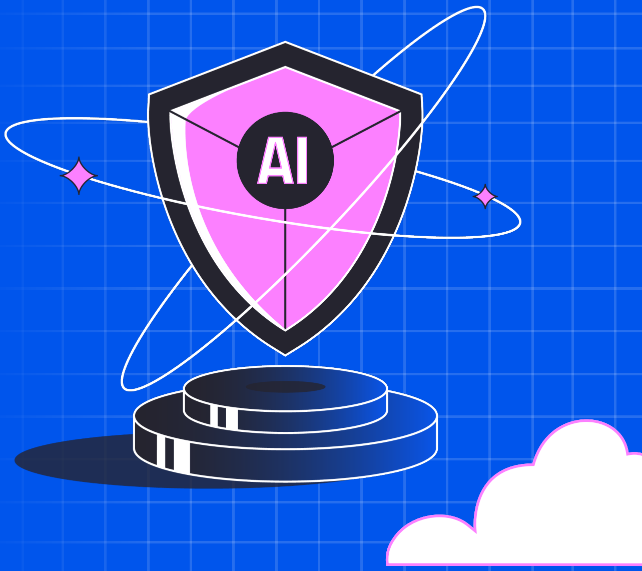
Cheat Sheet

GenAI Security Best Practices

AI and GenAI have become essential pillars for organizations aiming to enhance productivity and innovation. As the speed of AI releases and AI adoption accelerates, securing these AI pipelines takes on even greater importance.

AI risks are multi-faceted, and they require both existing SecOps techniques and new SecOps practices to evolve to address the unique challenges and characteristics of AI models and deployments.



The task for security teams is no small feat: Not only do you need to grasp the numerous technical aspects of AI but also understand the broader AI security landscape. You then need to gain visibility into AI usage within your organization before you can establish security-first processes to detect and respond to AI risks and threats effectively.

The key principles and objectives in this process include:

- **Visibility**
- **Zero critical risks**
- **Democratization**
- **Prevention**

This guide focuses on GenAI to provide you with a practical overview of best practices you can adopt to start fortifying your organization's GenAI security posture. Also, we'll share practical examples powered by Wiz's AI-SPM.

The concepts and techniques covered below apply to other AI use cases in general, and we encourage you to review our "Essential AI Security Best Practices" article for additional guidance on how to approach AI security.

Recap: What security risks come with GenAI?

GenAI is about generating new content from unstructured inputs (e.g., text, images, audio) that are extremely varied in format and inherently noisy. To allow for this creativity, GenAI carries new risks in terms of content anomalies, data protection, and AI application security.

Across these three categories, primary risks companies must keep an eye on include:

- **Data poisoning:** Maliciously altering training data to corrupt an AI model's outputs
- **Model theft:** Unauthorized access and duplication of proprietary AI models
- **Adversarial attacks:** Crafting inputs to steer AI models towards outputting incorrect, misleading, or harmful content.

- **Model inversion attacks:** Sending queries to an AI model to obtain sensitive training data
- **Supply chain vulnerabilities:** Exploiting weaknesses in the AI supply chain, such as third-party software dependencies, to compromise AI systems

Learn how these risks come to life from our interview with Clint Gibler, Head of Security Research at Semgrep.

The types of AI risk that apply to your organization depend on the AI that you employ and deploy. Understanding the security controls to set up around AI requires you to first have a good overview of the security risks that come with GenAI.

As you review common GenAI risks and threats, you should consider collecting your findings into an AI governance framework that can serve to both ensure regulatory compliance and set security standards.

Top 7 GenAI security practices

Below, you can find a brief introduction to seven of the main GenAI security practices for any enterprise organization.

1 Remove shadow AI

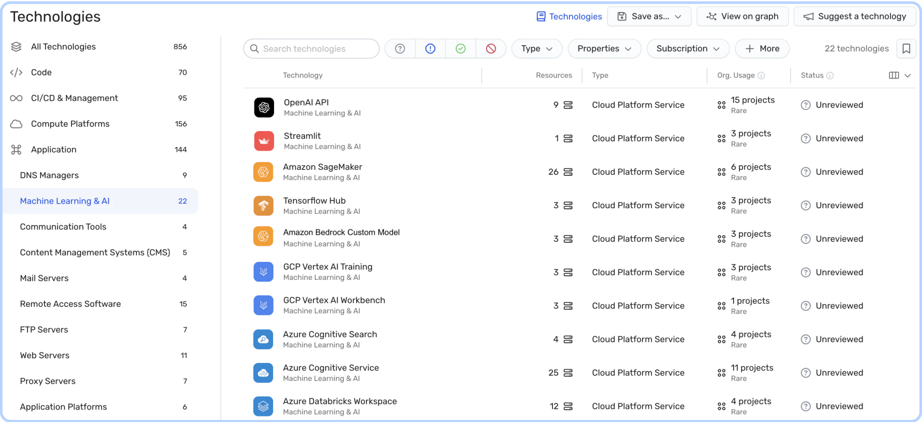
To ensure robust AI security, the first step is to achieve full visibility into what you are defending. This means eliminating any unauthorized AI usage within your organization.

Prerequisite: Everybody in your organization should know what they can and cannot do with GenAI. Make sure to add simple-to-follow GenAI security practices in the organization's general security guide.

Gaining visibility over all GenAI in the organization requires you to:

- Create an AI-BOM, i.e., a bill of materials collecting all your AI-related assets, ideally capable of automatically detecting new AI use.
- Set up relevant networking to ensure access for only whitelisted GenAI providers and software, or to block access to all those blacklisted.
- Foster education and awareness aimed at promoting a security-first mindset.

For example, you can review all the AI technologies used in your environment with a breakdown by project and resource through Wiz's AI-BOM.



All Technologies	856
Code	70
CI/CD & Management	95
Compute Platforms	156
Application	144
DNS Managers	9
Machine Learning & AI	22
Communication Tools	4
Content Management Systems (CMS)	5
Mail Servers	4
Remote Access Software	15
FTP Servers	7
Web Servers	11
Proxy Servers	7
Application Platforms	6

Technology	Resources	Type	Org. Usage	Status
OpenAI API Machine Learning & AI	9	Cloud Platform Service	15 projects Rare	Unreviewed
Streamlit Machine Learning & AI	1	Cloud Platform Service	3 projects Rare	Unreviewed
Amazon SageMaker Machine Learning & AI	26	Cloud Platform Service	6 projects Rare	Unreviewed
Tensorflow Hub Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
Amazon Bedrock Custom Model Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
GCP Vertex AI Training Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
GCP Vertex AI Workbench Machine Learning & AI	3	Cloud Platform Service	1 projects Rare	Unreviewed
Azure Cognitive Search Machine Learning & AI	4	Cloud Platform Service	4 projects Rare	Unreviewed
Azure Cognitive Service Machine Learning & AI	25	Cloud Platform Service	11 projects Rare	Unreviewed
Azure Databricks Workspace Machine Learning & AI	12	Cloud Platform Service	4 projects Rare	Unreviewed

By blocking Shadow AI, you minimize the chances of unexpected and unseen vulnerabilities for which you have no security controls in place; you also avoid any associated compliance issues.

2 Protect your data

Safeguarding sensitive information is crucial to maintaining organizational security and regulatory compliance. No sensitive information should be used in GenAI web and SaaS applications unless secured and approved, and no training data should be exposed and accessible through the GenAI model and application.

Prerequisite: Your team should agree with business and technical stakeholders on a definition of what constitutes sensitive information in your organization, possibly with different levels of criticality.

To protect your training and inference data:

- Discover and classify your data according to its security criticality.
- Use encryption for data at rest and in transit.
- Perform data sanitization such as removing or masking PII information for training data sets.
- Configure data loss prevention (DLP) policies to avoid sensitive data being used in end-user applications.
- Audit who has access to which data to understand effective access.

Wiz AI-SPM is fully integrated within Wiz's CNAPP, so you can use Wiz Data Security Posture Management (DSPM) to continuously monitor for sensitive data, detect exposures, eliminate attack paths for your data, and understand your IAM posture with Wiz Cloud Infrastructure Entitlements Management (CIEM).

For example, you can track the security of your storage buckets in all of your cloud deployments with Wiz to get early warnings of publicly exposed training data as well as publicly writable data that could be exploited for model poisoning.

By protecting access to your private data, you can prevent data breaches, protect intellectual property, secure models against poisoning, and ensure regulatory compliance.

3 Secure access to GenAI models

Unauthorized agents gaining access to GenAI models could deploy a variety of techniques to modify and misuse the model, such as introducing biases or harmful deceptions.

Prerequisite: A well-defined IAM configuration is a must-have for all assets associated with GenAI deployments and applications, with role-based access control (RBAC) recommended.

Whether the models are internal or external, you can add controls to secure GenAI models that allow you to:

- Set up authentication and rate limiting for API usage.
- Restrict access to model weights.
- Allow only required users to kickstart model training and deployment pipelines.

For example, you can easily review who has read access to a storage bucket or detect excessive permissions for a user by monitoring permissions with Wiz.

By controlling access to your GenAI models following the principle of least privilege, you can protect the integrity and reliability of your GenAI systems against model theft, tampering, and misuse.

4 Use LLM built-in guardrails

Following a multi-layer security-first mindset, it is always ideal to introduce security at the source by incorporating built-in guardrails of your GenAI models as security controls.

Prerequisite: Thoroughly review the documentation of GenAI providers and models to ensure they provide support for your designated guardrails.

Different GenAI providers and solutions offer different built-in security controls which may include:

- Content filtering to automatically remove or flag inappropriate or harmful content.
- Abuse detection mechanisms to uncover and mitigate general model misuse.
- Temperature settings to change AI output randomness to your desired predictability.

By setting up security controls against LLM misuse at the source, you minimize risk for both your organization and your application users.

5 Detect and remove AI risks and attack paths

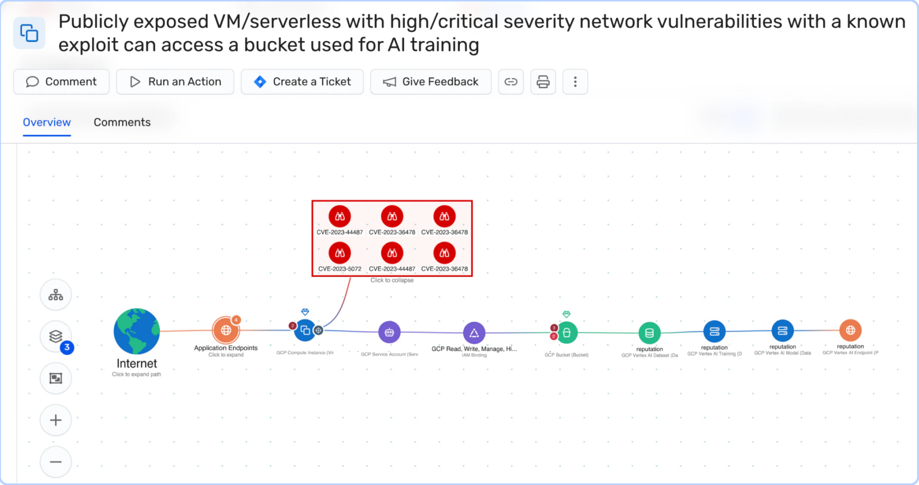
Attack path analysis (APA) preemptively identifies end-to-end attack paths composed of complex chains of exposures and lateral movement paths in your AI systems.

Prerequisite: End-to-end risk monitoring of your AI infrastructure with clear lineage and full context.

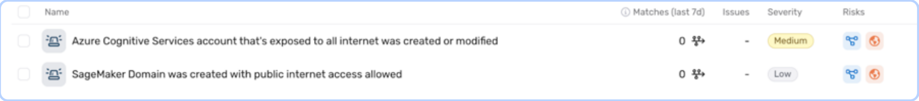
Your attack path analysis solution should:

- Continuously scan for and identify vulnerabilities in AI models.
- Verify all systems and components have the most recent patches to close known vulnerabilities.
- Scan for malicious models.
- Assess for AI misconfigurations, effective permissions, network exposures, exposed secrets, and sensitive data to detect attack paths.
- Regularly audit access controls to guarantee only authorized parties are granted access to critical systems.
- Provide context around AI risks so that you can proactively remove attack paths to models via remediation guidance.

For instance, scanning for malicious models with Wiz can help you discover risky pickling formats in your environment. Also, with Wiz Security Graph, you can gain context regarding what type of data the model can access, who can train the model, and if the model is exposed to the internet.



Publicly exposed VM/serverless with high/critical severity network vulnerabilities with a known exploit can access a bucket used for AI training



Name	Matches (last 7d)	Issues	Severity	Risks
Azure Cognitive Services account that's exposed to all internet was created or modified	0	-	Medium	
SageMaker Domain was created with public internet access allowed	0	-	Low	

By automatically detecting and removing attack paths, you are embracing a proactive security approach that actively prevents potential exploits, maintains the integrity of your AI systems, and safeguards sensitive data from malicious actors.

6 Monitor against anomalies

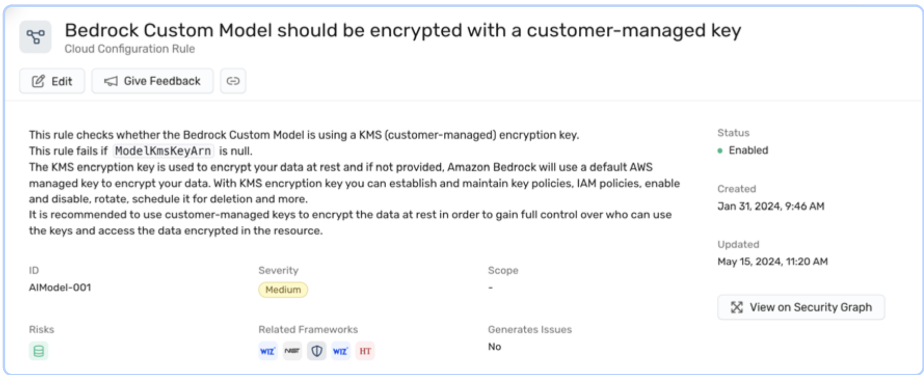
Continuous monitoring can help detect and address unusual activities in your AI systems promptly.

Prerequisite: A thorough monitoring solution should be put in place that provides extended detection for suspicious activity in GenAI applications.

Your monitoring solution should:

- Use anomaly detection and behavior analytics at both the input and output.
- Detect suspicious behavior in AI pipelines.
- Keep track of unexpected spikes in latency and other system metrics.
- Support regular security audits and assessments.

For example, Wiz's AI-SPM can help you detect suspicious access to your third-party AI endpoints such as an AWS Bedrock model linked to previously seen attacks.



Bedrock Custom Model should be encrypted with a customer-managed key

Cloud Configuration Rule

This rule checks whether the Bedrock Custom Model is using a KMS (customer-managed) encryption key.
This rule fails if `ModelKmsKeyArn` is null.
The KMS encryption key is used to encrypt your data at rest and if not provided, Amazon Bedrock will use a default AWS managed key to encrypt your data. With KMS encryption key you can establish and maintain key policies, IAM policies, enable and disable, rotate, schedule it for deletion and more.
It is recommended to use customer-managed keys to encrypt the data at rest in order to gain full control over who can use the keys and access the data encrypted in the resource.

ID: AIModel-001

Severity: Medium

Scope: -

Risks:

Related Frameworks:

Generates Issues: No

Status: Enabled

Created: Jan 31, 2024, 9:46 AM

Updated: May 15, 2024, 11:20 AM

Through advanced monitoring and alerting, you can rely on the ability of your security solution to alert you of any suspicious events in your AI pipelines so you can quickly reduce the blast radius.

7 Set up incident response

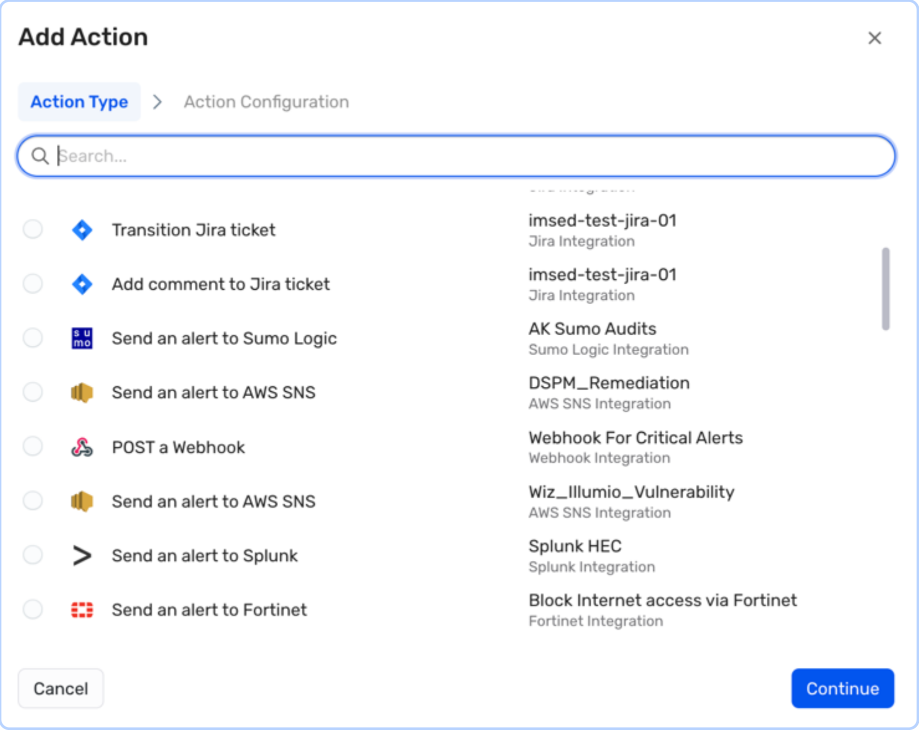
Preparing a swift incident response plan is critical to minimizing the blast radius of AI-related security incidents.

Prerequisite: A general incident response team should be available for critical AI systems and be able to rely on security tools designed for easy understanding of AI threats.

Incident response for GenAI systems can involve both automated and manual security controls, which include:

- Processes for isolation, backup, traffic control, and rollback.
- Integration with SecOps tools.
- Availability of an AI-focused incident response plan.

For example, Wiz supports integrations with all the most popular security tools. You can easily add a variety of actions such as sending alerts to Slack or transforming any discovered AI threats into a Jira ticket.



○	Transition Jira ticket	imsed-test-jira-01 Jira Integration
○	Add comment to Jira ticket	imsed-test-jira-01 Jira Integration
○	Send an alert to Sumo Logic	AK Sumo Audits Sumo Logic Integration
○	Send an alert to AWS SNS	DSPM_Remediation AWS SNS Integration
○	POST a Webhook	Webhook For Critical Alerts Webhook Integration
○	Send an alert to AWS SNS	Wiz_Illumio_Vulnerability AWS SNS Integration
○	Send an alert to Splunk	Splunk HEC Splunk Integration
○	Send an alert to Fortinet	Block Internet access via Fortinet Fortinet Integration

For example, Wiz supports integrations with all the most popular security tools. You can easily add a variety of actions such as sending alerts to Slack or transforming any discovered AI threats into a Jira ticket.

What’s next?

A proactive and agile approach to AI and GenAI security is necessary to keep up with the speed of development and adoption of these technologies.

An AI-SPM tool that supports security teams in setting up their AI posture with built-in best practices and automated processes is a big differentiator for ensuring that GenAI brings only the desired benefits to your organization.

Discover more about AI security and how Wiz can help. Sign up for a demo to see Wiz in action today.