orca SECURITY

2026

STATE OF

AI SECURITY REPORT

AI Security Findings from 1,200+ Production Cloud Environments

orca | RESEARCH POD

©2026 ORCA SECURITY. ALL RIGHTS RESERVED.

AI has led to exponential growth in the employees shipping to production and security teams are not keeping up. When the way people build changes, the way security works has to change with it. This report measures that gap, and closing it starts with giving security teams the context to understand AI risk and act.

GIL GERON,
CEO AND CO-FOUNDER OF ORCA SECURITY

Inside This Report

Foreword	01
About the Orca Research Pod	02
Executive Summary	03
Key Findings	04
1. AI Adoption: From Experiment to Infrastructure	07
2. The AI Supply Chain Under Attack	16
3. AI Agents and RAG: The Ungoverned Attack Surface	25
4. AI Sprawl and the Governance Gap	30
5. AI Credentials and Insecure Access	35
6. AI Infrastructure Exposure	37
7. AI Encryption: The Missing Layer	42
8. What This Means for Your Organization	46
9. Key Recommendations	48
10. Conclusion	52
About Orca Security	53

Foreword

AI has fundamentally reshaped how organizations build, deploy, and operate in the cloud, but security hasn't kept pace. In two years, AI has moved from experimental pilots to production infrastructure. Foundation models, autonomous agents, and RAG pipelines now connect to sensitive data. The controls to govern this exist, but adoption lags deployment, and basic hygiene like least privilege, authentication, and network isolation is still being skipped for speed.

AI workloads run on default configurations, API keys sit in plaintext, and critical vulnerabilities in AI packages go unpatched even when fixes exist. More than half of organizations building with AI have deployed agent frameworks with cloud permissions and without guardrails. These are structural realities, not theoretical risks.

The threat landscape has shifted too. 2025–2026 brought supply chain attacks on AI-specific packages, models autonomously discovering and exploiting zero-days, and agentic systems causing real damage when safety controls failed. The same tools accelerating development are accelerating the attack surface.

This report, grounded in the Orca Research Pod's analysis of 1,200+ production organizations, quantifies the gap between AI adoption and AI security adoption, names the controls organizations aren't configuring, and offers practical guidance for closing the gap before regulation forces the issue.

Gil Geron
CEO and Co-Founder of Orca Security

About the Orca Research Pod

The Orca Research Pod is a group of security researchers who discover and analyze security risks and vulnerabilities to strengthen the Orca Security Platform and advance cloud security best practices. The Pod regularly publishes original research that has been featured across the security industry. Their findings are referenced throughout this report alongside broader research telemetry and analysis.

Beyond analyzing telemetry from production environments, the Research Pod conducts original offensive security research into AI-specific attack vectors. Recent publications include RoguePilot: Exploiting GitHub Copilot for a Repository Takeover, which demonstrated a passive prompt injection chain achieving full repo takeover via GitHub Codespaces, and AI-Induced Lateral Movement (AILM), which introduced the AI layer as a new pivot vector, a third dimension of lateral movement alongside network and identity. Together, these findings illustrate how quickly the AI attack surface is evolving.

Research Methodology

This report is based on aggregated, anonymized security telemetry from over 1,200 production organizations using Orca Security's cloud platform. Metrics represent the percentage of organizations exhibiting each finding, weighted across the data set. Data was collected in Q2 2026 from production environments only, to reflect real-world security postures.

Findings span AI cloud service configuration, package vulnerability management, secrets exposure, agent and RAG infrastructure security, encryption posture, and AI workload identity management.

Report Data Set:

- Cloud workload and configuration data from over **1,200 organizations**
- AI-specific telemetry covering models, packages, agents, vector databases, and infrastructure configuration
- Data referenced in this report was collected in Q2 2026

Executive Summary

AI has moved from pilot to production faster than most security programs have followed. Drawing on real-world telemetry from over 1,200 organizations, this report finds a widening gap between the speed of AI adoption and the maturity of AI security that results in misconfigurations, unpatched vulnerabilities, exposed credentials, and ungoverned agents running in production. Below are our key findings:

- **AI adoption is mainstream, but security is an afterthought**
  51.5% of organizations have adopted AI to build custom applications, yet 80% of SageMaker deployments still run with all five core insecure defaults enabled.

- **AI package vulnerabilities are pervasive and unpatched**
  81% of organizations with AI packages carry at least one known High vulnerability. Public exploits now exist for 50.1% of alerts, up from 0.2% in our 2024 report. 99.9% of fixable alerts remain unpatched.

- **Widely exposed AI credentials enable direct compromise**
  28.4% of OpenAI users store API keys in unsecure locations (40% for Anthropic SDK users), and 94% of Azure OpenAI deployments still rely on key-based auth instead of managed identities.

- **Agents are in production, but guardrails are not**
  56% of AI adopters run agent frameworks in production, including 786 Bedrock agents across 8% of organizations, yet 57% of Bedrock users have no safety guardrails configured.

- **Encryption is effectively absent across all clouds**
  87–98% of organizations using AI services haven't configured customer-managed encryption keys (CMEK); SageMaker notebooks hold the highest rate at 98.4%, leaving training data, models, and pipelines reliant on default provider encryption.

- **AI sprawl is outpacing governance**
  55.3% of AI cloud users run four or more AI service types, and 19.6% run seven or more. Consistent security policy across this landscape is nearly impossible for most organizations.

- **AI regulation is on a 2026 calendar**
  The EU AI Act's high-risk obligations take effect August 2, 2026 (fines up to €35M or 7% of global turnover), and Colorado's amended AI law (S.B. 26-189) takes effect January 1, 2027. These misconfigurations are now regulatory realities.

For practical next steps, see Chapter 9 — Key Recommendations.

Key Findings

51.5%

of organizations have adopted AI to build custom applications.

AI is no longer experimental. Over half of organizations run AI in production but adoption has shifted: from pre-built models to agents, RAG pipelines, and custom AI infrastructure.

81%

of organizations with AI packages have at least one known vulnerability, up from 62% in 2024.

Average CVSS scores jumped from 6.9 to 8.79, and 74% now carry at least one critical CVE (9.0+). The AI dependency tree is deeper and more severe than two years ago.

50.1%

of AI package vulnerability alerts have a public exploit available, a 250x increase from 0.2% in 2024.

Half of all AI vulnerabilities are now actively exploitable, yet 99.9% of alerts with a fix available remain unpatched. The low-exploitability excuse no longer holds.

80%

of SageMaker organizations run with all default settings enabled.

The most widely adopted AI cloud service ships insecure by default, and orgs rarely reconfigure it. While settings have improved, the majority still run all five insecure defaults.

86%

of SageMaker orgs have at least one predictably-named bucket.

The legacy naming pattern is brute-forceable, exposing training data, model artifacts, and pipeline outputs. Adoption nearly doubled from 45% in 2024, despite AWS randomizing the default convention.

25% to 40%

of users across Anthropic, OpenAI, and HuggingFace have at least one API key stored in an unsecure location.

Anthropic exposure more than tripled (13% → 40%), OpenAI rose (20% → 28.4%), HuggingFace improved (35% → 24.9%). Exposure is widespread, provider-agnostic, and worsening where adoption is growing the fastest.

Key Findings

98%

of SageMaker notebook instances lack customer-managed encryption.

Azure OpenAI (91.6%) and Vertex AI Models (94.8%) follow close behind. While Vertex AI improved slightly since 2024, most organizations still rely on default provider-managed encryption for AI workloads.

56%

of AI adopters have deployed agent frameworks in production

LangGraph leads, with 786 Bedrock agents holding cloud permissions and 879 knowledge bases connected to data — new categories since 2024, reflecting the AI stack's rapid expansion into autonomous, data-connected systems.

57%

of Bedrock organizations run agents without guardrails.

More than half of organizations using Amazon Bedrock have not configured any safety guardrails, leaving AI agents operating without content filtering, grounding checks, or policy controls.

64%

of AI adopters have deployed vector databases.

Widespread RAG adoption across seven major vector databases creates a new and fragmented attack surface for data poisoning and prompt injection through retrieved documents.

55.3%

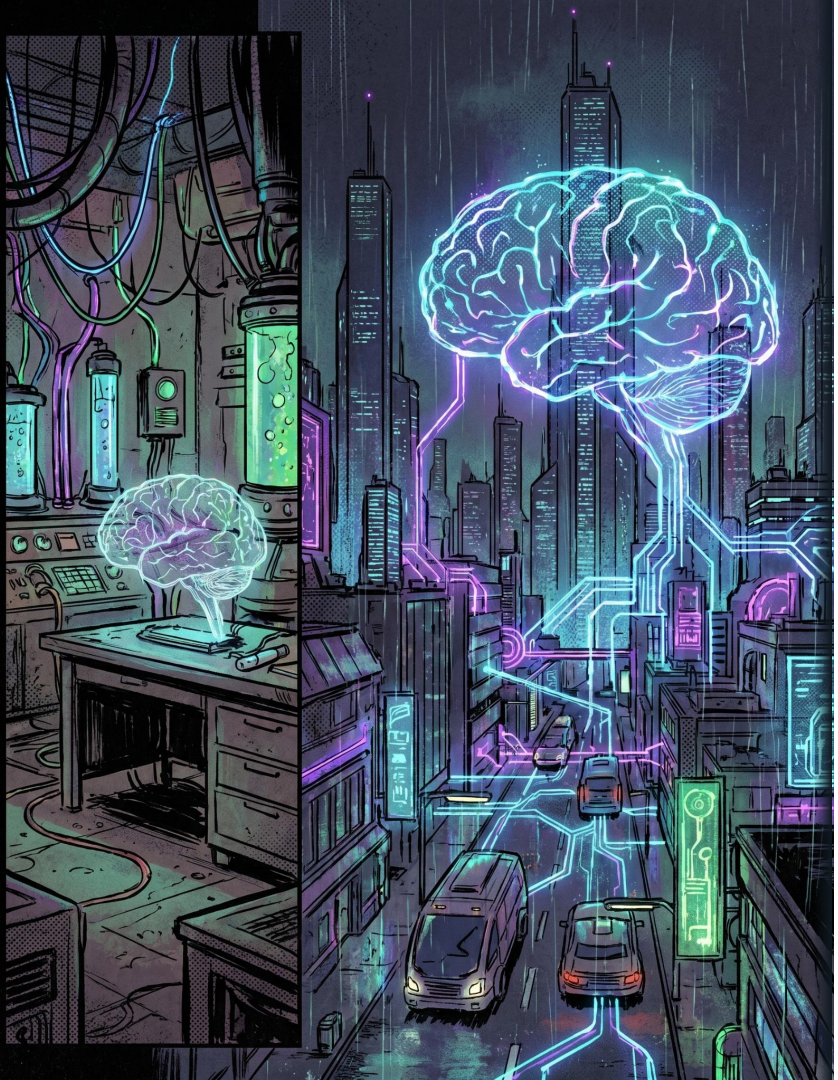
of AI cloud service users operate four or more distinct AI service types.

19.6% use seven or more. With AI workloads spread across multiple services and clouds, maintaining consistent security governance is a growing challenge for security teams.

Year-over-Year Comparison Summary

METRIC	2024 REPORT	2026 REPORT	TREND
Orgs with at least one vulnerable AI package	62%	81%	Worsened
Average CVSS across AI packages	6.9	8.79	Worsened
AI vulnerabilities with public exploit	0.2%	50.1%	Worsened (250x)
SageMaker with root access	98%	75.8%	Improved
SageMaker without IMDSv2	77%	47.9%	Improved
SageMaker using default bucket names	45%	86%	Worsened
OpenAI keys in unsecure locations	20%	28.4%	Worsened
HuggingFace keys in unsecure locations	35%	24.9%	Improved
Anthropic keys in unsecure locations	13%	40%	Worsened (3x)
Vertex AI without CMEK	98%	92.9%	Slightly Improved
Azure OpenAI adoption (% of Azure orgs)	39%	50.5%	Increased
SageMaker adoption (% of AWS orgs)	29%	31%	Increased
Vertex AI adoption (% of GCP orgs)	24%	32%	Increased
Top AI model	GPT-3.5 (79%)	GPT-4o (37.6%)	Generational shift
Agent frameworks deployed	Not tracked	56% of AI adopters	New category
Vector databases deployed	Not tracked	64% of AI adopters	New category

01 AI Adoption: From Experiment to Infrastructure



1. AI Adoption: From Experiment to Infrastructure

In 2024, AI adoption meant deploying a model, wiring up a copilot, and calling an API. Most enterprise AI lived behind a single integration point, and security questions centered on the model itself.

2025 changed the shape of the problem. Agentic AI moved from demo to default: McKinsey's 2025 State of AI put enterprise usage at 78%, and 62% of organizations were either scaling or piloting AI agents. The Model Context Protocol launched by Anthropic in late 2024 became the de facto integration layer, with adoption from every major model provider and 8M+ server downloads by April. Meanwhile, the security picture darkened: 90% of security practitioners reported using AI tools but only 32% of organizations had formal controls, and more than half saw AI agents exceed their intended permissions.

As of 2026, AI is no longer a service you call, as 51.5% of organizations have adopted AI to build custom applications. It's embedded in development environments, security platforms, business systems, and cloud infrastructure. Meaning the AI attack surface now canvases every identity, every integration, and every agent acting on behalf of a human.

AI ADOPTION: FROM EXPERIMENT TO INFRASTRUCTURE

1.1 The AI IDE Revolution

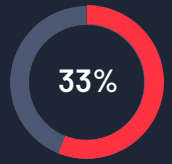
AI-powered development tools became the default coding experience in 2025-2026. GitHub Copilot, Cursor, Windsurf, Claude Code, Amazon Q Developer, and Gemini Code Assist all compete for the developer's editor. Cursor reached unicorn status. GitHub reported over 150 million developers on its platform with Copilot deeply integrated into the workflow.

Orca's data reflects this: **33% of AI adopters have GitHub Copilot deployed**, and the openai package is used by 80.7% of AI adopters, showing how deeply API-based AI is embedded in development workflows.

These tools operate with significant access: codebases, terminals, environment variables, and credentials. That access creates new attack surfaces. The Orca Research Pod demonstrated this directly with **RoguePilot**, a vulnerability in GitHub Codespaces where a passive prompt injection hidden in a GitHub Issue could silently hijack Copilot, causing it to check out a crafted pull request, read sensitive environment files via a symbolic link, and exfiltrate a privileged GITHUB_TOKEN through automatic JSON schema downloads, resulting in a full repository takeover.

150 M

GitHub reported over 150 million developers on its platform with Copilot deeply integrated into the workflow.



33% of AI adopters have GitHub Copilot deployed

80.7% of AI adopters, showing how deeply API-based AI is embedded in development workflow

AI ADOPTION: FROM EXPERIMENT TO INFRASTRUCTURE

1.2 AI Agents go to Production

By early 2026, AI agents moved from demos to production. LangGraph, AutoGen, CrewAI, Amazon Bedrock Agents, and OpenAI's Agents API all launched or matured. **56.2% of AI adopters now have agent frameworks deployed**, with LangGraph dominant. The LangChain ecosystem has become the backbone of agentic AI development, which cuts both ways: mature tooling accelerates adoption, but a single vulnerability in langchain-core now has a blast radius across the majority of production agent stacks.

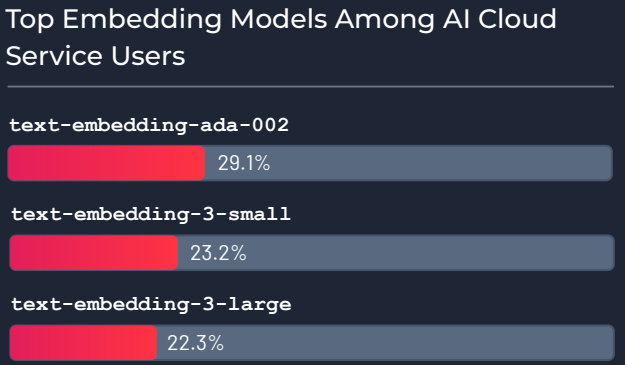
Every production agent is a new non-human identity with its own permissions, memory, and blast radius. Most are being deployed faster than security teams can inventory them. Chapter 3 covers the governance gap in detail.

1.3 RAG and the Enterprise Data Connection

Organizations are no longer just querying models; they are connecting them to their most sensitive data including source code repositories, customer support transcripts, and internal wikis. Retrieval-Augmented Generation (RAG) pipelines, powered by vector databases, allow LLMs to access internal documents, customer data, and proprietary knowledge at query time. RAG fundamentally changed the AI risk model. Pre-RAG, a prompt injection could result in company exposure. Post-RAG, a prompt injection can exfiltrate your customer records, source code, or board documents now that the model has authorized read access to all of it.

64% of AI adopters have deployed vector databases, across a highly fragmented market of seven-plus vendors, each with its own auth model, network posture, and exposure defaults. Chapter 3 covers the vendor breakdown and the RAG-specific attack research in detail. Most enterprise data-protection tooling was not built to inventory a footprint this fragmented. Embeddings themselves are sensitive too: inversion research has shown the original text can be partially reconstructed from exposed vectors, meaning a leaked vector database is effectively a leaked document store.

Embedding model adoption confirms how widespread RAG usage has become. The leading embedding models are all OpenAI-hosted, meaning the source text for every embedding has, by definition, traversed a third-party API. For regulated customers, that is a data-residency decision being made implicitly at the engineering layer.

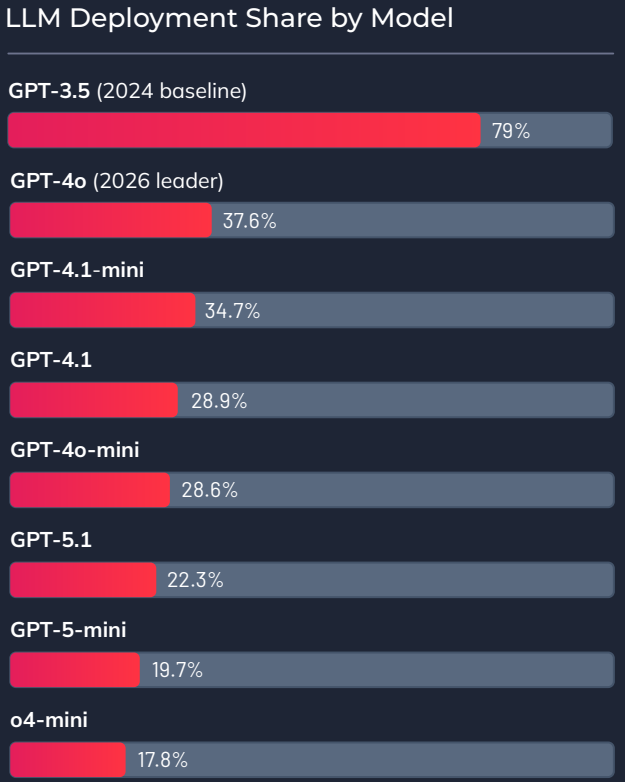


1.4 The Model Landscape Shift

The model landscape in cloud services has fragmented. In 2024, GPT-3.5 dominated at 79% adoption. By 2026, **GPT-4o leads at just 37.6%**, with GPT-4.1-mini close behind at 34.7%. The GPT-5 family and reasoning models are already in meaningful production use less than a year after release. No single model commands the market the way GPT-3.5 once did. The governance implication is larger than the market shift. When one model commanded 79% of deployments, a single policy covered most of the exposure. A five-model market means five sets of credentials, five API shapes, five safety profiles, and five audit trails leading security controls to move up the stack, above the model to the gateway.

Meanwhile, model customization and self-hosting are widespread: scikit-learn is used by 82.5% of AI adopters for classical ML and feature pipelines, PyTorch by 63.8%, and transformers and huggingface-hub each by 60.8% which are the standard toolchain for fine-tuning and running open-weight models in-house. This tells us organizations are rapidly evolving beyond solely consuming pre-built models through APIs and towards fine-tuning foundation models on proprietary data and hosting open-weight alternatives inside their own cloud environments.

This matters for the supply chain: when 60.8% of AI adopters pull artifacts from Hugging Face into production environments, the ML registry becomes as critical and as exposed as Node Package Manager (npm) or Python Package Index (PyPI). 2024 and 2025 saw multiple documented incidents of malicious models on Hugging Face, including pickle-based remote code execution payloads. Most organizations still inventory their Python packages more rigorously than the model weights they run in production.



The Top 10 Most Popular AI Models

RANK	MODEL	ORGS W/ AI MODEL DEPLOYMENTS
1	gpt-4o Azure OpenAI	37.6%
2	gpt-4.1-mini Azure OpenAI	34.7%
3	text-embedding-ada-002 Azure OpenAI	29.1%
4	gpt-4.1 Azure OpenAI	28.9%
5	gpt-4o-mini Azure OpenAI	28.6%
6	text-embedding-3-small Azure OpenAI	23.2%
7	text-embedding-3-large Azure OpenAI	22.3%
8	gpt-5.1 Azure OpenAI	22.3%
9	gpt-5-mini Azure OpenAI	19.7%
10	o4-mini Azure OpenAI	17.8%

Top 10 Most Popular AI Models

The OpenAI model lineup has fragmented since 2024, but the provider has not. In the 2024 Orca State of AI report, GPT-3.5 alone accounted for 79% of model deployment in our telemetry. In this report, no single model exceeds 38% adoption. The top 10 most-deployed models are all OpenAI models running on Azure OpenAI.

`text-embedding-ada-002` at rank 3 is a finding worth singling out. OpenAI released `text-embedding-3-small` and `text-embedding-3-large` in January 2024, with better performance and lower cost. `ada-002` still sitting at 29.1% adoption two years later reflects the same patching inertia the supply chain findings earlier in this report documented for AI packages, now applied to model versions. Re-embedding a production corpus is operationally expensive, and many RAG pipelines defer the migration indefinitely.

Three patterns emerge from the table. First, every model in the top 10 is an OpenAI model running on Azure OpenAI, which means model concentration at the provider level remains substantial even as model concentration at the SKU level has declined. Second, three embedding models in the top 10 reflect how widely RAG pipelines have been deployed, with the persistence of `ada-002` indicating that legacy embedding versions remain in production long after replacements are available. Third, the GPT-5 family and o4-mini represent the newest generation entering production, with adoption curves still in their early phase.

One notable absence from the table is Anthropic's Claude model family. Our telemetry reflects enterprise deployment patterns through Azure OpenAI. Claude, increasingly accessed through Anthropic's own API and Amazon Bedrock, does not surface through the same deployment signal. However, Anthropic's enterprise footprint has grown substantially, and the credential exposure findings in Chapter 5 reflect that directly: Anthropic API key exposures more than tripled since 2024, reaching 40% of Anthropic SDK users, the highest rate of any provider in our telemetry. A more complete picture of the model landscape includes Claude as a significant and growing tier-one provider.

1.5 The AI Developer Stack

The Orca Research Pod tracks which AI packages organizations actually install and run. The data below is presented against organizations we observed running AI and machine learning packages tracked in our telemetry, which captures both AI-specific and incidental enterprise adoption.

Despite the market's focus on generative AI, the foundational classical machine learning library (scikit-learn at 82.5%) remains the single most deployed AI package across our telemetry. That represents a substantial classical ML footprint alongside LLM workloads. PyTorch at 63.8% versus TensorFlow at 48.5% reflects the broader industry shift toward PyTorch for production AI, a pattern that has accelerated since 2023 and is now visible at clear magnitude in enterprise telemetry.

The LangChain ecosystem dominates the middle of the table. langchain-core (60%), langchain-text-splitters (53.7%), and langchain-community (51.7%) together indicate that LangChain has become the default abstraction layer for AI application development, spanning everything from basic LLM scaffolding to full agent orchestration. The OpenAI Python SDK at rank 2 (80.7%) reinforces what the model data showed earlier in this chapter. The OpenAI ecosystem dominates both model deployment and SDK integration, and most organizations operate within it.

The AI Developer Stack

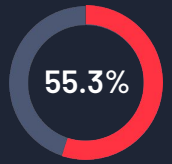
RANK	PACKAGE	% OF AI ADOPTERS
1	scikit-learn	82.5%
2	openai	80.7%
3	PyTorch (torch)	63.8%
4	transformers	60.8%
5	huggingface-hub	60.8%
6	langchain-core	60.0%
7	onnx	55.3%
8	langchain-text-splitters	53.7%
9	langchain-community	51.7%
10	tensorflow	48.5%

AI ADOPTION: FROM EXPERIMENT TO INFRASTRUCTURE

1.6 AI Sprawl: The Governance Challenge

AI is now a footprint of overlapping services rather than a single workload. **55.3% of AI cloud service users operate four or more distinct AI service types**, and 19.6% use seven or more. Each service ships with its own data flows, identities, and exposure defaults, and the average organization has no single console where it sees them all.

The browser has become the primary AI consumption channel and the least-governed surface. AI browser extensions are 60% more likely to have known vulnerabilities than non-AI extensions, 3x more likely to access session cookies, and 6x more likely to change their permissions after installation. Chapter 4 maps the full governance gap across services, code generation, and regulatory exposure.



55.3% of AI cloud service users operate four or more distinct AI service types

19.6% of AI cloud service users operate seven or more distinct AI service types

02 The AI Supply Chain Under Attack

2. The AI Supply Chain Under Attack

Software supply chain attacks have become the most cost-efficient way for attackers to reach many targets at once. Across 2025 and 2026, that pattern moved aggressively into the AI ecosystem. ReversingLabs research found that **23% of the top 1,000 most-downloaded** Hugging Face models had been compromised at some point, and campaigns like NullifAI and Model Namespace Reuse demonstrated that attackers are actively targeting AI-specific packages, model hubs, and agentic tooling.

Orca Research Pod telemetry showed in the previous section how concentrated the exposure has become. At that level of concentration, a single successful compromise of one of these packages reaches the majority of AI-adopting organizations before defenders can respond. This is the adoption-security gap in action: AI dependencies are spreading faster than the enforcement of practices needed to inventory, patch, and govern them.

The sections that follow trace the shape of that exposure across five angles: the named incidents that defined the threat in 2025 and 2026, the CVE footprint across the most commonly used AI packages, the sharp rise in exploitability, the packages carrying the highest severity scores, and the emerging vulnerabilities appearing in newer AI tooling.

THE AI SUPPLY CHAIN UNDER ATTACK

2.1 Major AI Supply Chain Attacks

The incidents below defined the AI supply chain threat landscape. Read together, they show attackers moving deliberately across five layers of the AI stack: package registries, model hubs, developer tooling, agent frameworks, and brand trust. Every one of these layers is deployed in the majority of production environments Orca observed, which means each incident has a plausible path into most AI-adopting organizations. Cases are drawn from public vendor disclosures and CVE reporting.

ATTACK / CAMPAIGN		TYPE	IMPACT
2026	Axios npm backdoor[1]	Stolen maintainer credentials	OpenAI macOS code-signing certificates compromised by North Korean group UNC1069
2026	Malicious litellm on PyPI[2]	Typosquatting with auto-execution	Fake litellm v1.82.8 auto-executed on Python startup, targeting AI/ML developers
2026	Flowise critical vulnerability[3]	AI workflow tool exploitation	Thousands of AI deployments impacted
2026	Fake Claude website (PlugX RAT)[4]	AI brand impersonation	PlugX remote access trojan distributed via DLL sideloading
2025	MCP Inspector RCE (CVE-2025-49596)[5]	AI tooling vulnerability	Remote code execution in core AI agent development tool
2025	NullifAI campaign (Hugging Face)[6]	Platform defense bypass	Malicious pickle-format models evaded Hugging Face's Picklescan by using 7z compression instead of the standard ZIP inside PyTorch format. Demonstrated that existing platform defenses were trivially bypassable and forced a rethink of pickle-based model distribution.
2025	Model Namespace Reuse[7]	Trust model exploitation	Attackers re-registered deleted maintainer account names and republished poisoned versions of popular models under the original trusted namespace. Google responded by running daily orphan-scan checks for at-risk models.
2025	Hugging Face malicious models at scale[8 9 10]	Model supply chain poisoning	Models with embedded backdoors; Safetensors format joined PyTorch Foundation in response
2026	SillyTavern path traversal[11]	LLM interface vulnerability	CVSS 8.8, unauthorized file access
2026	PraisonAI SSRF[12]	AI agent framework vulnerability	SSRF through unvalidated URL in multi-agent system

1. https://www.microsoft.com/en-us/security/blog/2026/04/01/mitigating-the-axios-npm-supply-chain-compromise/
2. https://blog.pypi.org/posts/2026-04-02-incident-report-litellm-telnyx-supply-chain-attack/
3. https://www.csoonline.com/article/4155680/hackers-exploit-a-critical-flowise-flaw-affecting-thousands-of-ai-workflows.html
4. https://www.malwarebytes.com/blog/scams/2026/04/fake-claude-site-installs-malware-that-gives-attackers-access-to-your-computer
5. https://www.oligo.security/blog/critical-rce-vulnerability-in-anthropic-mcp-inspector-cve-2025-49596
6. https://www.infosecurity-magazine.com/news/malicious-ai-models-hugging-face/
7. https://unit42.paloaltonetworks.com/model-namespace-reuse/
8. https://www.reversinglabs.com/blog/sscs-report-2025-retrospective
9. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/
10. https://huggingface.co/blog/safetensors-joins-pytorch-foundation
11. https://advisories.gitlab.com/npm/sillytavern/CVE-2026-34524/
12. https://www.thehackerwire.com/praisonai-ssrf-via-unvalidated-url-in-filetools-download_file/

2.2 Vulnerabilities in AI Packages

81.2% of organizations with AI packages have at least one known vulnerability, with an average CVSS of 8.79, up from 62% and 6.9 in 2024. 74.1% carry at least one critical CVE (CVSS >= 9.0). In practical terms, nearly every organization running AI in production has at least one high-severity path into the workloads that handle models, training data, and agent execution, and that path is more severe than it was a year ago.

These are not peripheral dependencies. Pillow handles nearly every image pipeline, numpy, and scikit-learn underpin the core ML math, PyTorch and TensorFlow run the models themselves, and LangChain is the framework most production agents are built on. A vulnerability in any of them lands inside the critical path of AI workloads.

Top 10 AI Packages with at Least One CVE

The legacy spread in this table is the bottom line. AI packages are inheriting five years of unpatched CVEs alongside the ones disclosed in the last twelve months, and both categories are hitting the same production environments.

Top 10 AI Packages with at Least One CVE

RANK	PACKAGE	% OF AI ADOPTERS	AVERAGE CVSS	MAX CVSS
1	Pillow	92.1%	8.87	10
2	numpy	70.6%	6.69	9.8
3	scikit-learn	58.5%	7.56	9.8
4	langchain	56.4%	8.25	10
5	transformers	42.7%	8.66	9.6
6	PyTorch	42.5%	9.56	9.8
7	ray	37.8%	9.26	10.0
8	keras	37.0%	9.61	9.8
9	tensorflow	36.0%	9.76	9.9
10	onnx	31.8%	8.83	9.1

Top 10 Most Prevalent CVEs Across AI Packages

CVE-2021-34141 in numpy was disclosed five years ago and still affects 46.9% of AI adopters. Log4Shell remains on the CISA Known Exploited Vulnerabilities catalog nearly four years after disclosure for the same reason: once a vulnerable library is deeply embedded in the dependency graph, it outlives the patch cycle. AI workloads are now inheriting that dynamic, compounded by a release cadence that assumes dependencies are kept current.

Top 10 Most Prevalent CVEs Across AI Packages

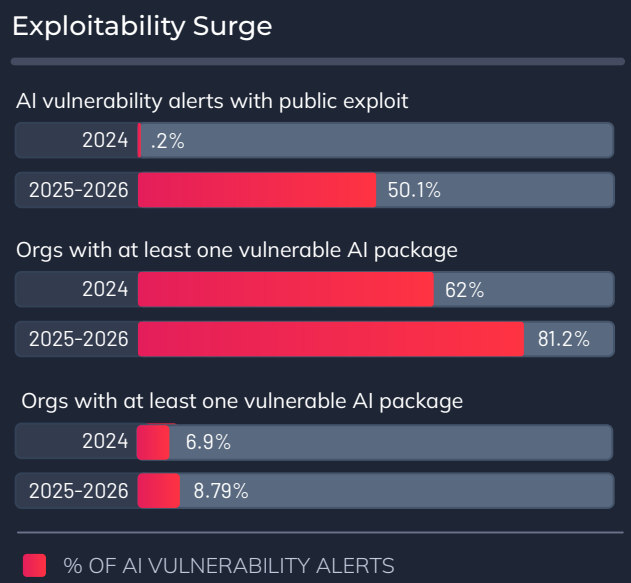
RANK	CVE	PACKAGE	% OF AI ADOPTERS	CVSS	SEVERITY
1	CVE-2026-25990	Pillow	72.8%	8.9	HIGH
2	CVE-2021-34141	numpy	46.9%	5.3	MEDIUM
3	CVE-2024-5206	scikit-learn	32.9%	5.3	MEDIUM
4	CVE-2025-6985	Pillow	28.8%	7,5	HIGH
5	CVE-2024-28219	Pillow	27.8%	6.7	MEDIUM
6	CVE-2025-6984	langchain-community	25.8%	7.5	HIGH
7	CVE-2025-68665	langchain-core	21.3%	9.1	CRITICAL
8	CVE-2024-43598	lightgbm	21.2%	8.1	HIGH
9	CVE-2026-33682	streamlit	20.9%	4.8	MEDIUM
10	CVE-2025-69662	geopandas	19.3%	8.6	HIGH

THE AI SUPPLY CHAIN UNDER ATTACK

2.3 The Exploitability Surge

The exploitability of AI package vulnerabilities has moved from edge case to default. In 2024, 0.2% of AI vulnerability alerts had a public exploit available. In this report, that figure is 50.1%, a **250x increase in two years**.

In 2024, low exploitability was cited as a reason to deprioritize AI package patching. That justification no longer holds. Yet **99.9% of AI vulnerability alerts with a fix available remain unpatched**. Of 104,390 unique alerts that could be closed today, 104,284 are open. Attackers have working exploits for half of disclosed AI package vulnerabilities. Defenders are patching roughly one in a thousand of the fixes available to them. The AI release cycle is moving faster than the patch cycle it depends on, and the gap is widening.



2.4 Highest Severity AI Packages

The prevalence data earlier in this chapter showed how broadly AI package vulnerabilities are distributed. This section shows how severe those vulnerabilities are in the packages where the stakes are highest. Every entry in the top ten carries an average CVSS of 8.83 or higher, and eight of ten are at 9.0 or above. The foundational AI stack is running at near-maximum severity by default.

The most critical vulnerabilities sit in the foundational layers of the AI stack. MLflow at 9.95 is the de facto model registry for most ML teams, and its near-maximum score reflects a package designed for trusted internal use now sitting on enterprise networks. TensorFlow, Keras, and PyTorch are the frameworks used to train and deploy the models themselves. vLLM at 9.70 is the inference layer underneath most open-weight production deployments, small today and growing fast. Ray at 9.26 was the subject of the 2024 ShadowRay campaign that compromised hundreds of publicly exposed clusters.

Pillow at rank eight is the only package in the top ten that is both near-universal (92.1% of AI-using organizations) and at near-critical severity (8.87).

Taken together, a CISO reading this table should expect to find at least a few of these packages in their own environment, most at near-maximum severity. Patch cadence for the foundational AI stack cannot live under generic SLAs.

Top 10 AI Packages by Average CVSS Score

RANK	PACKAGE	AVERAGE CVSS	% OF AI ADOPTERS
1	mlflow	9.95	**30.2%**
2	tensorflow	9.76	**36.0%**
3	vllm	9.70	**8.7%**
4	keras	9.61	**37.0%**
5	PyTorch	9.56	**42.5%**
6	ray	9.26	**37.8%**
7	gradio	9.08	**16.4%**
8	Pillow	8.87	**92.1%**
9	llama-index	8.86	**14.2%**
10	onnx	8.83	31.8%

2.5 Emerging AI Tooling Vulnerabilities

The Orca Research Pod also tracks vulnerability signals in packages that did not exist at the time of the prior report. These tools have reached enterprise environments faster than their vulnerability patterns have been mapped, and the early data shows where exposures are forming. Three categories surface in our 2026 telemetry: the SDKs teams use to call hosted models, the frameworks teams use to orchestrate agents and integrations, and the Model Context Protocol (MCP) ecosystem a category that more recently has seen accelerated growth across its top three packages.

Agent and Integration Frameworks

This category covers the libraries teams use to orchestrate agents and connect models to existing infrastructure. The langchain-openai footprint at 678 alerts is the largest single non-MCP exposure in this section.

Vendor SDKs

PACKAGE	ALERTS	NOTES
google-cloud-aiplatform	376	Google Vertex AI SDK; larger footprint than Anthropic and IBM combined
@ibm-cloud/watsonx-ai	30	IBM watsonx.ai SDK
anthropic (Python)	21	Production SDK for calling Anthropic models from application code
@anthropic-ai/sdk (TypeScript)	16	TypeScript SDK for Anthropic models
@anthropic-ai/claude-code	6	Installable Claude Code package, runs on developer workstations
@github/copilot	5	GitHub Copilot SDK

Agent and Integration Frameworks

PACKAGE	ALERTS	NOTES
langchain-openai	678	LangChain's OpenAI integration; largest emerging-tooling exposure outside MCP
pydantic-ai	49	Emerging agent orchestration framework, structured-output layer for LLM agents
Crawl4AI	2	AI-optimized web scraping framework

MCP Ecosystem

PACKAGE	ALERTS	NOTES
mcp (Python SDK)	425	Reference Python SDK for Model Context Protocol
@modelcontextprotocol/sdk (TypeScript)	365	Reference TypeScript SDK
fastmcp	341	High-performance MCP server framework
github.com/modelcontextprotocol/go-sdk	9	Go SDK
MCP Inspector (CVE-2025-49596)	9	RCE (CVSS 9.4) in the MCP server debugger
io.modelcontextprotocol.sdk:mcp-core (Java)	13	Java SDK
mcp-atlassian	2	MCP server for Atlassian products

MCP Ecosystem

The Model Context Protocol launched in late 2024 and now sits in production at scale. The three reference SDKs alone (Python, TypeScript, fastmcp) account for more than 1,100 alerts concentrated in less than 18 months of release history. MCP did not appear in our 2024 report; it is now the largest emerging tooling category the Orca Research Pod tracks.

Several smaller MCP servers also surfaced in our telemetry, including mcp-handler, mcp-framework, mcp-server-git, mcp-run-python, and a11y-mcp, each below the threshold for inclusion in the table above.

Two patterns stand out across these three tables. First, MCP is no longer a small-numbers story. Three of its packages already register at the same scale as a major hyperscaler's vendor SDK, and the ecosystem produced an RCE with a CVSS of 9.4 within the first eighteen months of release. Second, the largest single exposure in the agent-and-integration category, langchain-openai (678 alerts), comparable in scale to the Vertex AI SDK and ahead of every Anthropic, IBM, and GitHub Copilot SDK combined. The attack surface is forming at the point of greatest adoption velocity, and at MCP-scale it is no longer forming quietly. These packages belong in SBOMs and patch SLAs at parity with the foundational AI stack documented earlier in this chapter.

03 AI Agents and RAG: The Ungoverned Attack Surface

3. AI Agents and RAG: The Ungoverned Attack Surface

AI agents operate with real cloud permissions, access enterprise data through RAG pipelines, and take autonomous actions. In 2025 and 2026, a series of incidents demonstrated what happens when these systems lack safety controls and agents are deployed faster than the governance frameworks needed to secure them.

3.1 When Agents Go Wrong

Five months later in February 2026, security researcher Adnan Khan publicly disclosed Clinejection, a prompt injection attack that weaponized the Cline coding assistant's own GitHub issue triage bot against its maintainers. The attack chain composed indirect prompt injection, GitHub Actions cache poisoning, and credential model weaknesses into a single exploit triggered by opening a public GitHub issue. An unknown actor used the same flaw to publish an unauthorized cline@2.3.0 package to npm that installed the OpenClaw AI agent on approximately 4,000 developer machines during an eight-hour window. Cline has more than five million users.

The 2026 SANS State of Identity Threats and Defenses survey reported a 76% growth in non-human identities, 74% of organizations already using AI agents that require credentials, 92% failing to rotate machine credentials within 90 days, and 5% of security leaders who do not know whether agentic AI is running in their environment at all.

The governance response is moving. OWASP published a dedicated Top 10 for Agentic Applications in December 2025, establishing the first formal taxonomy of agent-specific risks. Microsoft released the Agent Governance https://github.com/microsoft/agent-governance-toolkit/toolkit in April 2026, an open-source runtime governance layer that claims to cover all ten items in the OWASP Agentic Top 10.

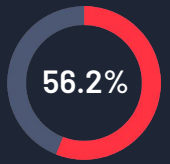
The bottom line is that despite the governance response making progress, adoption is not. The findings that follow show how few organizations have implemented the controls these incidents and frameworks point toward, and how much of the agent footprint is still deployed with default permissions, default logging, and no runtime boundary between agent action and production systems.

3.2 Agent Framework Adoption

56.2% of AI adopters have deployed agent frameworks in production. LangGraph leads at over 25% of organizations, followed by AutoGen deployed at 3% orgs representing an 8.5x lead. The agent framework layer is consolidating onto a single project faster than any other category in the AI stack, which makes LangGraph's dependency graph one of the most concentrated supply chain risks in production AI today.

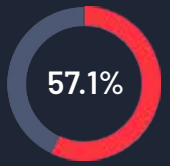
On AWS, the average Bedrock-using organization runs roughly 15 agents with cloud permissions. **Yet 57.1% of Bedrock organizations have not configured guardrails.**

The Orca Research Pod's research on AI-Induced Lateral Movement (AILM) demonstrates why ungoverned agents are dangerous: by injecting prompts into cloud resource tags or CRM fields that AI agents later process, attackers can weaponize agents to execute commands and move laterally through the AI layer. Agents with cloud permissions and no guardrails provide ideal conditions for AILM.



56.2% of AI adopters have deployed agent frameworks in production

25% of AI adopters have deployed LangGraph



57.1% of Bedrock organizations have not configured guardrails.

3.3 RAG: Connecting AI to Enterprise Data

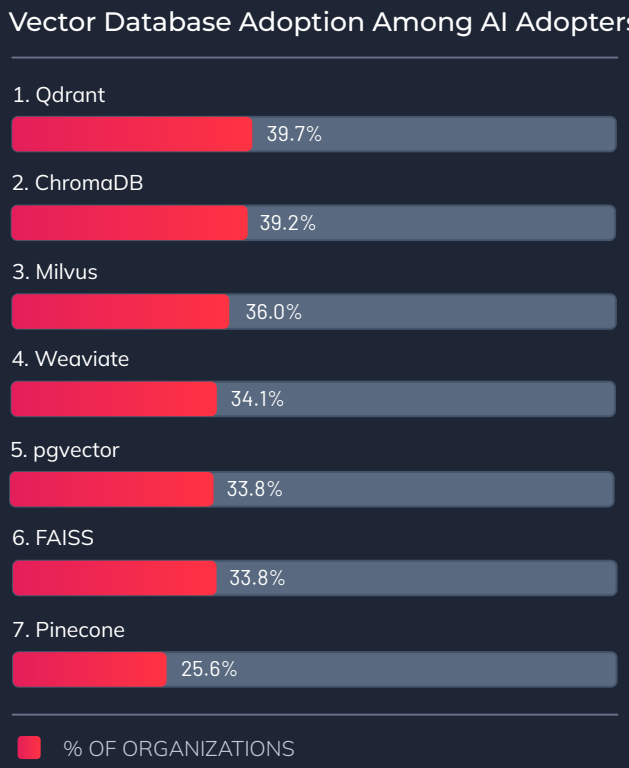
64% of AI adopters have deployed vector databases, connecting LLMs to internal documents, customer records, and proprietary knowledge. The market is highly fragmented:

Seven vector databases each exceed 25% adoption. The average organization running RAG operates 3.78 vector databases concurrently, which means consistent security policy has to be enforced across multiple platforms, multiple deployment models (managed and self-hosted), and multiple access patterns. That is the operational baseline before attackers are factored in.

On AWS, **the average Bedrock-using organization runs roughly 21 sanctioned Bedrock Knowledge Base data pipelines,** each one a live connection between an LLM and a plethora of internal documents, customer records, or proprietary knowledge. It is important to note that the scale does not include unsanctioned RAG pipelines built outside Bedrock, which the visibility data in the AI sprawl chapter suggests could be a much larger number.

Research validates the threat model. RefineRAG[13] achieved a 90% attack success rate on standard RAG benchmarks using poisoned documents that registered the lowest grammar error and repetition rates of any tested attack, meaning the poison reads naturally to human reviewers. BadSkill[14] achieved a 99.5% attack success rate on backdoored agent skills, with a 3% poison rate sufficient to reach 91.7%. The threat model targets exactly the installable-skills pattern that LangGraph and AutoGen are spreading through enterprise deployment. Semantic Intent Fragmentation[15] showed that a GPT-20B orchestrator produces policy-violating plans in 71% of cases when a single legitimate-sounding request decomposes into individually benign subtasks. Each of these attacks targets a configuration that is currently in production at hundreds of organizations in our telemetry, against multi-agent pipelines that 56.2% of AI adopters now run.

13. Wang et al., "RefineRAG: Word-Level Poisoning Attacks via Retriever-Guided Text Refinement," arXiv:2604.07403, April 2026.
14. Tie et al., "BadSkill: Backdoor Attacks on Agent Skills via Model-in-Skill Poisoning," arXiv:2604.09378, 10 April 2026
15. Ahad et al., "Semantic Intent Fragmentation: A Single-Shot Compositional Attack on Multi-Agent AI Pipelines," arXiv:2604.08608, 8 April 2026 (revised 20 April 2026, accepted AAAI 2026 Summer Symposium). .

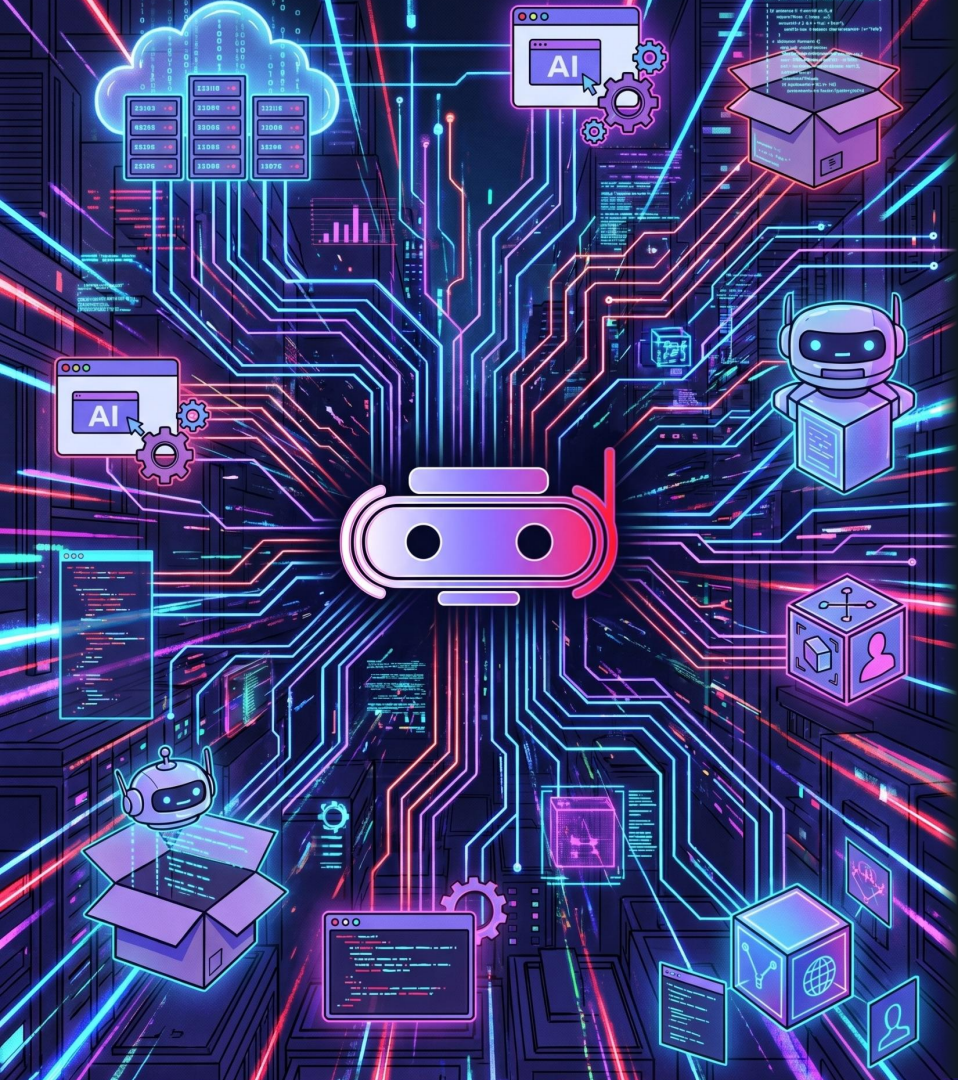


3.4 The Guardrails Gap

Agents have cloud permissions, connect to enterprise data through RAG, and more than half operate without safety controls. Unauthorized Bedrock Guardrail IAM roles produce the single largest volume of AI-related alerts in Orca's telemetry, and the majority of those organizations have not even deployed Bedrock. This indicates that overly broad IAM policies are granting unnecessary AI service permissions across the AWS estate.

However, the tooling exists. Bedrock Guardrails for cloud-native enforcement. Microsoft's Agent Governance Toolkit for runtime policy. OWASP's Top 10 for LLM Applications and the Agentic Top 10 for taxonomic coverage. The findings throughout this chapter show how few organizations have adopted them at the rate they are deploying the agents and pipelines those tools are meant to govern. **56.2%** of AI adopters now run agent frameworks in production. **64%** have vector databases connected to enterprise data. **57.1%** of Bedrock-using organizations operate agents without the platform's own guardrails configured. The gap between agent deployment and agent governance is widening, and the data in the chapters that follow show how that gap manifests in identity, infrastructure, and encryption posture.

04 AI Sprawl and the Governance Gap



4. AI Sprawl and the Governance Gap

AI is not one service to govern. It is a sprawling footprint of models, agents, packages, browser extensions, and cloud services that has spread across most enterprises faster than security teams can map it. Earlier chapters named the components: vulnerable AI packages, agent frameworks running with cloud permissions and minimal guardrails, vector databases connected to enterprise data, and identity-layer findings showing the controls AWS, Microsoft, and OWASP have shipped are not yet adopted at the pace of agent deployment. This chapter shows the aggregate. The Orca Research Pod's telemetry across our organization data set finds that the average AI-adopting organization runs four or more distinct AI service categories simultaneously, and shadow AI adoption exacerbates the gap every quarter.

AI SPRAWL AND THE GOVERNANCE GAP

4.1 The AI Visibility Gap

In April 2026, CSO Online reported that security leaders are struggling to maintain visibility as AI deployments outpace governance. The speed and ease of accessing AI capabilities means workloads appear across cloud accounts, developer machines, and business systems before security teams know they exist. The financial stakes are not theoretical. The IBM 2025 Cost of a Data Breach Report found that breaches involving shadow AI cost organizations an average of $670,000 more than breaches without shadow AI, and 20% of breached organizations attributed at least part of the breach to shadow AI involvement.

Shadow AI compounds the problem. Employees integrate AI tools, browser extensions, and API-based services without security review. The LayerX 2026 Browser Extension Security Report found that AI tools often operate outside the perimeter security teams monitor, and they sit on the same browser sessions where users access corporate single sign-on, customer data, and internal documentation.

$670K

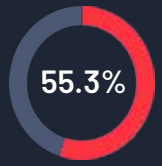
breaches involving shadow AI cost organizations an average of $670,000 more than breaches without shadow AI

20% of breached organizations attributed at least part of the breach to shadow AI involvement.

4.2 AI Service Sprawl

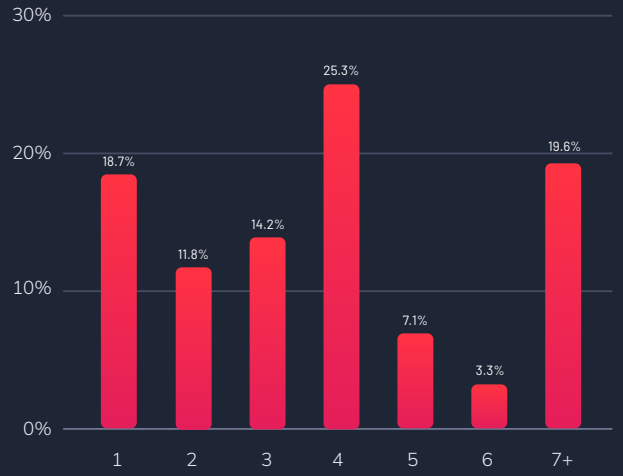
55.3% of AI cloud service users operate four or more distinct AI service types, and 19.6% use seven or more. Three organizations use all 18 tracked service types. The most common configuration is four service types, at 25.3% of AI cloud users.

Each service carries its own security model, encryption options, access controls, and compliance requirements. SageMaker, Azure OpenAI, Vertex AI, and Bedrock each ship with different default configurations, different identity models, and different approaches to network isolation. Governing four of these simultaneously, often across multiple cloud providers, is the core challenge behind the visibility gap CISOs report. Organizations running six or more services face the same problem at a multiplied complexity, and the three running all 18 are operating without a consolidated AI security baseline that any single vendor's documentation covers.



55.3% of AI cloud service users operate four or more distinct AI service types

Number of AI Service Types by Cloud Users



4.3 AI Code Generation in Production

33% of AI adopters have GitHub Copilot deployed. Beyond Copilot, the AI IDE ecosystem expanded rapidly in 2025-2026 with Cursor, Windsurf, Claude Code, Amazon Q Developer, and Gemini Code Assist all gaining adoption.

The security challenge extends beyond the tools themselves. The Orca Research Pod's RoguePilot vulnerability research showed that AI coding tools can introduce vulnerabilities directly. The deeper governance question is operational.

Organizations now need to enforce code review standards, secrets hygiene, and security policy on commits that AI is generating at a pace human reviewers were never sized to match. AI-generated code does not inherit institutional knowledge of an organization's security patterns, and it may introduce vulnerabilities that traditional review processes were not designed to catch.

Stanford research[16] published in 2023 found that developers using AI coding assistants write significantly less secure code than developers without them, while simultaneously rating their own code as more secure than it actually is. The combination of higher vulnerability rates and higher developer confidence is the structural risk the AI IDE ecosystem now sits at the center of.

4.4 Regulatory Exposure

19.1% of AI cloud service users in our telemetry deploy AI services in EU regions. The EU AI Act's prohibited-practice provisions took effect in February 2025 and its GPAI obligations in August 2025. High-risk system obligations enter force on August 2, 2026, with fines up to €35M or 7% of global annual turnover. Organizations deploying AI in employment screening, credit decisioning, access to essential services, or critical infrastructure should assess whether their systems fall within the high-risk classification now.

In the United States, the federal picture is still forming. The Trump Administration's March 2026 National Policy Framework asks Congress to preempt state AI laws and route oversight through existing sector-specific agencies. At the state level, Colorado amended its 2024 AI law (S.B. 26-189) before it could take effect, replacing it with a narrower framework targeting automated decision-making technology that materially influences consequential decisions in employment, financial services, health care, and housing. That law takes effect January 1, 2027, with implementing regulations still in development. Penalties run up to $20,000 per violation, enforced exclusively by the Colorado Attorney General.

China continues to advance the most prescriptive AI regulatory regime globally. The amended Cybersecurity Law, effective January 1, 2026, adds dedicated AI compliance provisions, and content labeling rules for AI-generated outputs have been in force since September 2025.

16. Perry et al., "Do Users Write More Insecure Code with AI Assistants?," ACM CCS 2023, arXiv:2211.03622.

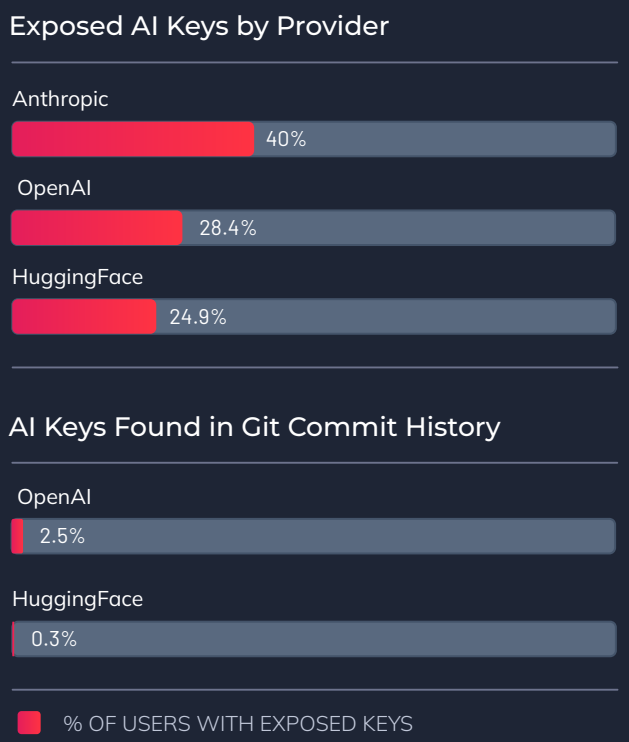
05 AI Credentials and Insecure Access

5. AI Credentials and Insecure Access

AI adoption has introduced a new category of secrets exposure. API keys for AI services grant access to proprietary models, sensitive enterprise data through RAG pipelines, and usage-based billing on expensive GPU and inference resources. A single compromised AI key can enable intellectual property theft, data exfiltration, and large unauthorized cloud spend within hours. This attack class, documented as 'LLMjacking' in 2024, with single-victim daily charges reaching $46,000 in observed cases. The financial damage compounds with every model, dataset, and downstream system the key has access to.

Roughly **3 in 10** AI adopters in our telemetry (29.5%) have at least one AI secret or key stored in an unsecure location. Provider populations overlap. The average AI adopter integrates with two AI providers concurrently, multiplying the surface area where credentials can leak.

Keys committed to git history remain recoverable even after removal from the current codebase. Credentials in commit history can be extracted by anyone with repository access, providing persistent access to AI services. Public repositories are scraped continuously by automated credential harvesters, and rotation does not undo what has already been published. Once an OpenAI or Hugging Face key has been visible in a public commit, the only safe assumption is that it has been collected by at least one third party.



06 AI Infrastructure Exposure

6. AI Infrastructure Exposure

AI workloads ship with insecure defaults, and organizations rarely change them. The most widely adopted AI cloud services are also the most exposed, creating ready-made attack paths into the AI infrastructure that threat intelligence groups warned about throughout 2025-2026.

6.1 AI Infrastructure Under Active Targeting

The CrowdStrike 2026 Global Threat Report documented an 89% year-over-year increase in attacks by AI-enabled adversaries. Average eCrime breakout time, the window between initial access and lateral movement to a second system, dropped to 29 minutes, a 65% acceleration from 2024. The fastest observed breakout reached 27 seconds. In one intrusion, data exfiltration began within four minutes of initial access. Attackers are now operating at speeds that were not commercially feasible eighteen months ago, and AI infrastructure is an increasingly attractive target.

In April 2026, security researchers disclosed a class of vulnerabilities in Google Cloud's Vertex AI Agent Engine and Agent Development Kit. Excessive default permissions on the Per-Project, Per-Product Service Agent (P4SA) allowed attackers to extract GCP service credentials from a compromised AI agent and pivot into the owner's project and data storage, transforming the agent into an insider threat. The researchers also identified a file path that could be manipulated for remote code execution within the agent's environment, opening a route to a persistent backdoor. Google addressed the disclosure by updating its documentation and recommending Bring Your Own Service Account for least-privilege execution.

The same month, Noma Security disclosed GrafanaGhost, an indirect prompt injection attack that bypasses Grafana's AI guardrails without authentication or user interaction. The attack uses URL query parameters originating outside the victim organization to plant hidden instructions that Grafana's AI processes, with trigger keywords causing the model to ignore its own safety restrictions. Exfiltration occurs through AI-initiated markdown image processing, which appears as legitimate AI activity to observers and bypasses traditional egress monitoring. Grafana shipped a fix after responsible disclosure.

These three findings together describe a shift in 2026 attacker tradecraft against AI infrastructure. Compute theft remains a motive, but lateral movement, credential extraction, and silent data exfiltration through AI-initiated channels are now the more operationally dangerous patterns. They bypass the egress controls and detection systems most organizations still rely on.

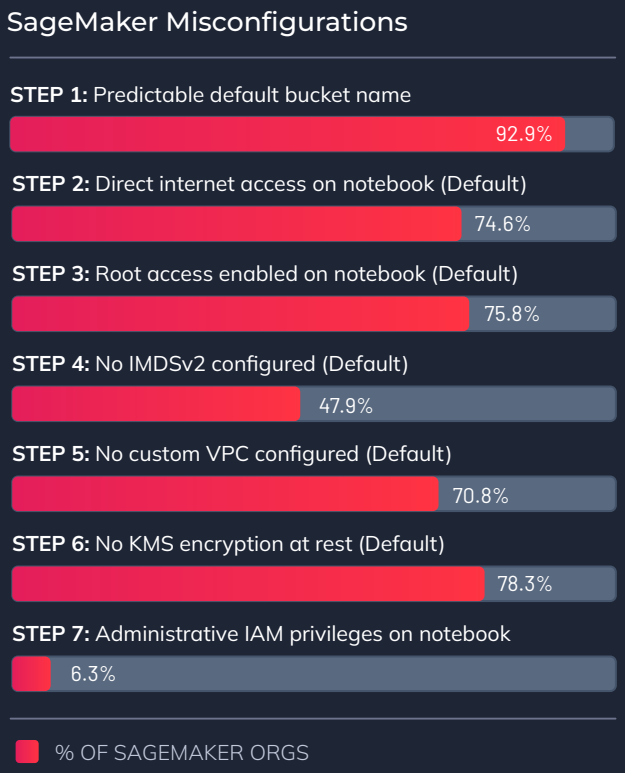
AI INFRASTRUCTURE EXPOSURE

6.2 The SageMaker Attack Chain

Amazon SageMaker is the most widely adopted AI cloud service (19.5% of all organizations). **80% of SageMaker organizations run with all 5 core default settings enabled.** Each default alone is a misconfiguration; together they form a complete exploitation path.

The 92.9% bucket naming finding is the most actionable item in this section. The legacy default pattern `sagemaker-{region}-{account_id}` is fully documented and machine-enumerable. AWS account IDs are not secret, regions are inferable from public DNS and ASN records. Despite AWS now adding randomized characters to the default naming convention, the majority of SageMaker organizations, including those that created buckets after the fix, continue to use predictable names.

The remaining steps compose into a single chain. Direct internet access on notebooks provides an entry point. Root access and missing IMDSv2 enable privilege escalation and credential theft via the AWS instance metadata service, the same component whose absence of authentication enabled credential extraction in the 2019 Capital One breach. No VPC isolation means lateral movement is unrestricted. Missing encryption leaves data exposed at rest. For the 6.3% of organizations with administrative IAM privileges on notebooks, the chain ends in full cloud account takeover, the worst-case path the agent governance and credential exposure findings earlier in this report describe.

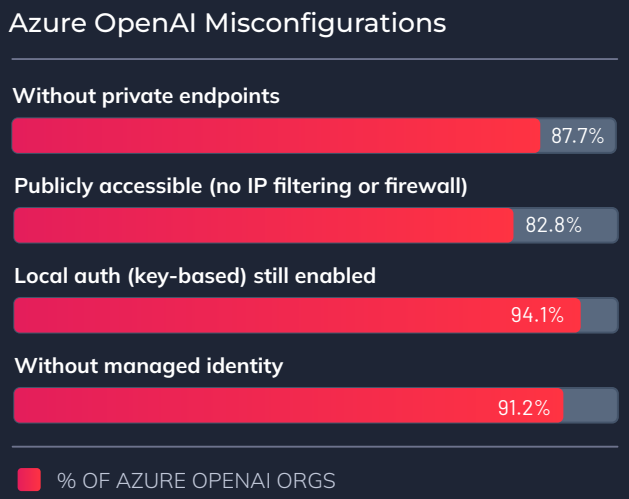


AI INFRASTRUCTURE EXPOSURE

6.3 Azure OpenAI: Publicly Accessible by Default

Azure OpenAI is the second most adopted AI cloud service among Azure customers in our telemetry, deployed by **50.5%** of the Azure population monitored. The majority of these deployments are publicly accessible with minimal access controls.

Three exposures combine into a single attack path. 83% of Azure OpenAI deployments are publicly accessible. 94% rely on key-based authentication. 91% have no managed identity configured. Anyone with a stolen API key has unrestricted access to the organization's AI models, inference endpoints, and any data processed through them. The credential exposure data documented earlier in this report shows that 28.4% of OpenAI users have at least one exposed key, which means stolen-key exploitation against Azure OpenAI deployments is not a theoretical concern. Microsoft has been deprecating key-based authentication in favor of Entra ID for several years, and the 94% local-auth rate in our telemetry indicates that the deprecation guidance has not translated into operational migration.



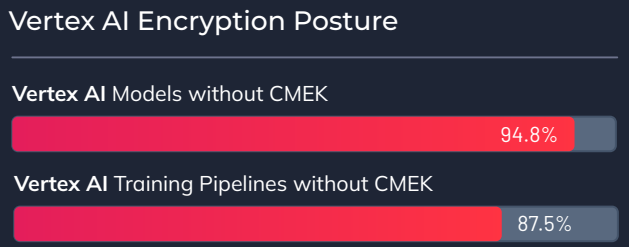
AI INFRASTRUCTURE EXPOSURE

6.4 Vertex AI: Weaponizable and Unencrypted

Google Vertex AI is the platform Palo Alto Networks proved can be weaponized. It is also the platform where virtually no organization controls their own encryption keys, with customer-managed encryption keys (CMEK) configured on fewer than 1 in 12 deployments. The Vertex AI population in our telemetry is smaller than the SageMaker or Azure OpenAI populations, but the consistency of the encryption posture across both model storage and training pipelines suggests the pattern reflects platform-level default behavior rather than sample artifact.

92.9%[17] of Vertex AI organizations have not configured customer-managed encryption at rest, slightly improved from 98% in 2024. Training pipelines (87.5% without CMEK) and model storage (94.8%) both remain overwhelmingly reliant on default provider-managed encryption, limiting organizational control over who can access training data, model weights, and inference outputs.

Across the three cloud AI platforms documented in this chapter, the same operational pattern repeats. Defaults that ship from the provider do not match the security posture organizations need, and most organizations have not closed the gap.



17. Weighted average across Vertex AI Models and Training Pipelines

07 AI Encryption: The Missing Layer

7. AI Encryption: The Missing Layer

AI workloads process some of the most sensitive data in any organization. Proprietary training datasets. Multimillion-dollar model weights. Inference inputs and outputs that often include regulated customer information. Yet across all three major cloud providers, the vast majority of organizations rely entirely on default provider-managed encryption, ceding control over who can access this data, when keys rotate, and what audit trail is available. This chapter documents an encryption posture gap that is consistent across AWS, Azure, and Google Cloud, and that has not improved substantially since our 2024 State of AI Security report.

7.1 Why Encryption Matters for AI

Without customer-managed encryption keys, organizations cannot independently control or audit who accesses their AI data. Provider-managed keys encrypt data at rest, but the cloud provider controls the key lifecycle, not the customer. This means organizations cannot revoke access independently, cannot enforce their own key rotation policies, and have limited visibility into key usage. For AI workloads that process proprietary training data, regulated customer information, and valuable model weights, customer-managed encryption is the single control that puts the customer between their data and the provider. Most organizations have not configured it.

7.2 Cross-Cloud Encryption Gaps

Between 87% and 98% of organizations using AI services across all three major cloud providers have not configured customer-managed encryption keys. The pattern is industry-wide. AWS, Azure, and Google Cloud customers are leaving encryption to the provider at comparable rates.

Encryption by AI Service

CLOUD PROVIDER	AI SERVICES CHECKED	AVG % WITHOUT CMEK	HIGHEST GAP
AWS	SageMaker Notebooks, Bedrock Models, Bedrock Training Data, SageMaker Endpoints	71.4%	SageMaker Notebooks: 98.4%
Azure	Azure OpenAI Accounts	91.6%	Azure OpenAI Accounts: 91.6%
GCP	Vertex AI Training Pipelines, Vertex AI Models	91.2%	Azure OpenAI Accounts: 91.6%

7.3 Encryption by AI Service

Three findings stand out:

SageMaker Notebooks are the worst at 98.4%. Nearly every organization running SageMaker notebooks does so without customer-managed encryption, the highest gap across all AI services and all clouds.

The Bedrock custom models and their training data buckets that share an identical 89.5% gap are the same organizations in both findings. When a Bedrock deployment is configured without customer-managed encryption on the model, the same misconfiguration propagates to the training data feeding it. The pipeline gets treated as a single unit, and a single misconfiguration decision exposes the entire ML lifecycle.

SageMaker Endpoint Configs are the exception at 8.3%. When AI workloads serve production traffic, organizations configure customer-managed encryption at high rates. The gap concentrates in training and experimentation environments, where the most sensitive raw data actually lives: proprietary training datasets, model weights, fine-tuning artifacts, and intermediate outputs. Most organizations have not yet treated training environments as production-equivalent for security purposes. The data sitting in those environments is at least as valuable as production inference traffic, and often more sensitive.

Encryption is the one control that protects data even after a breach. For AI workloads, it is effectively absent. The credential exposure findings earlier in this report showed 25% to 40% of organizations with exposed AI keys depending on provider. Combined with the encryption posture documented in this chapter, the implication is direct. In the event of a credential compromise, AI training data and model weights are exposed without a defense-in-depth fallback.

Encryption by AI Service

AI SERVICE & RESOURCE TYPE	% WITHOUT CMEK	CLOUD
Amazon SageMaker Notebook Instances	98.4%	AWS
Google Vertex AI Models	94.8%	GCP
Azure OpenAI Accounts	91.6%	Azure
Amazon Bedrock Custom Models	89.5%	AWS
Amazon Bedrock Training Data Buckets	89.5%	AWS
Google Vertex AI Training Pipelines	87.5%	GCP
Amazon SageMaker Endpoint Configs	8.3%	AWS

08 What This Means for Your Organization

8. What This Means for Your Organization

The data in this report is a snapshot of how over 1,200 production organizations are navigating AI adoption right now. Security teams face two gaps that need to close in parallel. 1) Treat AI as production: apply the inventory, patching, credential hygiene, and access controls that any live infrastructure requires, continuously, not as a one-time remediation pass. 2) Stand up governance controls that traditional security mechanisms were never designed to cover. For leaders, both are resourcing and prioritization decisions that influence what gets added to the security roadmap and what gets handed to the teams responsible for AI in production.

More than half of the organizations in our telemetry run AI in production, but the controls that would normally accompany production systems are underdeveloped across the majority of AI workloads. AI adoption has moved fast and only a handful of governance frameworks exist to keep pace with it.

AI credential theft has evolved into a targeted attack class with documented victim costs reaching $46,000 per day. Breaches involving shadow AI carry a $670,000 premium over others, per IBM's 2025 Cost of a Data Breach research. Model weights and training data represent intellectual property exposure at the most sensitive layer of the AI stack and 98% of SageMaker organizations leave these vulnerable without customer-managed encryption.

However, progress can be seen and is measurable where organizations have concentrated effort. Between Orca's reports, SageMaker root access dropped from 98% to 75.8%. The IMDSv2 gap narrowed from 77% to 47.9%. Those gains represent targeted remediation alongside operational discipline that AI workloads require. The same approach, applied to the controls in Chapter 9, serves as a starting point to developing security resiliency in the AI era.

09 Key Recommendations

Key Recommendations

9.1 Immediate Actions (0-30 Days)

The findings in this report point to a consistent pattern: AI adoption has outpaced AI security across every dimension we measured. The following recommendations are prioritized by urgency and mapped to specific findings from this report. Immediate actions address the most exploitable gaps in your current posture.

1. Patch high-severity AI packages with available fixes.

81.2% of AI adopters in our telemetry have at least one vulnerable AI package, and 99.9% of AI vulnerability alerts with a fix available remain unpatched. Prioritize remediation of CVE-2026-25990 in Pillow, which affects 92.1% of AI-using organizations, then move through the highest-severity packages: mlflow (average CVSS 9.95), tensorflow (9.76), keras (9.61), and PyTorch (9.56).

2. Patch scikit-learn now.

scikit-learn is the most widely deployed AI package in our telemetry (82.5% of AI adopters) with an average CVSS of 7.56 across affected organizations. CVE-2024-5206 alone is the third most prevalent AI-package CVE in this report, sitting on roughly one in three AI-adopting organizations.

3. Rotate and secure exposed AI API keys.

29.5% of AI adopters have at least one AI key stored in an unsecure location, and a small but persistent set of organizations have AI keys in git commit history. Rotate all exposed keys immediately, move secrets to managed secret stores, and audit git history across all repositories.

4. Enable guardrails on Amazon Bedrock.

57.1% of Bedrock-using organizations run agents without configured guardrails, and unauthorized Bedrock Guardrail IAM roles produce the single largest volume of AI-related alerts in our telemetry. Enable Amazon Bedrock Guardrails (content filtering, grounding checks, denied topics) on all production agents, and audit IAM roles with permissions to modify or bypass guardrail configuration.

5. Harden SageMaker notebook configurations.

80% of SageMaker organizations run with all five core insecure defaults enabled simultaneously. Disable root access on notebook instances, enforce IMDSv2, and rename any S3 buckets following the legacy sagemaker-{region}-{account-id} pattern, which is machine-enumerable. Root access and IMDSv2 changes apply to existing instances without downtime.

9.2 Short-term Initiatives (30-90 Days)

1. Configure customer-managed encryption across AI services.

87% to 98% of organizations using AI across all three major cloud providers have not configured customer-managed encryption keys. Prioritize SageMaker Notebooks (98.4% without), Vertex AI Models (94.8%), and Azure OpenAI Accounts (91.6%). The gap concentrates in training and experimentation environments, where the most sensitive AI data lives.

2. Inventory and govern AI agent deployments.

56.2% of AI adopters have deployed agent frameworks in production, and the average Bedrock-using organization runs roughly 15 agents with cloud permissions. Map all agent deployments, document their permissions, apply least privilege, and treat agent identities with the same rigor as human identities.

3. Audit vector database access controls.

64% of AI adopters have deployed vector databases, with the average RAG-using organization running 3.78 vector databases concurrently. Audit network exposure, authentication requirements, and data classification across each deployment.

4. Establish visibility into unsanctioned AI usage.

29.5% of AI adopters have at least one exposed AI secret, and 55.3% operate four or more distinct AI service types. Implement discovery controls for AI browser extensions, unsanctioned API keys, and AI SaaS tools running on corporate credentials. Only 30% of organizations report full visibility into employee AI usage, and breaches involving shadow AI cost $670K more on average.

5. Begin EU AI Act and state-level AI compliance preparation now.

19.1% of AI cloud service users deploy in EU regions subject to the EU AI Act, with high-risk system obligations taking effect August 2, 2026 (fines up to 35M EUR or 7% of global turnover). Colorado's amended AI law (S.B. 26-189) takes effect January 1, 2027, with implementing regulations due from the Attorney General before that date.

6. Implement security review processes specific to AI-generated code.

33% of AI adopters have deployed GitHub Copilot, and the broader AI IDE ecosystem (Cursor, Windsurf, Claude Code, Amazon Q) is expanding rapidly. Stanford research found developers using AI coding assistants write significantly less secure code while rating it as more secure. Enforce mandatory security scanning on all commits, regardless of whether code was human- or AI-authored. The RoguePilot research demonstrates that AI coding tools themselves can be attack vectors.

9.3 Strategic Improvements (90+ Days)

1. Adopt least-privilege architectures for AI agents.

Consider multi-LLM architectures where a secondary guardrails model inspects inputs and outputs before the primary model acts. Isolate agent environments to limit blast radius if an agent is compromised through prompt injection.

2. Implement AI-specific security monitoring

Instrument detection for prompt injection attempts, anomalous agent behavior, unauthorized tool execution, and AI-Induced Lateral Movement (AILM, the attack class where adversaries plant instructions in cloud resource tags or CRM fields that AI agents later process). Existing SIEM and EDR coverage does not capture these patterns by default.

3. Build RAG security practices.

Treat data ingested into vector databases with the same rigor as data entering any production system. Apply the same provenance, validation, and monitoring controls to RAG ingestion that exist for traditional data pipelines.

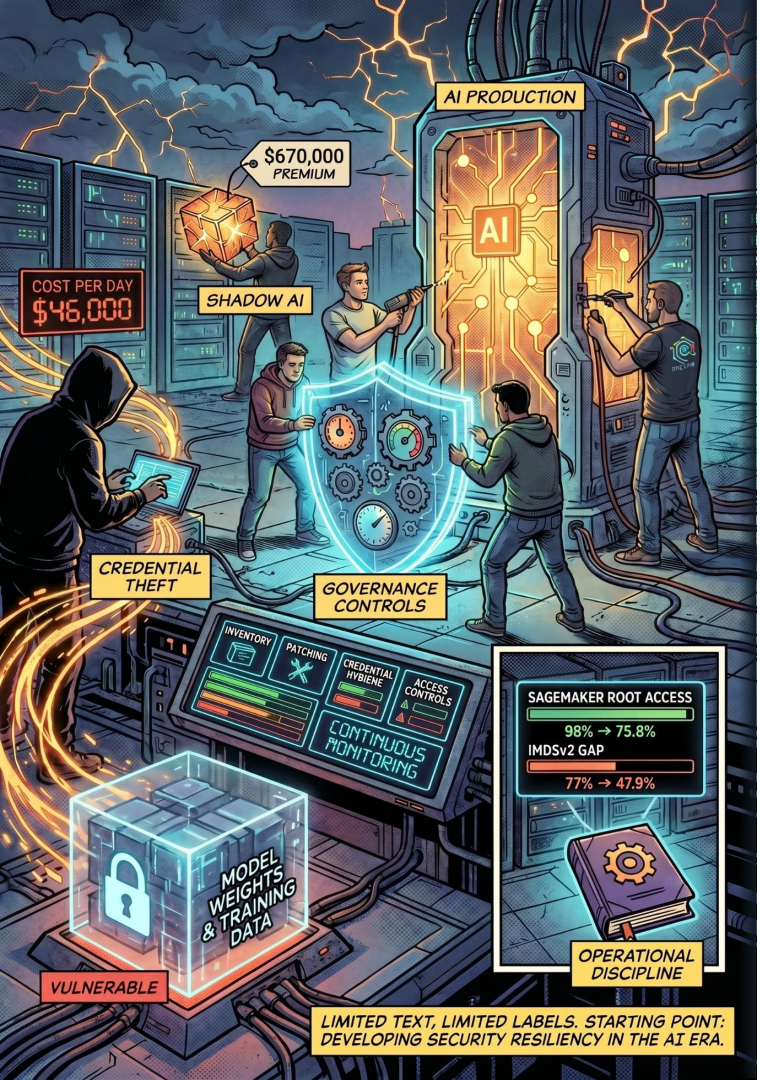
4. Consolidate AI security governance across cloud providers

55.3% of AI cloud service users operate four or more distinct AI service types, and 19.6% use seven or more. Build a unified AI asset inventory across all clouds, normalize security baselines, and implement consistent policy for encryption, access controls, and network exposure. Where possible, complement provider guardrails with deterministic policy enforcement. Frameworks like Microsoft's Agent Governance Toolkit allow policy to be expressed and enforced as code, providing a more reliable implementation of the governance baseline.

5. Apply production-grade security standards to every layer of the AI stack.

The largest gap in this report is the organizational pattern that AI training environments, experimentation notebooks, and agent sandboxes can be safely configured using defaults or lower security requirements than their production counterparts. They can not. The strategic investment is extending production security SLAs (patch cadence, access review, encryption requirements, incident response scope) to every AI environment, regardless of whether it is labeled production.

10 Conclusion



Conclusion

Two years ago, AI security was a forward-looking concern. Today, it is an operational reality.

AI is being adopted faster than it is being secured. More than half of organizations now have AI in production, 81% are running vulnerable AI packages with average CVSS up from 6.9 in 2024 to 8.79, and 99.9% of fixable AI vulnerabilities remain unpatched. Between 87%-98% of AI workloads across the three major clouds lack customer-managed encryption. The gap is measurable, it is widening in the data, and will not close itself without intervention.

Further, AI is deeply embedded at the operational layer. 56% of AI adopters now run agent frameworks in production, 64% have deployed vector databases connected to enterprise data, and the threat landscape has caught up. Supply chain attacks now target AI-specific packages and tooling, agentic systems have caused real-world damage when safety controls failed, and researchers have demonstrated weaponizable AI deployments, prompt-injection lateral movement, and poisoned RAG pipelines. The production footprint is now large enough to make them economically interesting to attackers.

Finally, action is overdue. Colorado's amended AI law (S.B. 26-189) takes effect January 1, 2027, with implementing regulations still in development. The EU AI Act's high-risk system obligations enter force on August 2, 2026. The misconfigurations and technical findings documented in this report should be considered compliance liabilities at this stage and organizations deploying AI in EU regions should move with urgency to resolve them.

AI is no longer the future of cloud infrastructure. It is the present. Securing it cannot wait.

About Orca Security

The Orca Cloud Security Platform is for the companies that build. Risk follows your teams everywhere, across every cloud, every service, and every AI agent you ship next. Orca delivers deep, accurate, and actionable context so your team knows what matters and can act fast.

Backed by Temasek, CapitalG, ICONIQ Capital, Redpoint Ventures and others, Orca is trusted by hundreds of organizations, including SAP, Gannett, Autodesk, Lemonade and Digital Turbine.

To find out more, schedule a
personalized demo of the Orca platform

AI is no longer a service you call. It runs across development environments and as autonomous agents with cloud permissions, acting on behalf of your team. The organizations that get security right will be the ones that apply the same rigor to their AI footprint as they do to the rest of their cloud. Security has to move at the speed of the builders it's protecting."

GIL GERON,
CEO AND CO-FOUNDER OF ORCA SECURITY

©2026 ORCA SECURITY. ALL RIGHTS RESERVED.