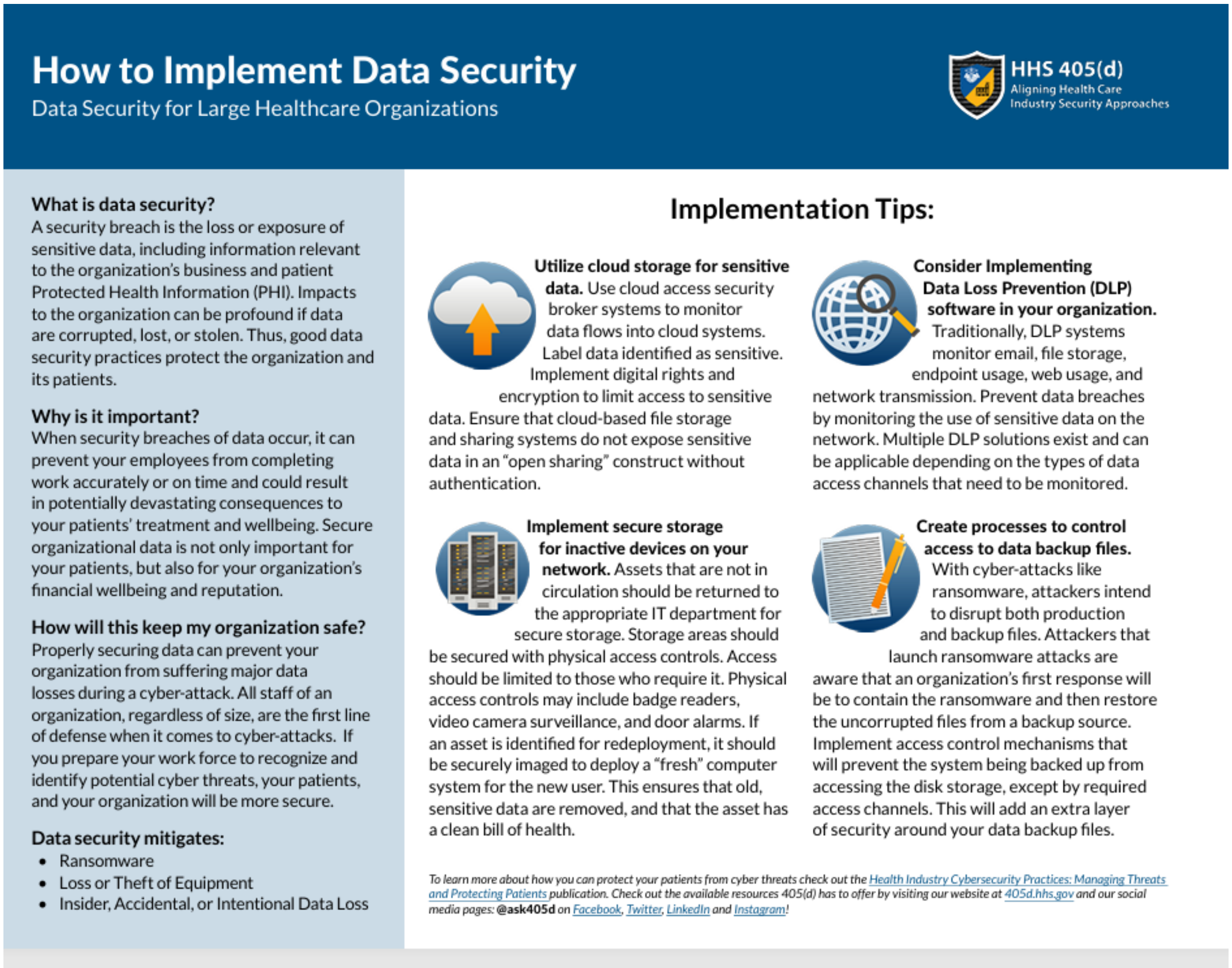

Figure 6 – HHS 405(d) Data Security Poster (Large Org Focus)

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.)*
