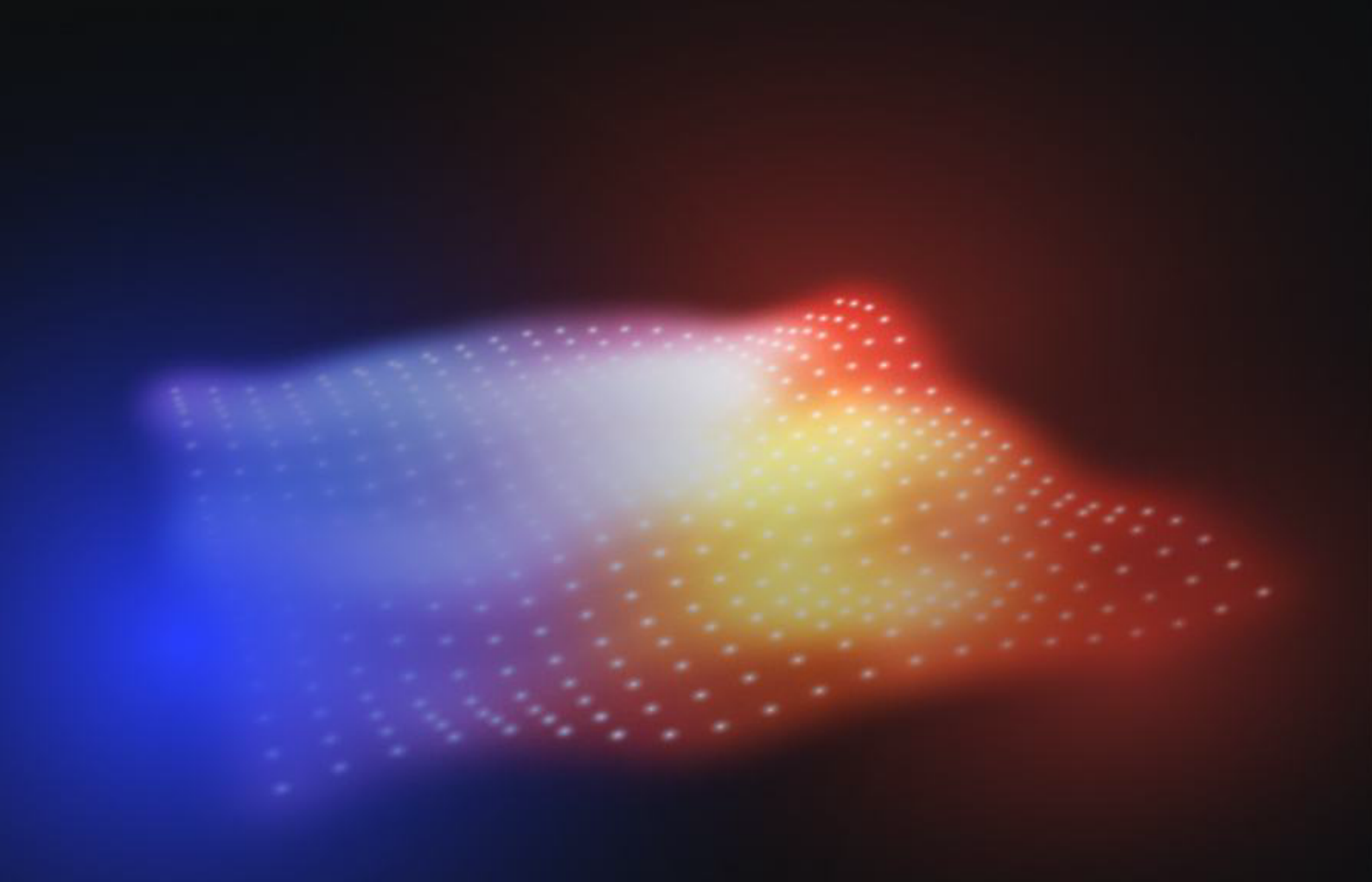
LAKERA

The AI Risk Map:

A Practical Guide to Frameworks, Threats, and GenAI Lifecycle Risks



Introduction: Navigating the AI Risk Landscape

The risks facing AI systems have evolved dramatically, especially with the rise of generative AI. What once lived in the world of abstract governance now demands urgent attention from security teams, product owners, and ML practitioners alike.

Yet as GenAI systems grow in complexity and capability, the guidance on how to secure them hasn't always kept up. A range of frameworks and standards now exist to define how we should evaluate and mitigate risk, but they vary widely in scope, focus, and relevance to the systems being built today.

This guide brings clarity to that landscape.

We focus on the frameworks that provide the clearest picture of today's GenAI risk surface, those that are:

- Actively maintained
- Purpose-built or adapted for LLMs, agents, or multimodal systems
- Used by teams building and securing real-world GenAI applications

Frameworks This Guide Is Built On

This section introduces the key frameworks that define the current landscape of AI and GenAI risks. Each one plays a different role: from technical threat modeling to secure development practices to regulatory guidance. Together, they form the foundation for the risk map we lay out in the rest of the guide.

1. MITRE ATLAS

MITRE's **Adversarial Threat Landscape for Artificial-Intelligence Systems** is the most detailed matrix of how AI systems (especially machine learning and GenAI) are attacked in practice. Based on the same model as MITRE ATT&CK, ATLAS maps attacker tactics and techniques across the full AI lifecycle, including model poisoning, evasion, extraction, and manipulation.

» Used throughout this guide as the foundation for understanding technical threats.

2. OWASP Top 10 for LLM Applications (2025)

First released in November 2024, this is the most up-to-date, community-vetted list of the top vulnerabilities affecting LLM-powered applications. It includes risks like **prompt injection, system prompt leakage, excessive agency, and vector-based attacks** that arise from modern practices like Retrieval-Augmented Generation (RAG).

» Serves as a frontline reference for anyone deploying LLMs in production environments.

3. OWASP LLM Security Verification Standard (LLMSVS)

A newly released verification standard providing **concrete security requirements** for building and evaluating LLM-backed systems. Structured across verification layers and control domains: from model lifecycle and real-time learning to plugin security and anomaly detection, it supports both development and audit efforts.

» Adds practical, testable controls that complement risk-oriented frameworks like OWASP Top 10 and MITRE ATLAS.

Explore Lakera's OWASP LLMSVS Cheatsheet for a quick and convenient overview of the standard.

4. NIST AI Risk Management Framework (NIST AI RMF)

While broader in scope and more governance-oriented, NIST AI RMF offers a structured approach to responsible AI development and risk management. Its value lies in **alignment with policy, risk categorization,** and high-level **accountability practices.**

» Referenced selectively in this guide where governance intersects with security.

5. EU AI Act (Regulatory Landscape)

The EU AI Act defines risk-based requirements for AI systems, including **transparency, data quality, documentation**, and additional obligations for **foundation models**. Though not a technical framework, it plays a key role in compliance planning.

» Included in this guide to help contextualize regulatory drivers of AI security work.

Read about the differences and similarities between the EU AI Act and the White House's AI Bill of Rights.

These frameworks don't exist in isolation, they intersect in ways that matter to real-world GenAI development and deployment. Throughout the guide, we draw on their strengths to highlight where risks emerge, how they evolve across the lifecycle, and what practical steps teams can take to address them.

Whether you're building, deploying, or securing GenAI systems, the goal is the same: make the landscape clearer, the risks more tangible, and the path to mitigation more actionable.

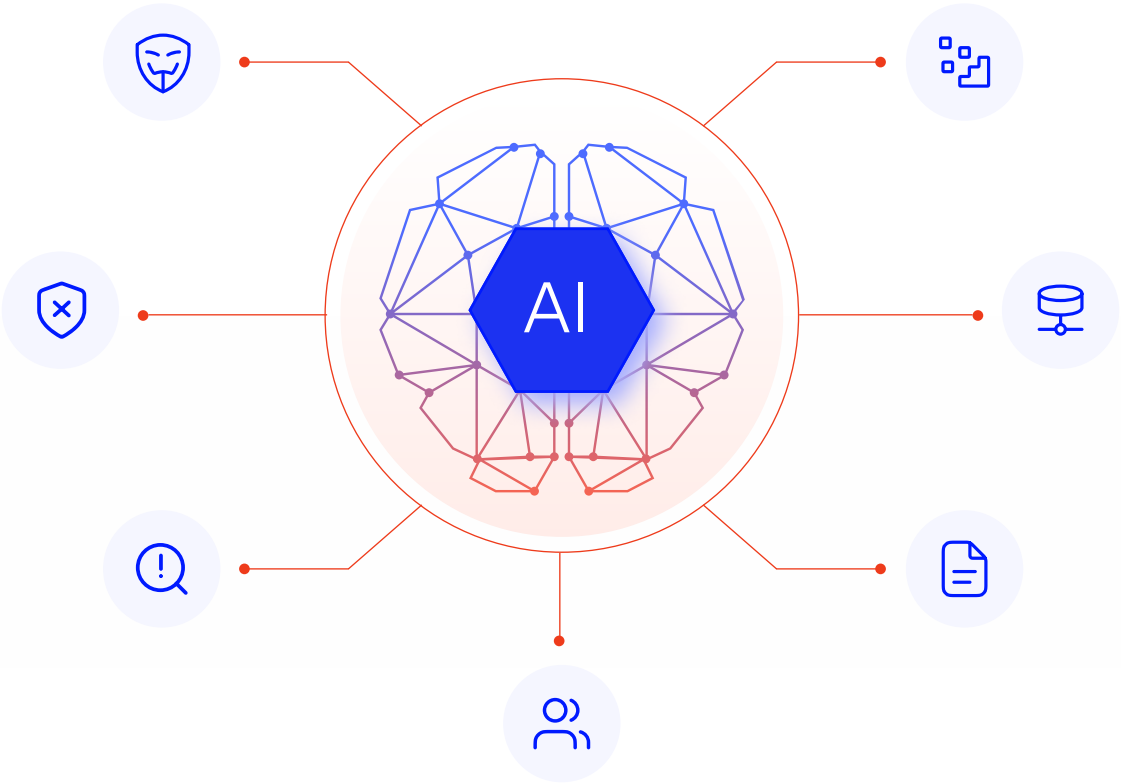
AI Risk Categories:
What Today's Frameworks Actually Show

Not all AI risks are created equal, and not all frameworks define them the same way.

This section distills the most **prominent categories of risk** that emerge when you line up the latest guidance from:

- **MITRE ATLAS** (latest release: 14 tactics, 100+ techniques)
- **OWASP Top 10 for LLM Applications (2025):** the most current view of LLM vulnerabilities in practice
- Selected references to **NIST AI RMF** and the **EU AI Act**, where policy meets security

These seven categories reflect how adversaries exploit GenAI systems, where the systems themselves fail, and where responsibility often falls through the cracks.



ADVERSARIAL MANIPULATION

Attacks designed to subvert or deceive the model, whether through inputs, poisoned data, or control over its behavior.

Key threats include:

- Prompt injection
- LLM jailbreaks
- Adversarial example crafting
- Poisoned training data
- Hallucination manipulation (e.g., publishing false RAG entries)

Where this shows up:

- **MITRE ATLAS:** Execution, Defense Evasion, ML Attack Staging
- **OWASP 2025:** LLM01:2025 – Prompt Injection, LLM04:2025 – Data and Model Poisoning, LLM06:2025 – Excessive Agency

This is the most technical and actively evolving category of risk—*and for good reason*.

GenAI models are increasingly embedded in complex systems, granted access to tools and APIs, and tasked with autonomous behavior. That means more surface area, more decision points, and more attack vectors. While safety layers can block obvious misuse, creative prompt injection and indirect attacks are designed to slip through the cracks.

That's why **Prompt Injection (LLM01:2025)** isn't just first in OWASP's list, it's also the foundation of many real-world breaches, jailbreaks, and misuses.

Explore prompt attacks step-by-step, learn how to spot and understand them, as well as assess their impact on GenAI systems: Understanding Prompt attacks: A Tactical Guide

MODEL AND DATA LEAKAGE

Risks tied to exposing proprietary models, sensitive training data, or hidden system prompts through normal use or attack.

Key threats include:

- Model extraction
- Training data leakage
- System prompt exposure
- Inference-based data reconstruction

Where this shows up:

- **MITRE ATLAS:** ML Model Access, Exfiltration
- **OWASP 2025:** LLM02:2025 – Sensitive Information Disclosure, LLM04:2025 – Data and Model Poisoning

These are **intellectual property** and **privacy risks**, especially critical for companies training or fine-tuning custom models. Leaked model artifacts or system prompts can also enable follow-on attacks, serving as a foothold for adversaries to craft more targeted exploits.

MISUSE AND OVERRELIANCE

When users or systems place too much trust in GenAI output, or when applications misuse models in ways that weren't intended or safe.

Key threats include:

- Blind trust in generated output
- Lack of human oversight
- Unsafe delegation to agents or plugins
- Overly permissive system behavior

Where this shows up:

- **OWASP 2025:** LLM05:2025 – Improper Output Handling, LLM06:2025 – Excessive Agency, LLM10:2025 – Unbounded Consumption
- **MITRE ATLAS:** Persistence, Privilege Escalation (via plugin abuse)

This is where **design flaws meet user behavior**, and where security often fails quietly.

INFRASTRUCTURE AND ACCESS RISKS

Technical weaknesses in how GenAI systems are integrated, hosted, or exposed, especially around APIs, plugins, or cloud environments.

Key threats include:

- Insecure model APIs
- Weak access controls
- Compromised plugins
- Supply chain threats in datasets or libraries

Where this shows up:

- **MITRE ATLAS:** Initial Access, Credential Access, ML Model Access
- **OWASP 2025:** LLM03:2025 – Supply Chain, LLM05:2025 – Improper Output Handling

This connects traditional **AppSec and cloud security** concerns to the GenAI stack.

OBSERVABILITY AND RESPONSE GAPS

Risks that arise when AI systems behave unpredictably, and no one is watching closely enough to detect or respond.

Key threats include:

- Lack of output validation
- No logging or monitoring
- Inability to detect abnormal usage patterns
- Failure to retrain or revalidate over time

Where this shows up:

- **MITRE ATLAS:** Discovery, Impact
- **OWASP 2025:** LLM07:2025 – System Prompt Leakage, LLM10:2025 – Unbounded Consumption
- **NIST AI RMF:** Continuous monitoring and impact evaluation

These are **operational blind spots**, often overlooked until it's too late.

ACCOUNTABILITY AND GOVERNANCE RISK

Structural risks from a lack of explainability, documentation, or clarity over who is responsible for what the AI does.

Key threats include:

- No audit trail of system decisions
- No model documentation
- Gaps in testing, validation, or red teaming
- Deployment of unverified models

Where this shows up:

- **NIST AI RMF, EU AI Act, ISO 42001**
- **OWASP 2025:** LLM06:2025 – Excessive Agency, LLM10:2025 – Unbounded Consumption
- **MITRE ATLAS:** indirectly via ML Attack Staging and Discovery

This is where **compliance, safety, and security intersect**, especially in enterprise and regulated environments.

BIAS, HARM, AND SOCIAL IMPACT

Risks tied to outputs that cause harm to individuals or groups due to bias, toxic content, misinformation, or unintended use.

Key threats include:

- Discriminatory decisions
- Toxic or offensive output
- Hallucinated legal/medical advice
- Misinformation propagation

Where this shows up:

- **OWASP 2025:** LLM04:2025 – Data and Model Poisoning, LLM09:2025 – Misinformation, LLM10:2025 – Unbounded Consumption
- **EU AI Act:** Prohibitions and transparency for high-risk use cases
- **NIST AI RMF:** Fairness and harm reduction focus

These risks often sit at the **intersection of safety, trust, and reputation**, and are increasingly tied to public and legal accountability.

Top 25 AI Risks Mapped

These 25 risks show up again and again in real-world GenAI systems. We picked them based on what comes up most often in leading AI security frameworks like OWASP and MITRE ATLAS, and what causes the biggest problems if you don't catch them early.

Technique / Risk	Definition	Category	Lifecycle Stage(s)	OWASP 2025	MITRE ATLAS*
Prompt Injection	Manipulating LLMs via crafted inputs to override instructions.	Adversarial Manipulation	Deployment	LLM01:2025 Prompt Injection	AML.T0051 – LLM Prompt Injection
Data Poisoning	Injecting malicious data into training to corrupt model behavior.	Adversarial Manipulation	Data, Training	LLM04:2025 Data and Model Poisoning	AML.T0020 – Poison Training Data
Model Extraction	Recreating model functionality via repeated querying.	Model and Data Leakage	Deployment	LLM02:2025 Sensitive Information Disclosure	AML.T0024.002 – Extract AI Model
Model Inversion	Reconstructing training data from model outputs.	Model and Data Leakage	Deployment	LLM02:2025 Sensitive Information Disclosure	AML.T0024.001 – Invert AI Model
Overreliance on Output	Assuming model responses are correct without validation.	Misuse and Overreliance	Deployment, Monitoring	LLM10:2025 Unbounded Consumption	AML.T0048.003 – External Harms: User Harm
Backdoor Injection	Training a model with hidden triggers for malicious behavior.	Adversarial Manipulation	Training	LLM04:2025 Data and Model Poisoning	AML.T0018 – Manipulate AI Model
Insecure Output Handling	Not sanitizing or controlling model responses before display.	Misuse and Overreliance	Deployment	LLM05:2025 Improper Output Handling	AML.T0067 – LLM Trusted Output Components Manipulation
Training Data Leakage	Exposure of sensitive or proprietary data used during training.	Model and Data Leakage	Data, Training	LLM02:2025 Sensitive Information Disclosure	AML.T0057 – LLM Data Leakage
Insufficient Monitoring	Failure to track and respond to model behavior post-deployment.	Observability and Response Gaps	Monitoring		AML.T0063 – Discover AI Model Outputs
Lack of Auditability	Inability to trace how an AI system made a decision.	Accountability and Governance Risk	Evaluation, Monitoring		AML.T0015 – Evade AI Model
Logic Bombs	Embedded triggers in a model that activate under specific conditions.	Adversarial Manipulation	Training, Deployment	LLM04:2025 Data and Model Poisoning	AML.T0018.000 – Poison AI Model
Hallucinated Advice	Confident but fabricated responses that mislead users.	Misuse and Overreliance	Deployment	LLM09:2025 Misinformation	AML.T0067.000 – LLM Trusted Output Manipulation

Technique / Risk	Definition	Category	Lifecycle Stage(s)	OWASP 2025	MITRE ATLAS*
Insecure Access Control	Lack of fi ne-grained access restrictions to AI systems or data.	Infrastructure and Access Risks	Deployment	LLM05:2025 Improper Output Handling	AML.T0012 – Valid Accounts
Unsafe Fine-tuning	Fine-tuning models without proper evaluation or safeguards.	Infrastructure and Access Risks	Training	LLM04:2025 Data and Model Poisoning	AML.T0017.000 – Develop Capabilities: Adversarial AI Attacks
Lack of Human Oversight	No human-in-the-loop review of high-risk AI decisions.	Accountability and Governance Risk	Deployment, Monitoring	LLM10:2025 Unbounded Consumption	AML.T0048.003 – External Harms: User Harm
Supply Chain Compromise	Insertion of malicious artifacts or poisoned data via dependencies.	Infrastructure and Access Risks	Build, Train, Deploy	LLM03:2025 Supply Chain	AML.T0010 – AI Supply Chain Compromise
Jailbreak via Prompt Injection	Using prompt injections to bypass LLM guardrails.	Adversarial Manipulation	Deployment	LLM01:2025 Prompt Injection	AML.T0054 – LLM Jailbreak
System Prompt Leakage	Exfi ltrating hidden instructions embedded in LLMs.	Model and Data Leakage	Deployment	LLM07:2025 System Prompt Leakage	AML.T0056 – Extract LLM System Prompt
RAG Poisoning	Contaminating external knowledge sources in retrieval-augmented generation.	Adversarial Manipulation	Training, Deployment	LLM08:2025 Vector and Embedding Weaknesses	AML.T0070 – RAG Poisoning
False RAG Entry Injection	Injecting documents that trick the LLM into treating false data as real context.	Adversarial Manipulation	Deployment	LLM08:2025 Vector and Embedding Weaknesses	AML.T0071 – False RAG Entry Injection
Plugin Exploitation	Abusing or compromising plugins connected to LLMs.	Infrastructure and Access Risks	Deployment	LLM06:2025 Excessive Agency	AML.T0053 – LLM Plugin Compromise
Unsecured Credentials	Exposed API keys or credentials used in model pipelines.	Infrastructure and Access Risks	Deployment	LLM02:2025 Sensitive Information Disclosure	AML.T0055 – Unsecured Credentials
Cost Harvesting	Using adversarial input to inflate model inference costs.	Impact / Abuse of Service	Deployment	LLM10:2025 Unbounded Consumption	AML.T0034 – Cost Harvesting
Evasion via Adversarial Input	Crafting input that intentionally bypasses model predictions.	Adversarial Manipulation	Deployment	LLM01:2025 Prompt Injection	AML.T0015 – Evade AI Model
Bias and Discrimination	Generation of harmful, unfair, or biased outputs.	Bias, Harm, and Social Impact	Training, Deployment	LLM09:2025 Misinformation	AML.T0048.002 – External Harms: Societal Harm

* To learn more about each of the ATLAS techniques and explore them easily, you can use the ATLAS Navigator.

Lifecycle Coverage Map

AI systems don't become risky only at runtime.

Security risks and policy blind spots emerge at **every stage** of the GenAI lifecycle: from dataset curation and fine-tuning to plugin integrations and user interaction. Understanding when these risks appear is just as important as understanding what they are.

This section maps the seven risk categories introduced earlier to the typical lifecycle stages of a GenAI product.

It helps teams pinpoint:

- Where in the lifecycle each risk emerges
- When guardrails, red teaming, or governance actions are most effective
- How OWASP and ATLAS view risk timing differently

LIFECYCLE STAGES

These are the high-level lifecycle stages used across this guide:

Stage	Description
Data	Collection, curation, and storage of raw and labeled datasets
Training	Model pretraining, fine-tuning, and internal evaluation
Build	Integration of models with applications, APIs, plugins, and infrastructure
Deployment	External exposure of the model to users or other systems
Monitoring	Ongoing observation, validation, and feedback after deployment
Governance	Auditability, documentation, and policy-setting across all other stages

Category-by-Stage Coverage Map

The table below maps each **AI risk category** to the **stages of the GenAI lifecycle** where that risk is most likely to **emerge or require mitigation.**

- A ❗ means the risk **typically originates** or is most **effectively addressed** at that stage.
- An empty cell means that category is **not strongly associated** with that lifecycle stage.

This overview helps teams see not just *what* the risks are, but *when* they need to pay attention to them.

Risk Category	Data	Training	Build	Deployment	Monitoring	Governance
Adversarial Manipulation	❗	❗	❗	❗	❗*	
Model and Data Leakage	❗	❗		❗	❗	❗
Misuse and Overreliance			❗	❗	❗	❗
Infrastructure Risks		❗	❗	❗		❗
Observability Gaps				❗	❗	❗
Accountability & Governance				❗	❗	❗
Bias and Harm	❗	❗		❗	❗	❗

** The risk itself usually begins earlier.*

Notes:

- **Adversarial threats** span nearly every stage, from poisoning during data collection to prompt injection at runtime.
- **Observability and governance gaps** often show up **only after deployment**, when it's often too late to patch them without significant disruption.
- **OWASP** is heavily focused on **Deployment and Build**, while **MITRE ATLAS** spans **all stages** in greater detail.
- Some risk categories (like **Bias and Harm**) are **cross-cutting** and require attention across the entire lifecycle, especially in regulated environments.

Who Should Use What?

Security risks in GenAI systems touch multiple teams: security engineers, ML teams, product leads, compliance officers, and not everyone needs the same framework or the same level of detail.

This section helps you match the **right framework to the right stakeholder.** Whether you're red teaming an AI product, shaping security policy, or ensuring compliance with new regulations, there's a framework that speaks your language (and your needs).

Framework Fit Matrix

Framework	Security Engineers & Red Teams	ML Practitioners & Builders	Product & Platform Teams	Compliance & Governance
MITRE ATLAS	✅	✅	⚠️	
OWASP Top 10 for LLM Applications	✅	✅	✅	⚠️
NIST AI RMF	⚠️	⚠️	⚠️	✅
EU AI Act		⚠️	⚠️	✅

Explanation & Guidance

MITRE ATLAS

Ideal for teams conducting technical threat modeling, red teaming, or LLM-specific risk assessments. Covers a wide range of attack techniques and lifecycle stages.

- ✅ Great for hands-on security practitioners
- ⚠️ Less actionable for policy or product teams without translation

OWASP Top 10 for LLM Applications (2025)

Practical, high-level risk checklist that bridges the gap between AppSec, ML, and platform teams. Helps identify common pitfalls in real-world GenAI systems.

- ✅ Easy to use across teams
- ⚠️ May lack depth for nuanced or low-level threats

NIST AI Risk Management Framework (RMF)

Provides governance, documentation, and accountability scaffolding. Most useful for policy-driven orgs or compliance-heavy industries.

- ✅ Strong alignment with responsible AI and ISO standards
- ⚠️ Light on actionable technical threat detail

EU AI Act

Focused on legal and regulatory obligations, especially around foundation models and high-risk use cases.

- ✅ Essential for compliance and policy planning
- ⚠️ Requires translation into security or dev team actions

Key Takeaways

The AI risk landscape is evolving fast and no single framework captures everything.

But taken together, frameworks like MITRE ATLAS, OWASP Top 10 for LLM Applications (2025), and NIST AI RMF give us a solid foundation for understanding how GenAI systems break, where risks emerge, and what defenses actually matter.

Here's what to keep in mind as you navigate this space:

1. You don't need to use every framework

Choose the ones that match your goals.

- Building LLM apps? Start with OWASP.
- Running red teams or risk assessments? Dive into MITRE ATLAS.
- Need governance and accountability? Look to NIST or the EU AI Act.

2. Risk happens across the entire lifecycle

Don't wait until deployment to start thinking about security. LLMs are intrinsically hackable so security needs to be designed into the system from the start.

Data curation, model fine-tuning, and plugin integration are all entry points for adversaries.

3. GenAI risks aren't just technical

Prompt injection is just one piece of the puzzle.

Overreliance, observability gaps, and governance blind spots are equally dangerous, and often easier to overlook.

4. The risks are new but the mindset isn't

Good AI security draws from both software security and ML safety. It often looks more like social engineering though than technical exploitation.

Think like an attacker. Think like a compliance officer. Think like a user.

5. You now have a map

This guide gives you a shared language for assessing AI risks, aligned with the best and most current frameworks.

Use it to evaluate your systems, guide red teaming, train your team, and stay ahead of what's next.