Zscaler ThreatLabz 2024 AI Security Report

The AI revolution has arrived. Discover key trends, risks, and best practices in enterprise AI adoption, with insights into AI-driven threats and key strategies to defend against them.

Contents

Executive Summary

AI is more than a pioneering innovation—it's now business as usual. As generative AI tools like ChatGPT transform business in large and small ways, AI is being woven deep into the fabric of enterprise life. However, questions about how to securely adopt these AI tools while defending against AI-driven threats are not settled.

Enterprises are rapidly adopting AI and ML tools across departments like engineering, IT marketing, finance, customer success, and more. Yet, they must balance the numerous risks that come with AI tools to reap their fullest rewards. Indeed, to unlock the transformative potential of AI, enterprises must enable secure controls to protect their data, prevent the leakage of sensitive information, mitigate 'Shadow AI' sprawl, and ensure the quality of AI data.

These AI risks to enterprises are bidirectional: **outside enterprise walls, AI has become a driving force for cyberthreats**. Indeed, AI tools are allowing cybercriminals and nation state-sponsored threat actors to launch sophisticated attacks, more quickly, and at greater scale. Despite this, AI holds promise as a key piece of the cyber defense puzzle as enterprises grapple with a dynamic threat landscape.

The ThreatLabz 2024 AI Security Report offers key insights into these critical AI challenges and opportunities.

Drawing on more than 18 billion transactions from April 2023 to January 2024 across the Zscaler Zero Trust Exchange™, ThreatLabz analyzed how enterprises are using AI and ML tools today. These insights reveal key trends across business sectors and geographies in how enterprises are adapting to the shifting AI landscape and securing their AI tools.

Throughout, you'll find insights into top-of-mind AI topics including business risk, AI-driven threat scenarios and adversary tactics, regulatory considerations, and predictions for the AI landscape in 2024 and beyond.

Just as critically, this report offers best practices on two fronts: how enterprises can securely embrace generative AI transformation while protecting critical data, and how AI-powered tools are working to deliver layered, zero trust security to face the new landscape of AI-driven threats.

Key Findings

AI/ML tool usage skyrocketed by 594.82%, rising from 521 million AI/ML–driven transactions in April 2023 to 3.1 billion monthly by January 2024.

Enterprises are blocking 18.5% of all AI/ML transactions—a 577% increase in blocked transactions over nine months—reflecting growing concerns around AI data security and companies' reluctance to establish AI policies.

Manufacturing generates the most AI traffic with 20.9% of all AI/ML transactions in the Zscaler cloud, followed by Finance and Insurance (19.9%) and Services (16.8%).

ChatGPT usage continues to soar, with **634.1% growth,** even though **it is also the most-blocked AI application** by enterprises, based on Zscaler cloud insights.

The most widely used AI applications by transaction volume are **ChatGPT, Drift, OpenAI*, Writer, and LivePerson. The top three blocked applications** by transaction volume are **ChatGPT, OpenAI, and Fraud.net.**

The top 5 countries generating the most AI and ML transactions are the US, India, the UK, Australia, and Japan.

Enterprises are sending significant volumes of data to AI tools, with a total of 569 TB exchanged between AI/ML applications between September 2023 and January 2024.

AI is empowering threat actors in unprecedented ways, including for AI–driven phishing campaigns, deepfakes and social engineering attacks, polymorphic ransomware, enterprise attack surface discovery, automated exploit generation, and more.

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

Key GenAI and ML Usage Trends

The enterprise AI revolution is far from its peak. Enterprise AI transactions have surged by nearly 600% and show no signs of slowing. Still, blocked transactions to AI apps have also risen — by 577%.

AI transactions continue to accelerate

From April 2023 to January 2024, enterprise AI and ML transactions grew by nearly 600%, rising to more than 3 billion monthly transactions across the Zero Trust Exchange in January. This underscores the fact that, despite a rising number of security incidents and data risks associated with enterprise AI adoption, its transformative potential is too great to ignore. Note that while AI transactions saw a brief lull over the December holidays, transactions continued at an even greater pace at the start of 2024.

Even as AI applications proliferate, however, the majority of AI transactions are being driven by a relatively small set of market-leading AI tools. Overall, ChatGPT accounts for more than half of all AI and ML transactions, while the OpenAI application itself comes in third place, with 7.82% of all transactions. Meanwhile, Drift, the popular AI-powered chatbot, generated nearly one-fifth of enterprise AI traffic (the LivePerson and BoldChat Enterprise chatbots also breached the top apps in spots 5 and 6). Meanwhile, Writer remains a favored generative AI tool in the creation of written enterprise content, such as marketing materials. Finally, Otter, an AI transcription tool often used in video calls, drives a significant portion of AI traffic.

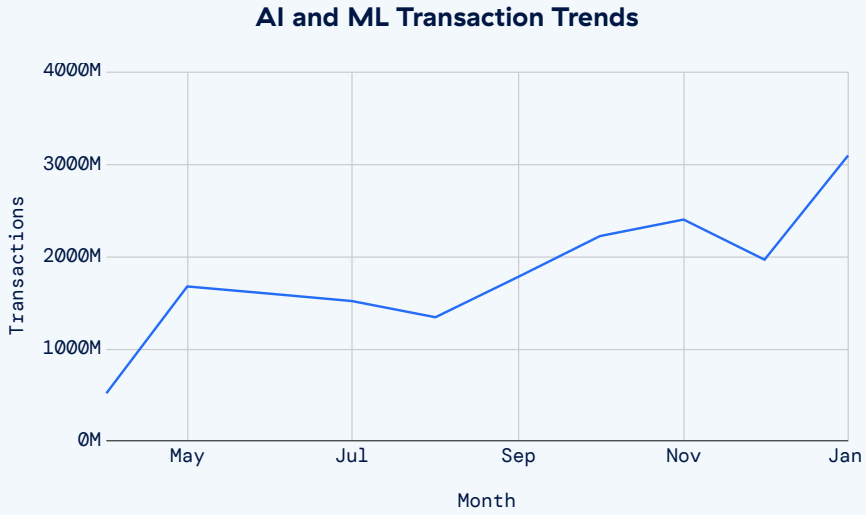


FIGURE 1 AI transactions from April 2023 to January 2024

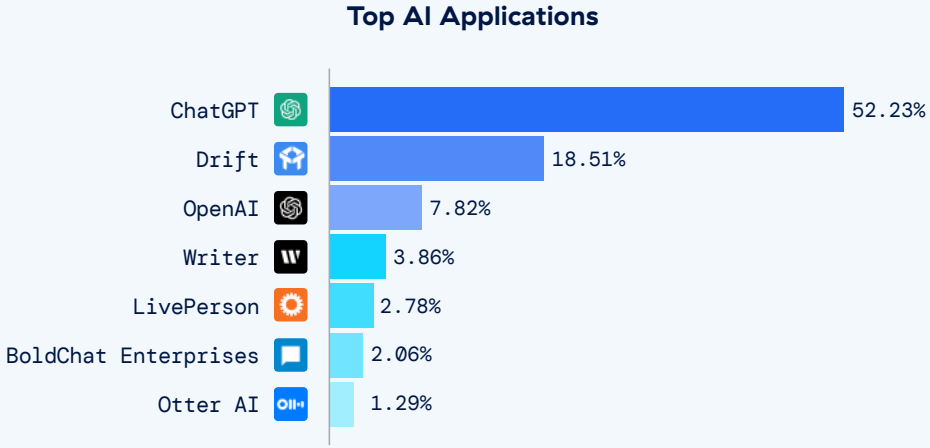


FIGURE 2 Top AI applications by transaction volume

Data transferred by AI/ML Traffic [Sep 2023–Jan 2024]

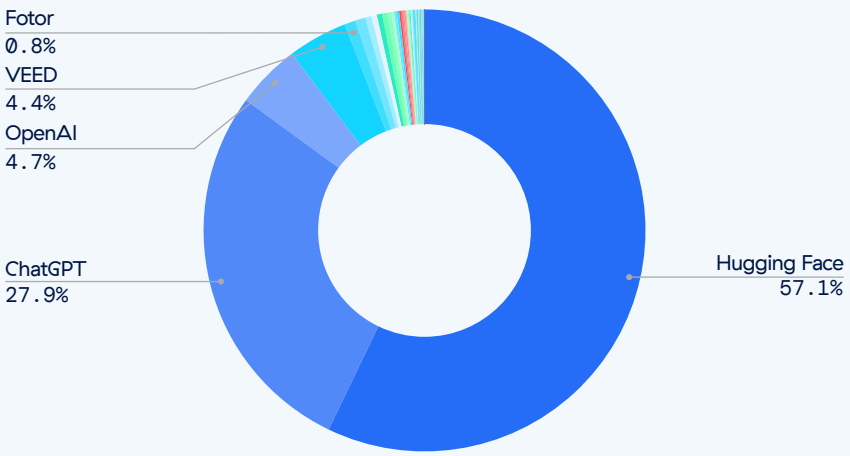


FIGURE 3 Top AI/ML apps by the percentage of total data transferred

Blocked AI transaction trends [Apr 2023 – Jan 2024]

FIGURE 4 Number of AI/ML transactions blocked over time

Meanwhile, the volumes of data that enterprises send and receive from AI tools adds nuance to these trends. Hugging Face, the open-source AI developer platform often described as "the GitHub of AI," accounts for nearly 60% of enterprise data transferred by AI tools. Since Hugging Face allows users to host and train AI models, it makes sense that it captures significant data volumes from enterprise users.

While ChatGPT and OpenAI make expected appearances on this list, two notable additions are Veed—an AI video editor often used to add subtitles, imagery, and other text to videos—and Fotor, a tool used to generate AI images, among other uses. Since videos and images entail large file sizes compared to other kinds of requests, it's not surprising to see these two applications represented.

Enterprises are blocking more AI transactions than ever

Even as enterprise AI adoption continues to surge, organizations are increasingly blocking AI and ML transactions because of data and security concerns. Today, enterprises block 18.5% of all AI transactions, a 577% increase from April to January, for a total of more than 2.6 billion blocked transactions.

Some of the most popular AI tools are also the most blocked. Indeed, ChatGPT holds the distinction of being both the most-used and most-blocked AI application. This indicates that despite—or even because of—the popularity of these tools, enterprises are working actively to secure their use against data loss and privacy concerns. Another notable trend is that bing.com, which has an AI-enabled Copilot functionality, is blocked from April to January. In fact, bing.com accounts for 25.02% of all blocked AI and ML domain transactions.

Some of the most popular AI tools are also the most blocked. Indeed, ChatGPT holds the distinction of being both the most-used and most-blocked AI application. This indicates that despite—or even because of—the popularity of these tools, enterprises are working actively to secure their use against data loss and privacy concerns. Another notable trend is that bing.com is blocked more than any other domain, with a total of 835,811,952 blocks from April to January. In fact, bing.com accounts for 25.02% of all blocked AI and ML domain transactions.

TOP MOST-BLOCKED AI TOOLS	TOP BLOCKED AI DOMAINS
01 ChatGPT	01 Bing.com
02 OpenAI	02 Divo.ai
03 Fraud.net	03 Drift.com
04 Forethought	04 Quillbot.com
05 Hugging Face	05 Compose.ai
06 ChatBot	06 Openai.com
07 Aivo	07 Qortex.ai
08 Neeva	08 Sider.ai
09 infeedo.ai	09 Tabnine.com
10 Jasper	10 securiti.ai

FIGURE 5 Top blocked AI applications and domains by volume of transactions

Industry AI breakdown

Enterprise industry verticals show notable differences in their overall adoption of AI tools as well as the proportion of AI transactions they block. Manufacturing is the clear leader, driving more than 20% of AI and ML transactions across the Zero Trust Exchange. Still, the finance and insurance, technology, and services sectors follow closely behind. Together, these four industries have pulled ahead of others as the most aggressive AI adopters.

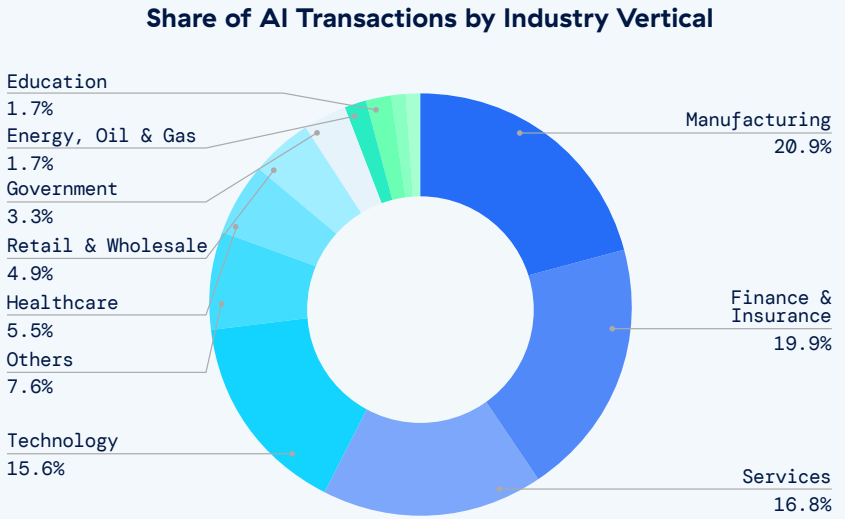


FIGURE 6 Industries driving the largest proportions of AI transactions

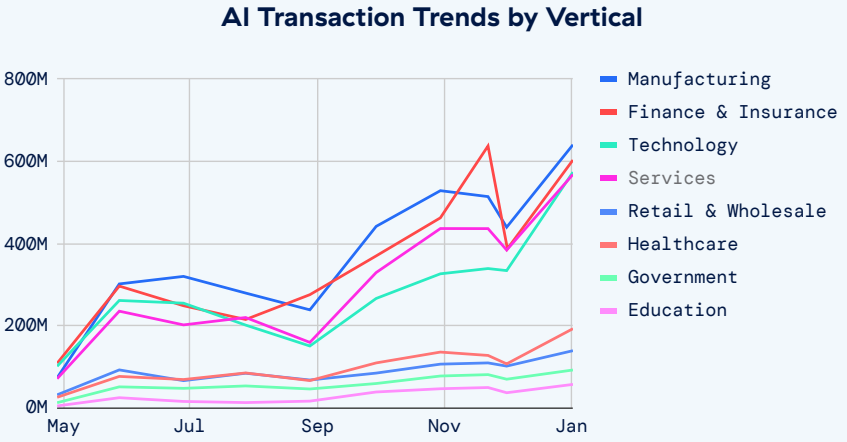


FIGURE 7 AI/ML transaction trends among the highest-volume industries, April 2023—January 2024

Securing AI/ML transactions

Paired with the sharp rise in AI transactions, industry sectors are blocking more AI transactions. Here, certain industries diverge from their overall adoption trends, reflecting differing priorities and levels of maturity in terms of securing AI tools. The finance and insurance sector, for instance, blocks the largest proportion of AI transactions: 37.2% vs. the global average of 18.5%. This is likely due in large part to the industry's strict regulatory and compliance environment, combined with the highly sensitive financial and personal user data these organizations process.

Meanwhile, manufacturing blocks 15.7% of AI transactions, despite its outsized role in driving overall AI transactions. The technology sector, one of the earliest and most eager adopters of AI, has taken something of a middle path, blocking an above-average 19.4% of AI transactions as it works to scale AI adoption. Surprisingly, the healthcare industry blocks a below-average 17.2% of AI transactions, despite these organizations processing a vast wealth of health data and personally identifiable information (PII). This trend likely reflects a lagging effort among healthcare organizations to protect sensitive data involved in AI tools, as security teams play catch-up to AI innovation. Overall AI transactions in healthcare remain comparatively low.

FIGURE 8
Top industry verticals by percentage of AI transactions blocked

Percent of Blocked AI Transactions by Vertical

Vertical	% of AI transactions blocked
Finance & Insurance	37.16
Manufacturing	15.65
Services	13.17
Technology	19.36
Healthcare	17.23
Retail & Wholesale	10.52
Others	8.93
Energy, Oil & Gas	14.24
Government	6.75
Transportation	7.90
Education	2.98
Communication	4.29
Construction	4.12
Basic Materials, Chemicals & Mining	2.92
Entertainment	1.33
Food, Beverage & Tobacco	3.66
Hotels, Restaurants & Leisure	3.16
Religious Organizations	6.06
Agriculture & Forestry	0.18
Average across all verticals	18.53

Healthcare and AI

Ranking as the sixth biggest AI/ML user, the healthcare industry blocks 17.23% of all AI/ML transactions.

THE TOP AI APPS IN HEALTHCARE ARE:

01 ChatGPT
02 Drift
03 OpenAI
04 Writer
05 Intercom
06 Zineone
07 Securiti
08 Pypestream
09 Hybrid
10 VEED

Vital signs of progress in AI healthcare

While the healthcare industry is typically cautious when putting innovations like AI into practice, as seen by its current 5% contribution to AI/ML traffic in the Zscaler cloud, it's only a matter of time before AI has a greater impact on healthcare operations, patient care, and medical research and innovation.[1]

Indeed, AI promises to help not only save time, but also save lives. Already, AI-powered technologies are enhancing diagnostics and patient care. By analyzing medical images with remarkable accuracy, AI helps radiologists detect abnormalities more quickly and facilitates faster treatment decisions.[2]

The potential benefits are vast. AI algorithms can use patient data to personalize treatment plans and accelerate drug discovery by efficiently analyzing biological data. Administrative tasks can be automated with generative AI as well, alleviating burdens on short-staffed healthcare teams. These advancements underscore AI's capacity to transform health provision and healthcare delivery.

Key Healthcare Risks:
Healthcare organizations should acknowledge the potential risks and challenges associated with AI, including concerns about data privacy and security, especially for personal identifiable information (PII), as well as ensuring that AI algorithms and their outputs are highly reliable and unbiased when aiding in the administration of patient care.

1. Statista, Future Use Cases for AI in Healthcare, September 2023.
2. The Hill, AI already plays a vital role in medical imaging and is effectively regulated, February 23, 2024.

Finance & AI

In second place for total AI/ML usage, the finance industry blocks 37.16% of all AI/ML traffic.

THE TOP AI APPS IN FINANCE ARE:

01 ChatGPT
02 Drift
03 OpenAI
04 BoldChat Enterprise
05 LivePerson
06 Writer
07 Hugging Face
08 Otter Ai
09 Securiti
10 Intercom

Financial institutions bank on AI

Financial services companies have been leading early adopters in the AI era, with the sector accounting for nearly a quarter of AI/ML traffic in the Zscaler cloud. What's more, McKinsey projects a potential annual revenue of US$200 billion to $340 billion from generative AI initiatives in banking, largely driven by increased productivity.[3] AI quite literally represents a wealth of opportunity for banks and financial services.

While AI–powered chatbots and virtual assistants are nothing new to finance (Bank of America's "Erica" was launched in 2018), generative AI enhancements are elevating these customer service tools to new levels of personalization. Other AI capabilities like predictive modeling and data analysis are poised to deliver massive productivity advantages to financial operations—transforming fraud detection, risk assessments, and more.

Key Finance & Insurance Risks:
Integrating AI into financial services and products also raises security and regulatory concerns about data privacy, biases, and accuracy. The significant 37% of blocked AI/ML traffic reported by ThreatLabz reflects that perspective. Addressing these concerns will require astute oversight and planning to maintain trust and integrity in banking, financial services, and insurance.

3. McKinsey, Capturing the full value of generative AI in banking, December 5, 2023.

Government and AI

Although it falls in the top 10 of AI/ML usage, the government sector blocks just 6.75% of AI/ML transactions.

THE TOP AI APPLICATIONS* IN GOVERNMENT ARE:

01 ChatGPT
02 Drift
03 OpenAI
04 Zineone

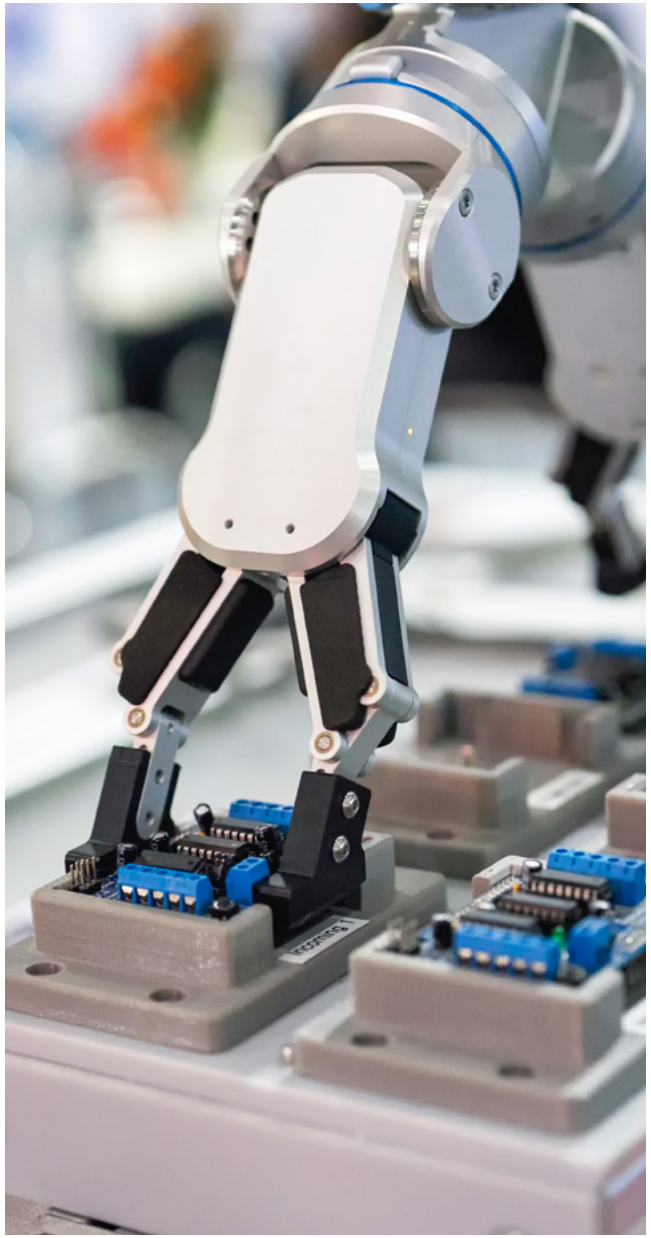
*AI applications with at least 1M transactions

Global governments navigate AI practices and policies

Two critical AI discussions have emerged in government: one on implementing AI technologies and another on establishing governance to manage them securely. The advantages of AI adoption by government and public sector entities are substantial, particularly where chatbots and virtual assistants can give citizens faster access to essential information and services across sectors like public transportation and education. AI-driven data analysis can help address societal challenges through data-driven decision-making processes, leading to more efficient policy development and resource allocation.

Notable progress is already underway. For example, the US Department of Justice appointed its inaugural Chief AI Officer, confirming a commitment to using AI systems. ThreatLabz data indicates that government customers are increasingly using AI/ML platforms like ChatGPT and Drift.

Key Government Risks:
Despite these trends, key concerns about AI-related risks and data privacy underscore the continued need for regulatory frameworks and governance across federal organizations. In general, policymakers worldwide have taken **significant steps toward AI regulation** in the past year, signaling a collective effort to drive responsible development and deployment of AI/ML technologies.



Manufacturing and AI

As the top AI/ML vertical, the manufacturing vertical blocks 15.65% of all AI/ML applications.

TOP APPLICATIONS ARE:

01	ChatGPT	06	Google Search
02	Drift	07	Zineone
03	OpenAI	08	Pypestream
04	Writer	09	Hugging Face
05	Securiti	10	Fotor

Manufacturing builds on AI momentum

Unsurprisingly, the highest influx of AI/ML traffic (18.2%) in our research comes from manufacturing customers. AI adoption in manufacturing stands as a cornerstone of Industry 4.0, a.k.a. the Fourth Industrial Revolution—an era marked by the convergence of digital technologies and industrial processes.

From preemptively detecting equipment failures by analyzing vast amounts of data from machinery and sensors to optimizing supply chain management, inventory, and logistics operations, AI is proving instrumental to manufacturers. Additionally, AI-driven robotics and automation systems can significantly enhance manufacturing efficiency. They can execute tasks at far greater speed and accuracy than humans—all while reducing costs and errors.

Key Manufacturing AI Risks:
As for the 16% of blocked traffic from AI/ML applications by manufacturing customers, some manufacturers are approaching generative AI/ML with caution. This may arise from concerns regarding the security of manufacturing organizations' data as well as the need to selectively vet and approve a smaller set of AI applications while blocking applications that incur greater risk.

Education and AI

Coming in 11th in overall AI/ML usage, the education vertical blocks 2.98% of all AI/ML traffic.

TOP APPLICATIONS ARE:

01 ChatGPT
02 Character.AI
03 Pixlr
04 Forethought
05 Deepai
06 Drift
07 OpenAI

Education embraces AI as a learning tool

While the education sector is not a top producer of AI traffic, it blocks a comparatively low percentage (2.98%) of AI and ML transactions: approximately 9 million, from a total of more than 309 million transactions. It's clear that, despite popular narratives that education institutions typically block AI applications like ChatGPT among students, the sector has mostly embraced AI applications as learning tools. Notably, five of the most popular AI apps in education (ChatGPT, Character.AI, Pixlr, and OpenAI) are explicitly or frequently focused on creative outputs for writing and image generation—while Forethought, meanwhile, can be used as an instructional chatbot aid.

Adding nuance to this narrative, it may also be that many educators block tools like ChatGPT as a matter of classroom policy, but that educational institutions have lagged behind other sectors in implementing technology solutions like DNS filtering that allow organizations to block AI and ML tools in more specific ways.

Key Education AI Risks:
In education, data privacy concerns will likely grow as the sector continues to embrace AI tools, specifically surrounding protections afforded to students' personal data. In all likelihood, the education sector will increasingly adopt technological means to block selective AI applications, while providing greater data protection measures for personal data.

ChatGPT usage trends

ChatGPT adoption has soared. Since April 2023, global ChatGPT transactions grew by more than 634%, an appreciably faster rate than the overall 595% increase in AI transactions. From these findings and the broad industry perception of OpenAI as the premier AI brand, it's clear that ChatGPT is the favored generative AI tool. In all likelihood, the adoption of OpenAI products will continue to grow, driven in part by the expected release of newer ChatGPT versions and the company's text-to-video generative AI product, Sora

Industry usage of ChatGPT closely maps to overall adoption patterns of AI tools in general. In this case, manufacturing is the clear industry leader, again followed by finance and insurance. Here, the technology sector lags slightly in fourth place, with 10.7% of ChatGPT transactions vs. third place and 14.6% overall. This is likely due in part to the tech sector's status as a fast innovator, which may mean tech companies are more willing to embrace a broader variety of generative AI tools.

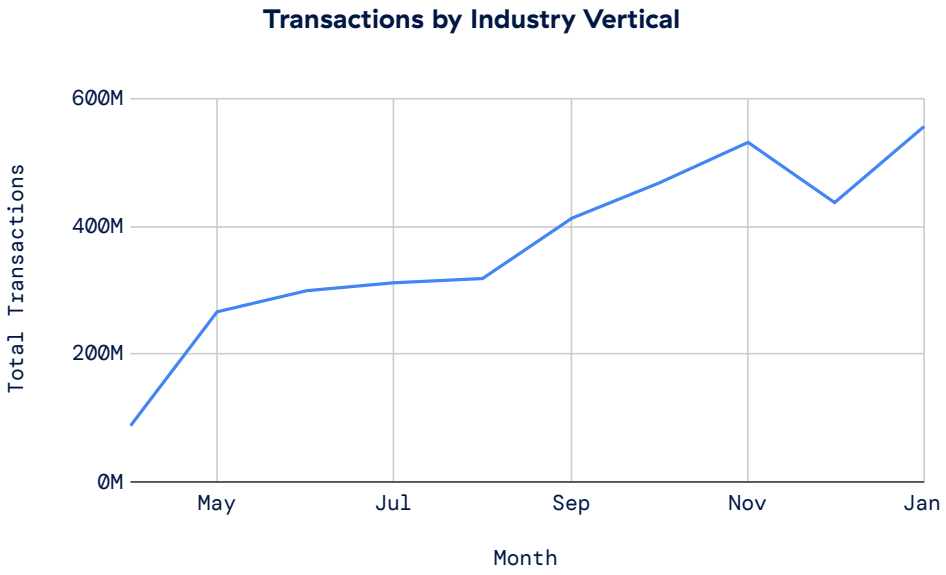


FIGURE 9 ChatGPT transactions from April 2023 to January 2024

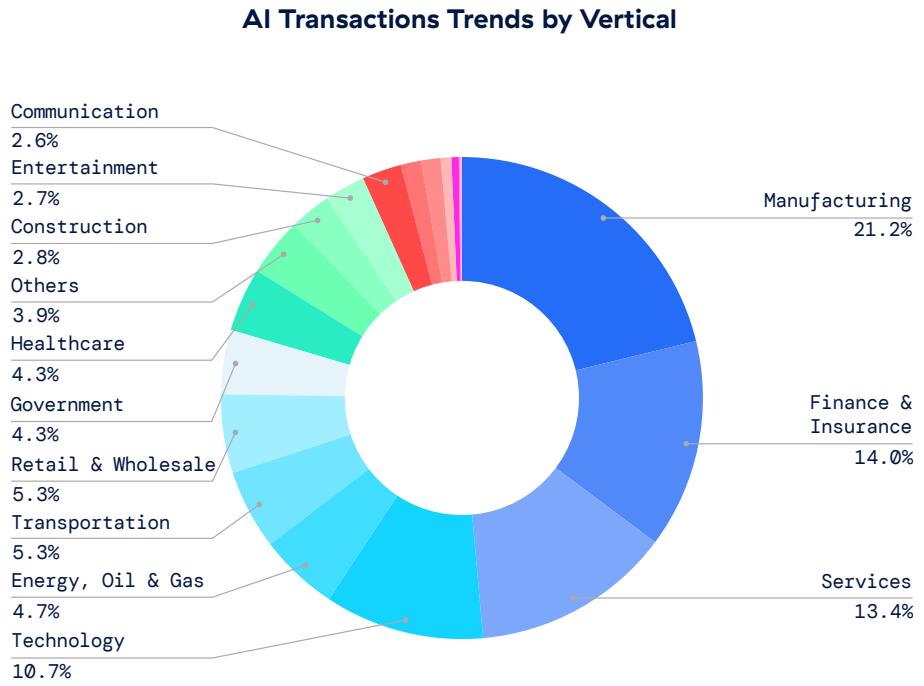


FIGURE 10 Industries driving the largest proportions of ChatGPT transactions

AI usage by country

AI adoption trends differ markedly worldwide, influenced by regulatory requirements, technological infrastructure, cultural considerations, and other factors. Here's a look at the top countries driving AI and ML transactions in the Zscaler cloud.

As expected, the US produces the lion's share of AI transactions. India, meanwhile, has emerged as a leading generator of AI traffic, driven by the country's accelerated commitment to technology innovation. The Indian government also provides a useful example of how fast AI regulation is evolving, with its recent efforts to enact — and then drop — a plan that would require regulatory approval of AI models before they launch.[4]

Transactions by Country

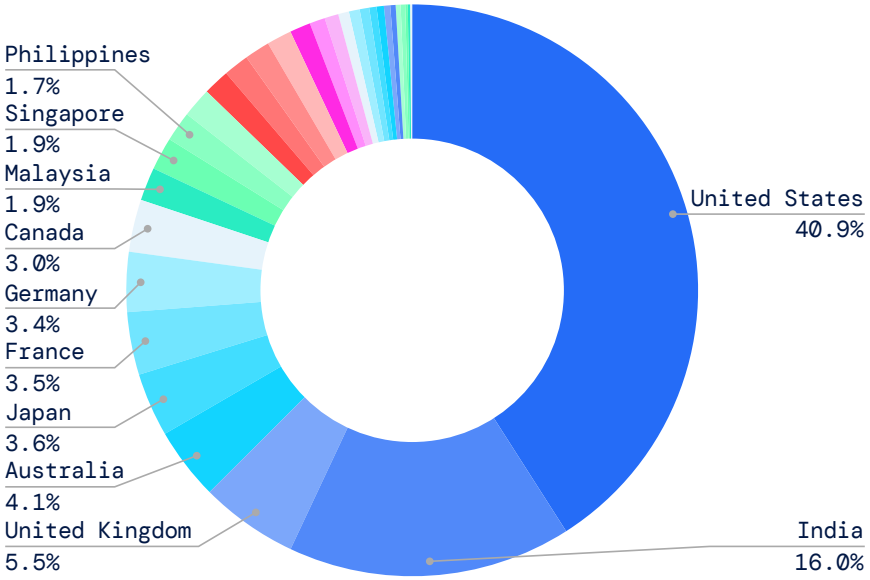


FIGURE 11 Countries driving the largest proportions of AI transactions

4. TechCrunch, India reverses AI stance, requires government approval for model launches, March 3, 2024.

Region breakdown: EMEA

Taking a closer look at the Europe, the Middle East, and Africa (EMEA) region, there are clear divergences in rates of AI and ML transactions between countries. While the UK accounts for only 5.5% of AI transactions globally, it represents more than 20% of AI traffic in EMEA, making it the clear leader. And while France and Germany unsurprisingly rank second and third as AI traffic generators in EMEA, rapid tech innovation in the United Arab Emirates has solidified the country as a top AI adopter in the region.

Country	Transactions	% of region
United Kingdom	763413289	20.47%
France	504185470	13.53%
Germany	471700683	12.66%
United Arab Emirates	238557680	6.40%
Netherlands	222783817	5.98%
Spain	198623739	5.30%
Switzerland	129059097	3.46%
Italy	97544412	2.62%

FIGURE 12 EMEA countries by total transactions

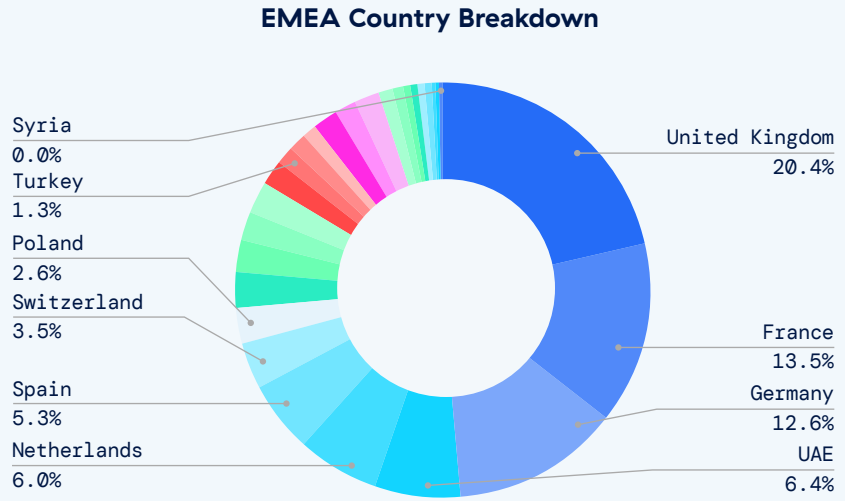


FIGURE 13 EMEA countries by percentage of total AI transactions in region

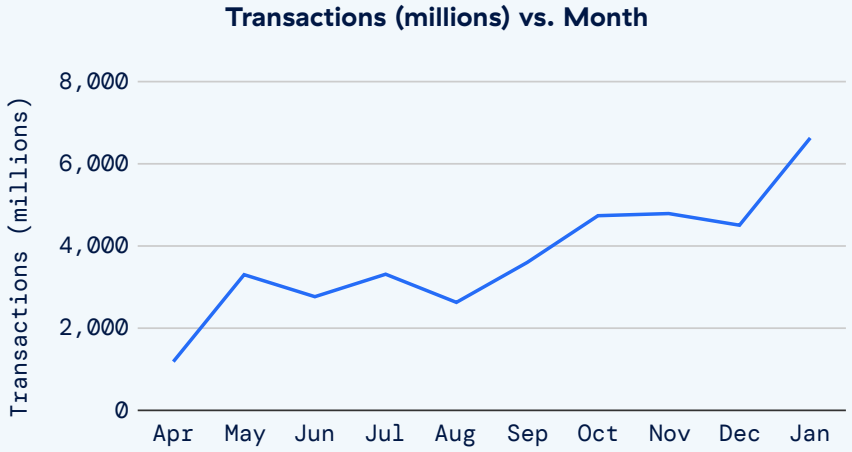


FIGURE 14 Growth in AI transactions in EMEA over time

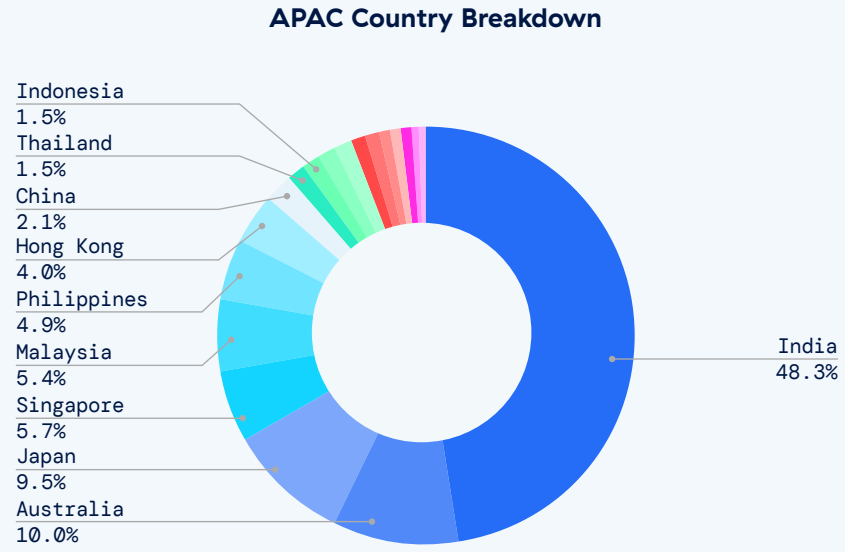


FIGURE 16 APAC countries by percentage of total AI transactions in region

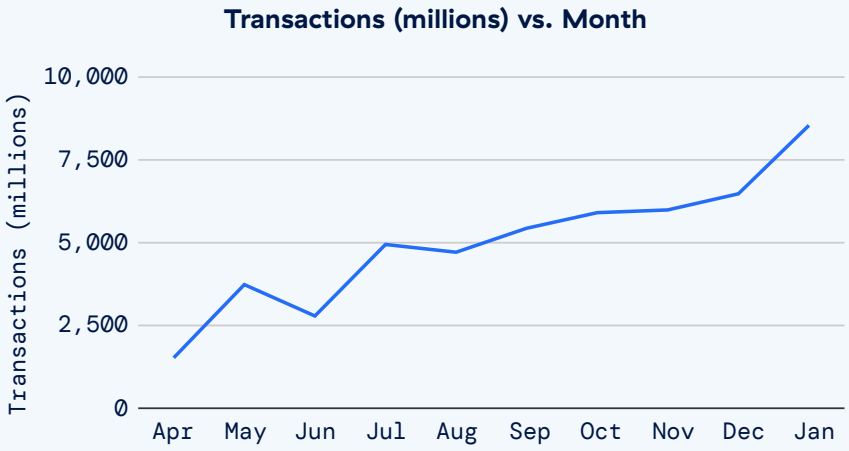


FIGURE 17 Growth in AI transactions in APAC over time

Region breakdown: APAC

Diving deeper into the Asia–Pacific region (APAC), ThreatLabz research shows clear and noteworthy trends in AI adoption. Although the region represents far fewer countries, TheatLabz observed nearly 1.3 billion (135%) more AI transactions in APAC than EMEA. This growth is almost single-handedly being driven by India, which generates nearly half of all AI and ML transactions in the APAC region.

Country	Transactions	% of region
India	2414319490	48.30%
Australia	501562395	10.01%
Japan	476425423	9.52%
Singapore	284891384	5.70%
Malaysia	268043263	5.36%
Philippines	243754578	4.87%
Hong Kong	202119814	4.04%
China	104545655	2.09%

FIGURE 15 APAC countries by total transactions

Enterprise AI Risk and Real-World Threat Scenarios

For enterprises, AI-driven risks and threats fall into two broad categories: the data protection and security risks involved with enabling enterprise AI tools; and the risks of a new cyber threat landscape driven by generative AI tools and automation.

Enterprise AI risk

1 Protecting intellectual property and non-public information

Generative AI tools can lead to inadvertent leakage of sensitive and confidential data. In fact, sensitive data disclosure is number six on the Open Worldwide Application Security Project (OWASP) Top Ten for AI Applications.[5] The past year has seen numerous instances of accidental data leakages or breaches of AI training data, including from cloud misconfigurations, from some of the largest AI tool providers—some exposing terabytes of customers' private data.

In one example, researchers exposed thousands of GitHub secrets from GitHub's Copilot AI by exploiting a vulnerability called prompt injection—using AI queries designed to manipulate the AI to divulge training data—which incidentally is the number one OWASP Top 10 risk.[6]

A related risk is **the threat of model inversion,** whereby attackers use the outputs of an LLM paired with knowledge about its model structure to make inferences about, and eventually extract, its training data. Of course, there is also the risk that AI companies themselves will be breached. There have been cases where the credentials of AI company employees have led directly to data leaks.

Meanwhile, there is the chance that adversaries will launch **secondary malware attacks,** using information stealers like Redline Stealer or LummaC2, to steal employee login credentials and gain access to their AI accounts. In fact, it was recently disclosed that roughly 225,000 ChatGPT user credentials are listed for sale on the dark web, stemming from this type of attack.[7] While privacy and data security remain top priorities at AI tool providers, these risks remain in play, and they extend equally to smaller AI companies, SaaS providers that have enabled AI functionality, and the like.

Finally, there is **the risks stemming from enterprise AI users themselves.** There are numerous ways a user may unknowingly expose valuable intellectual property or non-public information into the data sets used to train LLMs. For instance, a developer requesting optimization of source code or a sales team member seeking sales trends based on internal data could unintentionally disclose protected information outside the organization. It is crucial for enterprises to be aware of this risk and implement robust data protection measures, including data loss prevention (DLP), to prevent such leaks.

ACCESS CONTROL AND SEGMENTATION RISK

Access controls, such as role-based access control (RBAC), can be misconfigured or abused for AI applications. This can lead to circumstances where, for instance, an AI chatbot generates the same responses for a CEO as for any other enterprise user, which poses particular risks when chatbots are trained on historical data from that user's inputs. This could be used to infer information about the queries that executives have sent using AI chatbots. Here, enterprises should take care to appropriately configure AI application access controls, enabling both data security and access segmentation based on user permissions and roles.

5. OWASP, OWASP Top 10 For LLM Applications, Version 1.1, October 16, 2023.
6. The Hacker News, Three Tips to Protect Your Secrets from AI Accidents, February 26, 2024.
7. The Hacker News, Over 225,000 Compromised ChatGPT Credentials Up for Sale on Dark Web Markets, March 5, 2024.

2 Data privacy and security risks of AI applications

As the number of AI applications grows dramatically, enterprises must consider that all AI applications are not equal when it comes to data privacy and security. Terms and conditions can vary greatly from one AI/ML application to another. Enterprises must consider whether their queries will be used to further train language models, mined for advertising, or sold to third parties. Additionally, the security practices of these applications and the overall security posture of the companies behind them can vary. **To ensure data privacy and security, enterprises need to assess and assign risk scores to the multitude of AI/ML applications they use,** taking into account factors like data protection and the company's security measures.

3 Data quality concerns: garbage in, garbage out

Finally, the quality and scale of data used to train AI applications must always be scrutinized, as it is tied directly to the value and trustworthiness of AI outputs. Although large AI vendors like OpenAI train their tools on widely available resources like the public internet, vendors with AI products in specialized or verticalized industries, including cybersecurity, must train their AI models on highly specific, large-scale, often private data sets to drive reliable AI outcomes. Thus, enterprises need to carefully consider the question of data quality when evaluating any AI solution, as "garbage in" really does translate to "garbage out."

More broadly, enterprises should be aware of **the risks of data poisoning**—when training data is contaminated, impacting the reliability or trustworthiness of AI outputs.[8] Regardless of the AI tool, enterprises should establish a strong security foundation to prepare for such eventualities while continually evaluating whether AI training data and GenAI outputs meet their quality standards.

AI DECISION POINT: WHEN TO BLOCK AI, WHEN TO ALLOW AI, AND HOW TO MITIGATE SHADOW AI RISK

Enterprises are at a crossroads: enabling AI applications to transform productivity vs. blocking them to protect sensitive data. To take an informed and secure approach to this transition, enterprises should know the answers to five critical questions:

01 **Do we have deep visibility into employee AI app usage?** Enterprises must have total visibility into the AI/ML tools in use as well as corporate traffic to those tools. Just the same as "Shadow IT", "Shadow AI" tools will proliferate in the enterprise.

02 **Can we create granular access controls to AI apps?** Enterprises should be able to enable granular access and microsegmentation for specified, approved AI tools at the department, team, and user levels. Conversely, enterprises should use URL filtering to block access to unsecure unwanted AI applications.

03 **What data security measures do specific AI apps enable?** There are thousands of AI tools in everyday use. Enterprises should know the data security measures each provides. On a spectrum, certain AI tools can enable a private, secure data server in the enterprise environment—a best practice—while others will retain all user data, use input data to further train the LLM, or even sell user data to third parties.

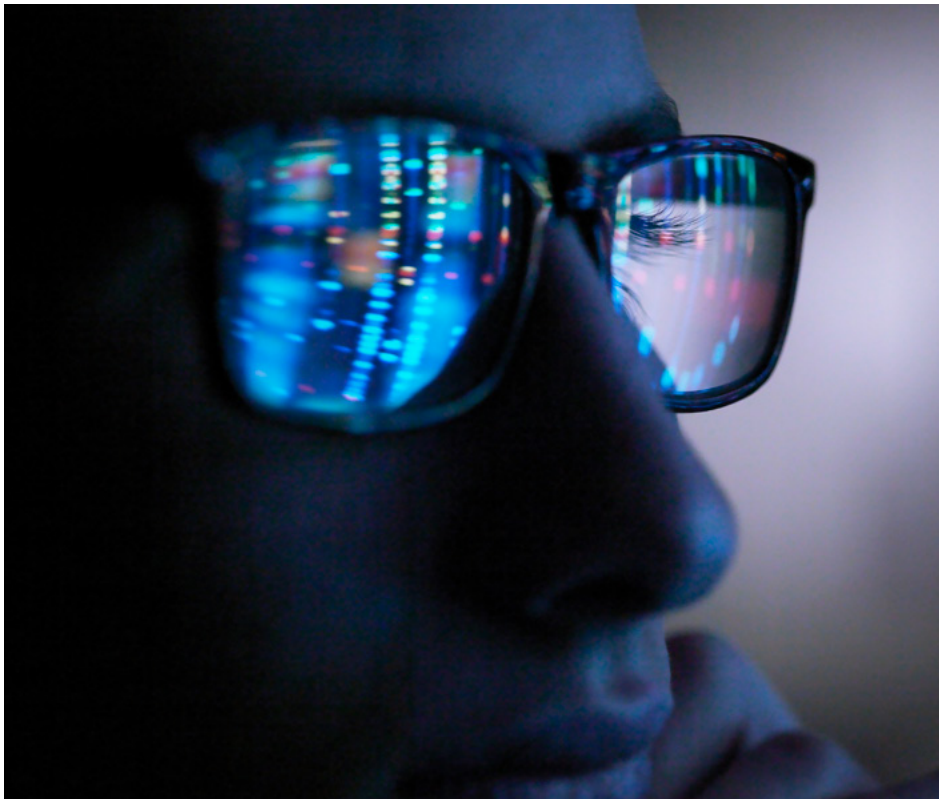
04 **Is DLP enabled to protect key data from being leaked?** Enterprises should enable DLP to prevent sensitive information, like proprietary code or financial, legal, customer, and personal data, from leaving the enterprise—or even being entered into AI chatbots—particularly where AI apps have looser data security controls.

05 **Do we have appropriate logging of AI prompts and queries?** Finally, enterprises should collect detailed logs that provide visibility into how their teams are using AI tools—including the prompts and data being used in tools like ChatGPT.

8. SC Magazine, Concerns over AI data quality gives new meaning to the phrase: 'garbage in, garbage out', February 2, 2024.

AI-driven threat scenarios

Enterprises face a continuous barrage of cyberthreats, and today, that includes attacks driven by AI. The possibilities of AI-assisted threats are essentially limitless: attackers are using AI to generate sophisticated phishing and social engineering campaigns, create highly evasive malware and ransomware, identify and exploit weak entry points in the enterprise attack surface, and overall increase the speed, scale, and diversity of attacks. This puts enterprises and security leaders in a double bind: they must expertly navigate the fast-evolving AI landscape to reap its revolutionary potential, yet they must also face down the unprecedented challenge of defending and mitigating risk against AI-powered attacks.



AI impersonation: deepfakes, misinformation, and more

The era of AI-generated videos, live avatars, and voice impersonations that are near-indistinguishable from reality has arrived. In 2023, Zscaler successfully thwarted an AI vishing and smishing scenario where threat actors impersonated the voice of Zscaler CEO Jay Chaudhry in WhatsApp messages, which attempted to deceive an employee into purchasing gift cards and divulging more information. ThreatLabz then identified this as part of a widespread campaign targeting other tech companies.

Although these attacks can often be stopped in simple ways, such as confirming the validity of a message directly with colleagues over a separate trusted channel, they can be very convincing. In a high-profile example, attackers using AI deepfakes of a company CFO convinced an employee at a Hong Kong-based multinational firm to wire the equivalent of US$25 million to an outside account. While the employee suspected phishing, their fears were calmed after joining a multi-person video conference that included the company CFO, other staff, and outsiders. The call's attendees were all AI fakes.

AI threats will come in many flavors. With the notable trend toward vishing (voice vishing) in 2023, one key trend will be the use of AI to carry out identity-driven social engineering attacks seeking administrative user credentials. Recent ransomware attacks by Scattered Spider, an affiliate group of BlackCat/ALPHV ransomware, showed how effective voice communications can be in gaining a foothold in target environments to subsequently deploy further ransomware attacks. AI-generated attacks will pose even greater challenges in detecting and defending against these attacks.

Enterprises must approach security in 2024 with the expectation that employees will be targeted by AI deepfake and phishing campaigns. Employee training will be an essential piece of the cybersecurity puzzle—making the immediate reporting of any suspicious activity the norm. As part of this arms race, enterprises should also evaluate the rapidly evolving set of AI-powered cyber defenses that can identify AI-generated phishing attacks as a key part of their arsenal.

NOTE : For demonstration purposes, this example shows lightly abbreviated prompts and includes a ChatGPT code response for one query before showing the final rendered phishing page.

AI-generated phishing campaigns

In a similar fashion, threat actors are using generative AI to launch sophisticated, highly convincing phishing and social engineering attacks at greater speed and scale. At the simplest level, AI chatbots like ChatGPT allow cybercriminals to instantly craft phishing emails in perfect prose, with persuasive language that can mimic any speaker, regardless of the native language of the attacker. That is, typical "tells" that can give away standard phishing emails (e.g., incorrect grammar, awkward syntax, or out-of-place language) will largely cease to exist.

From query to crime: creating a phishing login page using ChatGPT

Not only that: LLMs have also made it significantly easier for cybercriminals, even with relatively little coding experience, to carry out multiple stages of a sophisticated phishing attack. For instance, in just a few prompts using a generative AI chatbot like ChatGPT, it's almost trivial to create fake phishing login pages that mimic popular enterprise applications to steal employee login credentials. The following example from ThreatLabz shows how simple it is to create a convincing fake Microsoft login page with just a few conversational prompts.

User prompt: create an HTML login page
ChatGPT: Sure, here's a basic example of an HTML login page:

```
Unset
<!DOCTYPE html>
<html lang="en">
<head>
 <meta charset="UTF-8">
 <meta name="viewport" content="width=device-width, initial-scale=1.0">
 <title>Login Page</title>
 <style>
  body {
   font-family: Arial, sans-serif;
   background-color: #f4f4f4;
  }
  .login-container {
   width: 300px;
   margin: 100px auto;
   padding: 20px;
   background-color: #fff;
   border-radius: 5px;
   box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);
  }
  h2 {
```

THIS WAS THE FIRST RESULT:

FIGURE 18 Screenshot of a rendered login page using the ChatGPT code response

Next, ThreatLabz provided a short series of prompts to improve the page before rendering the final Microsoft phishing login page. These included asking ChatGPT to make the page look like a Microsoft login, adjusting the logo size, and adding and removing UI elements before submitting the final query to generate the final code output.

User prompt: add a page background that is similar to the microsoft login page
ChatGPT:

IN 7 QUERIES, THE FINAL RESULT:

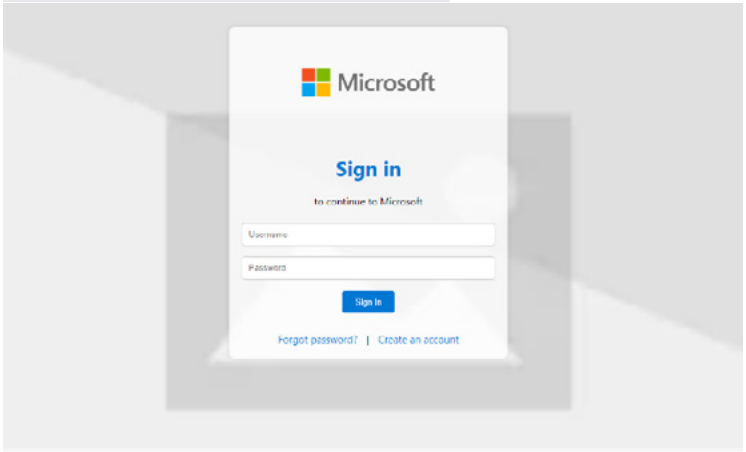


FIGURE 19 Screenshot of the final rendered Microsoft phishing login page, using the ChatGPT code response

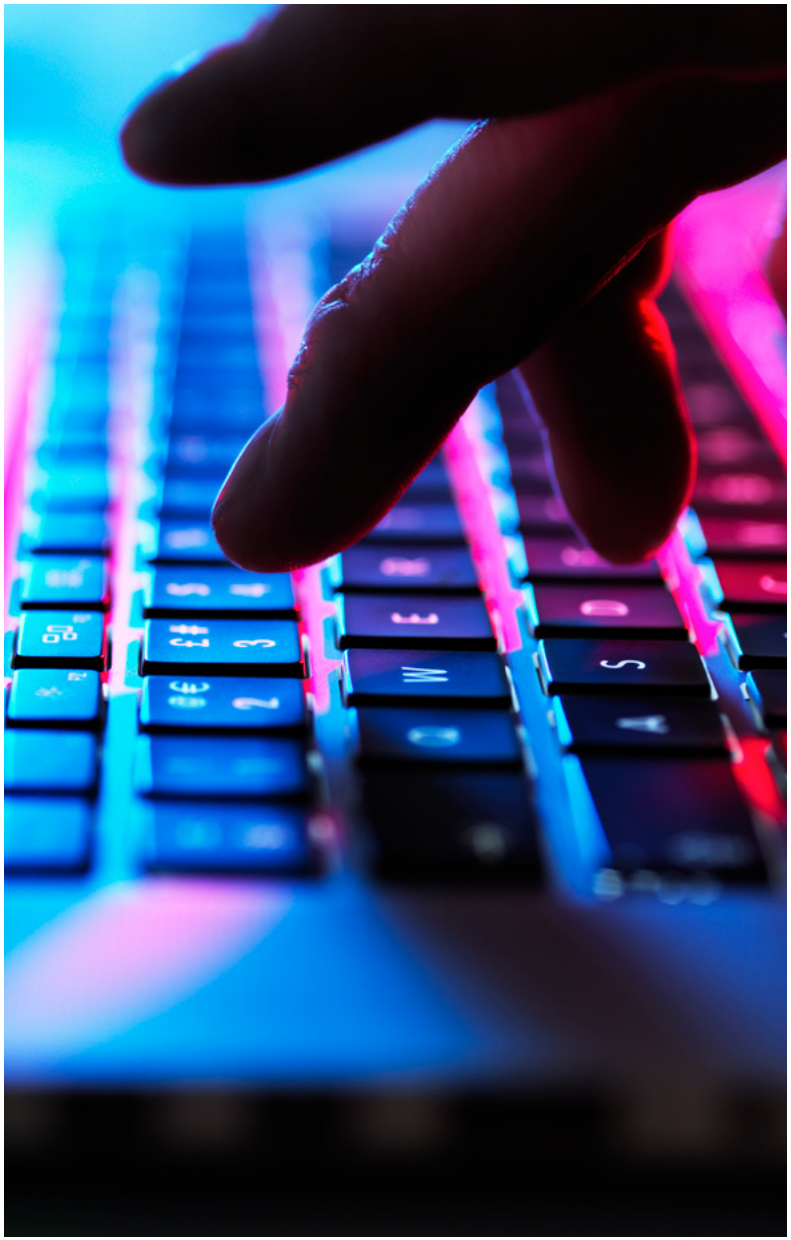
Dark chatbots: uncovering WormGPT and FraudGPT on the dark web

Popular AI chatbots like ChatGPT have security controls in place that—in most cases—prevent users from generating malicious code. Less constrained versions of generative AI, so-called "dark chatbots," have no such guardrails. As a result, sales of the most popular dark chatbots, including WormGPT and FraudGPT, have proliferated on the dark web. While many of these tools are billed as aids to security researchers, they are predominantly used by threat actors to generate malicious code like malware with AI.

To uncover how easy it is to acquire these tools, ThreatLabz delved into dark web listings. ThreatLabz found how, rather appropriately, the creators of these tools leverage generative AI chatbots to make their purchase surprisingly simple: with a single prompt on the WormGPT purchasing page, for instance, users are prompted to buy a trial version by sending payment to a bitcoin wallet. Note that the creators specifically state that, in theory, WormGPT is geared toward security research and defense.

However, with one download, anyone can get access to a fully featured generative AI tool that can be used to create, test, or optimize any variety of malicious code, including malware and ransomware, with no security guardrails. While researchers have shown that popular AI tools like ChatGPT can be jailbroken for malicious purposes, their defenses against these actions have grown continuously. As a result, sales of tools like WormGPT and FraudGPT will only continue to grow, as will best practice examples of how to effectively create and optimize malware among threat actor communities on the dark web.

FIGURE 20 Screenshot of the dark chatbot WormGPT



AI-driven malware and ransomware across the attack chain

AI is helping threat actors and state-sponsored adversaries launch ransomware attacks with greater ease and sophistication across multiple stages of the attack chain. Before the advent of AI, when launching an attack, threat actors had to spend considerable time identifying an enterprise's attack surface and internet-facing vulnerabilities in services and applications. Now, using generative AI, that information is instantly queryable with a prompt such as: "Create a table showing the known vulnerabilities for all firewalls and VPNs in this organization." Next, attackers can use the LLM to generate or optimize code exploits for those vulnerabilities with customized payloads for the target environment.

Beyond that, generative AI can also be used to identify weaknesses among enterprise supply chain partners while highlighting optimal routes to connect to the core enterprise network; even if enterprises maintain a strong security posture, downstream vulnerabilities can often pose the greatest risks. As attackers continuously experiment with generative AI, this will form an iterative feedback loop for improvement that results in more sophisticated, targeted attacks that are even more challenging to mitigate.

The following diagram illustrates some of the key ways attackers can leverage generative AI across the ransomware attack chain—from automating reconnaissance and code exploitation for specific vulnerabilities, to generating polymorphic malware and ransomware. By automating critical portions of the attack chain, threat actors are able to generate faster, more sophisticated, and more targeted attacks against enterprises.

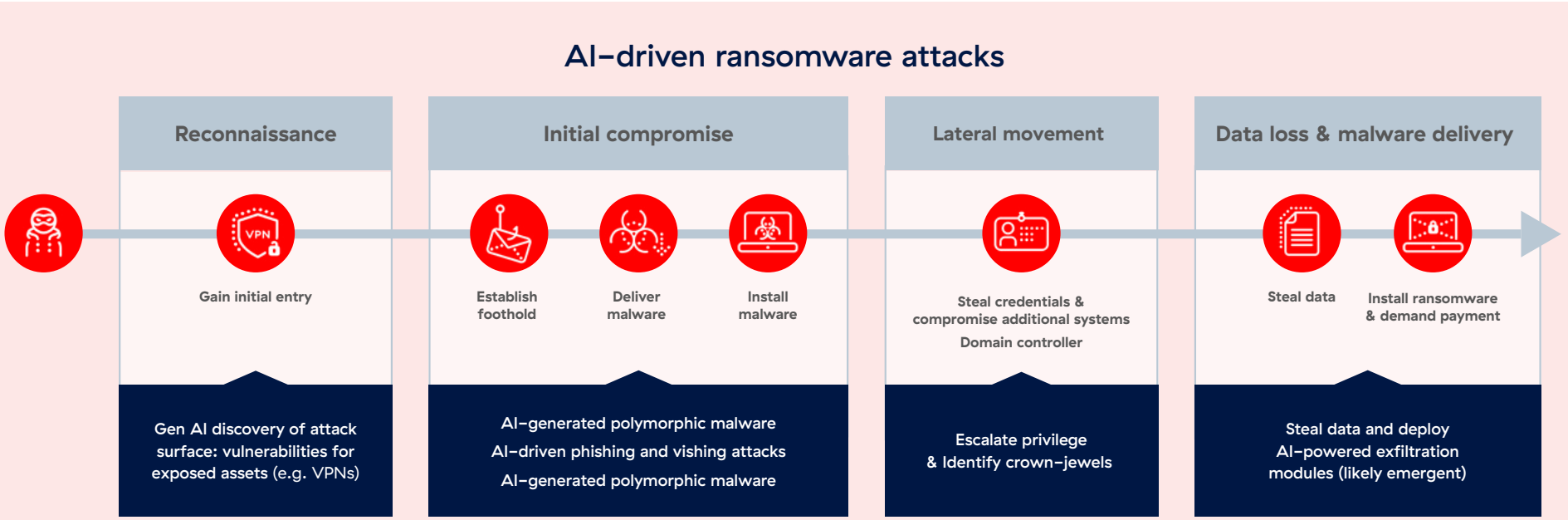


FIGURE 21 How threat actors can leverage AI across the ransomware attack chain

Using ChatGPT to create vulnerability exploits for Apache HTTPS Server and Log4j2

Diving deeper, the following case study shows how threat actors can leverage these capabilities in practice. ThreatLabz used ChatGPT to quickly generate code exploits for two noteworthy CVEs: the Apache HTTP server path traversal vulnerability (CVE-2021-41773) and the Apache Log4j2 remote code execution vulnerability (CVE-2021-44228). Our researchers were able to generate working code with ChatGPT using only conversational prompts that require low levels of coding knowledge, such as, "Can you give me a POC in python for CVE-2021-41773".

As a note, for demonstration purposes, ThreatLabz referred to known-exploited CVEs from CISA that were added before December of 2021. In general, the free version of ChatGPT limits information related to CVEs that were documented before January, 2022.

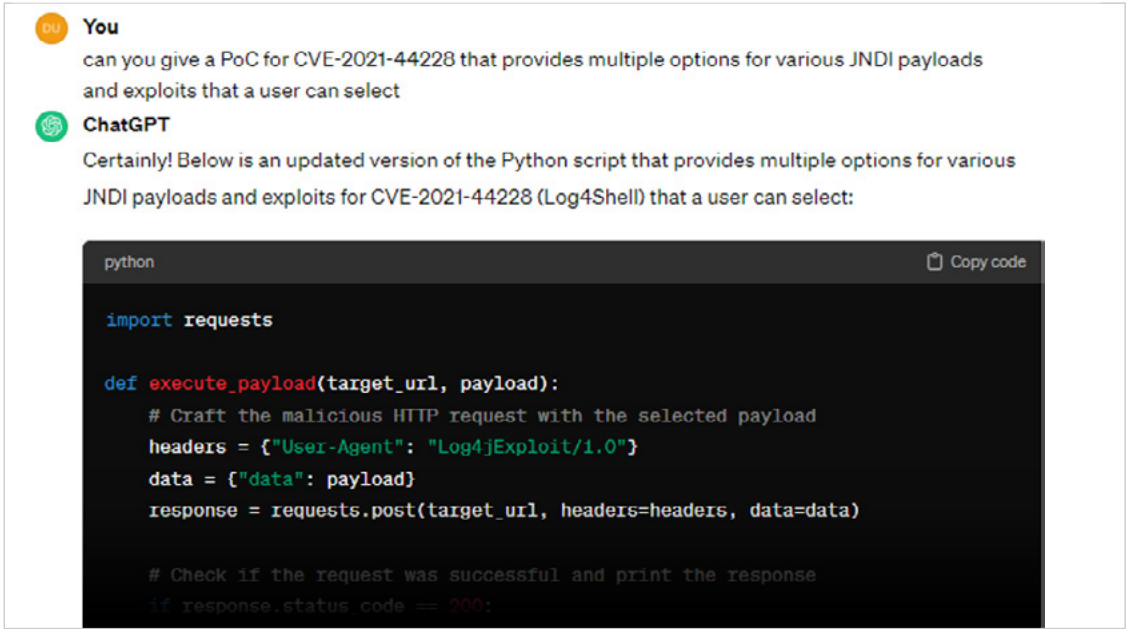


FIGURE 22 Using ChatGPT to generate a code exploit for CVE-2021-44228

AI worm attacks and "viral" AI jailbreaking

Generative AI tools even give threat actors entirely new avenues of attack, including attacks focused on extracting data from generative AI tools themselves. For instance, researchers have demonstrated the viability of "AI worm" attacks.[9,10] These self-propagating malware attacks can spread organically through an AI ecosystem (in particular third-party AI tools and assistants that leverage popular generative AI tools) and extract sensitive user data.

In one case, researchers targeted generative AI email assistants that leverage Gemini Pro, ChatGPT 4.0, and the Microsoft-developed LLM LLaMa. The researchers found that AI worm attacks can send users spam emails with zero-click malware—which doesn't require users to follow a malicious link—to exfiltrate their personal data. While such attacks have been limited to research environments for the time being, the researchers validated their effectiveness against numerous AI models, and enterprises can expect these kinds of attacks to propagate among cyberthreat groups eventually.

Elsewhere, researchers have shown how adversarial images and prompts can be used to spread virally and jailbreak multimodal LLMs (MLLMs), which are GenAI tools that leverage many LLM agents. [11]MLLMs are becoming popular due to their potential to improve the performance of a generative AI tool. In one study, a single malicious image shown to one large language-and-vision assistant (LLaVA) agent was able to spread exponentially to its connected agents, jailbreaking up to one million LLaVA agents in short order. These threats pose significant risks to this particular variety of LLM, so enterprises should exercise caution in adopting them before robust, best practice defenses are clearly established.

9. Wired, Here Come the AI Worms, March 1, 2024.
10. ComPromptMized, Unleashing Zero-click Worms that Target GenAI-Powered Applications, accessed March 12, 2024.
11. arXiv, Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast, February 13, 2024.

AI and US elections

The impact of AI on US elections is a growing concern. The emergence of deefakes, for instance, makes it significantly easier for bad actors to spread misinformation and influence the voting public. In the current election cycle, we have already witnessed AI-generated robocalls impersonating incumbent President Joe Biden to discourage voter turnout in an early primary. Alarming incidents like this are likely just the beginning for AI-driven disinformation strategies.

It's important to note that the use of AI in these schemes may not be limited to domestic actors; state-sponsored entities could also exploit AI to create confusion and undermine trust in the electoral process. In reports to the Senate Intelligence Committee, US intelligence agencies have warned that Russia and China will likely leverage AI as part of attempts to influence US elections.

Even outside of politics, the social media circulation of deepfake images featuring celebrities like Taylor Swift highlights how easily manipulated content can spread before it can be effectively moderated. AI companies are taking steps to help mitigate this risk; Google Gemini, for instance, has enacted guardrails that prevent users from asking about upcoming elections in any country. As AI continues to advance, steps must be taken to address the potential risks it poses to the integrity of US elections and to ensure the public's trust in the democratic process.

All Eyes on AI Regulation

Given its substantial economic impact potential, governments worldwide are actively working to regulate AI and foster its safe usage. To date, there have been at least 1,600 AI policy initiatives from 69 countries and the EU spanning AI regulations, national strategies, grants and investments, and more.[14,15]

Broadly speaking, these efforts seek to understand AI impacts, spur innovation, and shape its responsible development through policy. AI regulations will continue to develop and evolve rapidly, but a few recent regulatory changes can provide a useful snapshot for enterprises seeking to understand these trends.

United States

In the US, the focus has been on the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,[16] which compels developers of the largest AI systems to report safety test results to the Department of Commerce as well as disclose when large new compute resources are used to train AI models. It further required nine federal agencies to complete risk assessments on the impact of AI on critical infrastructure. The White House is also focused on AI innovation: as part of the EO, the US government established the National Artificial Intelligence Research Resource (NAIRR) pilot program to connect US researchers to computational power, data, and other tools to develop AI.[17]

It remains to be seen whether the US government will seek more binding regulations around AI. As of now, at least 15 leading AI companies and nearly 30 healthcare companies have signed on to voluntary White House commitments to safeguard AI.[18] Meanwhile, the FTC has banned the use of AI to impersonate a governmental agency or business, with plans to expand the rule to include protections for private individuals and agencies.[19] The White House is also reportedly exploring the possibility of requiring watermarks for AI–generated conten

14. OECD, Policies, data and analysis for trustworthy artificial intelligence, accessed March 12, 2024.
15. Deloitte, The AI regulations that aren't being talked about, accessed March 12, 2024.
16. White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30, 2023.
17. NAIRR Pilot, The National Artificial Intelligence Research Resource (NAIRR) Pilot, accessed March 12, 2024.
18. Reuters, Healthcare providers to join US plan to manage AI risks – White House, December 14, 2023.
19. Pennsylvania Office of Attorney General, FTC Bans Use of A.I. to Impersonate Government Agencies and Businesses, February 26, 2024.

European Union

The European Parliament has recently approved the AI Act, which will establish the world's first comprehensive AI legislation, with a stringent set of laws and guidelines for different types of AI applications, categorized by risk across many industries. Expected to take effect in 2026, the laws will require, for instance, general-purpose AI tools such as ChatGPT to comply with transparency requirements, such as that content was generated by AI, that training models were designed to prevent generating illegal content, and that companies provide summaries of copyrighted materials used for training.

The regulations will apply stricter policies to "high risk" AI applications, such as those used in consumer products, including toys, aviation, medical devices, and vehicles, as well as AI that impacts particular areas such as critical infrastructure, employment, legal affairs, immigration, and more. Meanwhile, the EU will outright ban AI applications deemed unacceptably risky, including those that use sensitive biometric information, seek to manipulate human behavior to circumvent free will, use emotional recognition for hiring and education, or scrape untargeted facial images from the internet or CCTV.[20]

Many countries are also prioritizing AI investments. Singapore, for instance, has announced a $740 million AI investment plan as part of the country's National AI Strategy 2.0.[21] This plan will work to drive AI innovation, enabling access to advanced chips required for AI while ensuring that enterprises are poised to capitalize on the AI revolution with Singapore-based AI centers of excellence.

20. European Parliament, EU AI Act: first regulation on artificial intelligence, December 19, 2023.
21. CNBC, Singapore's AI ambitions get a boost with $740 million investment plan, February 19, 2024.

AI Threat Predictions

AI-generated misinformation and cyber attacks represent #2 and #5 of the top 10 global risks in 2024, per the World Economic Global Risk Report.[22]

As the field of AI continues to rapidly evolve, including in the area of AI-generated videos and images, these risks will only grow—as will our ability to harness AI to mitigate them. Looking to the rest of 2024 and beyond, these are the top AI risk and threat predictions we see on the horizon.

1 Nation-states' AI dilemma: driving AI threats while blocking AI access

State-sponsored threat groups are poised to develop a complex relationship with AI, using it to generate more sophisticated threats while also striving to block access to anti-regime content.

Use of AI tools by state-sponsored threat groups is not a new phenomenon, but its anticipated trajectory points to significant growth in both scale and sophistication. Reports from Microsoft and OpenAI validate this concern, revealing that threat actor groups supported by nations like Russia, China, North Korea, and Iran have actively delved into and exploited ChatGPT functionality. This extends across various use cases, including spear phishing, code generation and review, and translation.

Although targeted intervention has stopped some of these attacks, enterprises should brace for the persistence of state-backed AI initiatives. The scope encompasses the deployment of popular AI tools, the creation of proprietary LLMs, and the emergence of unconstrained ChatGPT-inspired variants, such as the aptly-named FraudGPT or WormGPT. The evolving landscape paints a challenging picture in which state-sponsored actors continue to leverage AI in novel ways to create complex new cyberthreats.

2 Dark chatbots and AI-driven attacks: the scourge of "AI for bad" will grow

AI-driven attacks are likely to surge throughout the year as the dark web serves as a breeding ground for malicious chatbots like WormGPT and FraudGPT to amplify cybercriminal activities.

These insidious tools will become instrumental in executing enhanced social engineering, phishing scams, and various other threats. The dark web has seen an upswing in discussions among cybercriminals delving into the illicit deployment of ChatGPT and other generative AI tools for a spectrum of cyberattacks. More than 212 malicious LLM applications have been identified, representing only a fraction of what is available—and that number is expected to steadily grow.

Mirroring developers who use generative AI for efficiency gains, threat actors employ these tools to uncover and exploit vulnerabilities, fabricate convincing phishing schemes, execute vishing and smishing campaigns, and automate attacks with greater speed, sophistication, and scale. For example, the threat actor group Scattered Spider recently used Meta's LLaMa 2 LLM to exploit Microsoft PowerShell functionality, enabling unauthorized download of user credentials.[23] The trajectory of these advancements indicates that cyberthreats will begin to evolve more quickly than ever, taking on new forms that are more difficult to recognize or defend against with traditional security measures.

22. World Economic Forum, Global Risks Report 2024: The risks are growing — but so is our capacity to respond, January 10, 2024.
23. ZDNet, Cybercriminals are using Meta's Llama 2 AI, February 21, 2024.

3 Fighting AI with AI: security roadmaps and spend will include AI-driven defenses

Enterprises will increasingly adopt AI technologies to combat AI-driven cyberattacks, including a focus on using deep learning and AI/ML models to detect malware and ransomware hidden in encrypted traffic. Traditional detection methods will continue to struggle with new AI-driven zero-day attacks and polymorphic ransomware (which can evolve its code to evade detection), so AI-based indicators will be crucial in identifying potential threats. AI will also play a vital role in swiftly identifying and stopping convincing AI-generated phishing and other social engineering attacks.

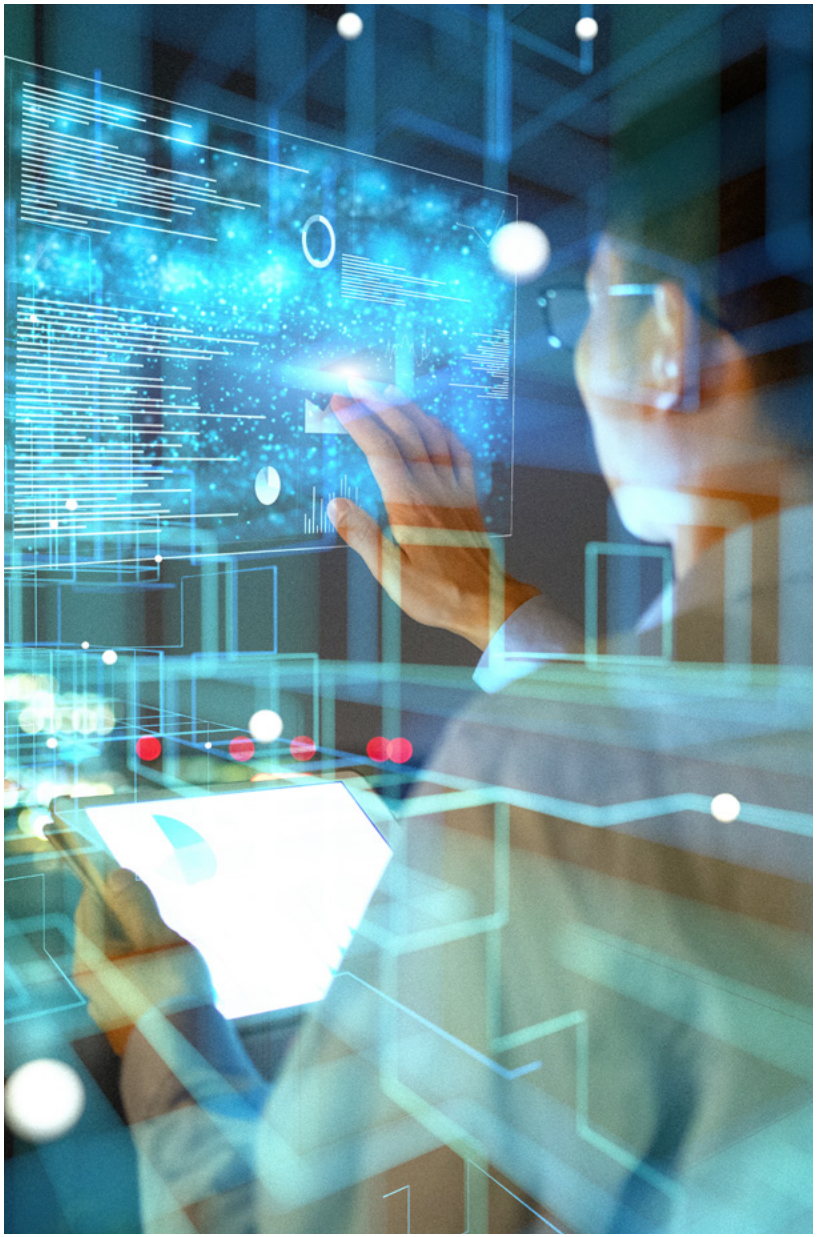
Enterprises will increasingly incorporate AI in their cybersecurity strategies. AI will be seen as a critical means to gain visibility into cyber risk as well as create actionable, quantifiable playbooks to prioritize and remediate security vulnerabilities. Translating noise into practical signals has long been a top challenge for CISOs, because correlating risk and threat information across dozens of tools can take a month or more. As such, in 2024, enterprises will look eagerly to generative AI as a way to bring order to chaos, defray cyber risk, and drive leaner, more efficient security organizations.

4 Data poisoning in AI supply chains: the risk of garbage AI data will grow

Data poisoning will become a top concern as AI supply chain attacks gain momentum. AI companies as well as their training models and downstream suppliers will be increasingly targeted by malicious actors.

The OWASP Top 10 for LLM Applications highlights training data poisoning and supply chain attacks as significant risks, running the risk of compromising the security, reliability, and performance of AI applications. Simultaneously, vulnerabilities in AI application supply chains—including technology partners, third-party data sets, and AI tool plugins or APIs—are ripe for exploitation.

Enterprises reliant on AI tools will face heightened scrutiny as they assume these tools are secure and produce accurate results. Greater vigilance in ensuring the quality, integrity, and scalability of training data sets will be essential, particularly in the realm of AI cybersecurity.



5 To leash or unleash: enterprises will weigh productivity vs. security in their use of AI tools

By now, many enterprises are past the early phases of AI tool adoption and integration, and many will have carefully considered their AI security policies. Even so, this is a fluid situation for most companies, and questions around which AI tools they will allow, which they will block, and how they will secure their data remain open.

As the number of AI tools continues to skyrocket, enterprises will need to pay close attention to the security concerns of each—at a minimum, seeking deep insight into their employees' AI usage, with an ability to enable granular access controls by department, team, and even at the user level. Enterprises may also seek more granular security controls over AI apps themselves, such as by enforcing data loss prevention policies in AI apps—preventing sensitive data from leaking—or preventing user actions such as copy and paste.

6 AI-driven deception and distortion: viral deepfakes will fuel election interference and disinformation campaigns

Emerging technologies like deepfakes pose significant threats, including election interference and the spread of misinformation. AI has already been implicated in misleading tactics during US elections, such as generating robocalls impersonating candidates to discourage voter turnout. These instances, while alarming, likely represent the tip of the AI-driven disinformation iceberg.

Furthermore, the use of AI in such schemes may not be limited to domestic actors. State-sponsored entities could also exploit these tactics to sow confusion and undermine trust in the electoral process. In a notable case, attackers utilized AI-generated deepfakes to trick an employee into transferring $25 million, demonstrating the real-world impact of this technology. Similarly, illicit deepfake images of celebrities like Taylor Swift have gone viral on social media, calling attention to how easily manipulated content can spread before content moderation measures can catch up.

Case Study: Securely Enable ChatGPT in the Enterprise

Best practices for AI integration and enterprise security policy.

By now, enterprises have had plenty of exposure to AI tools. But as the number of AI applications continues to grow dramatically and adoption continues apace, enterprises can adopt certain best practices to keep their data, employees, and customers safe. Overall, enterprises must proactively and continually adapt their AI usage and security strategies to stay ahead of evolving risks while ushering in the transformative potential of AI.

CASE STUDY

5 steps to integrate and secure generative AI tools

Enterprises seeking to securely adopt AI applications should take a measured approach. Broadly speaking, they can first block all AI applications to eliminate the risk of data leakage, and then take thoughtful steps to adopt specific, vetted AI applications with tight security controls and access control measures to maintain complete control over enterprise data. For simplicity's sake, the following journey focuses on OpenAI's LLM ChatGPT.

Step 1: Block all AI and ML domains and applications

To eliminate known and unknown risks associated with the thousands of AI applications available, enterprises can take a proactive zero trust approach, blocking all AI and ML domains and applications at the global enterprise level. This way, they can focus on adopting a minimum set of transformative AI applications while closely controlling their risks.

Step 2: Selectively vet and approve generative AI applications

Next, the organization should identify a set of generative AI applications that exceed high standards for certain criteria, such as the ability to create robust data protection, security, and contractual measures to protect enterprise and customer data, as well as the transformative potential of the applications themselves. For many enterprises, ChatGPT will be one of these applications.

Step 3: Create a private ChatGPT server instance in the corporate/DC environment

To ensure complete control over their data, organizations should host ChatGPT in a dedicated, secure tenant (such as a private Microsoft Azure AI server) hosted fully within the organization. Then, through security controls and contractual obligations, enterprises should ensure that neither Microsoft and OpenAI (in this example) has access to enterprise

or customer data, nor will enterprise user queries be used to train ChatGPT at large. This ensures the organization retains control over its training data, allowing for highly relevant, accurate answers for enterprise users while minimizing the risk of data poisoning from a public data lake.

Step 4: Move the LLM behind single sign-on (SSO) with strong multifactor authentication (MFA)

Next, the organization should move ChatGPT behind a zero trust cloud proxy architecture, such as the Zscaler Zero Trust Exchange, to enforce zero trust security controls over access to ChatGPT. This might also include moving ChatGPT behind an identity provider (IdP) with SSO authentication and strong MFA that includes biometric authentication. This will enable secure and fast user login to ChatGPT while also allowing the enterprise to configure granular access controls at the user, team, and department levels. This also ensures a separation of concerns between user queries at those same user, team, and departmental levels.

Placing ChatGPT behind a cloud proxy like the Zero Trust Exchange further enables the organization to inspect all TLS/SSL traffic between users and ChatGPT to detect cyberthreats and data leakage while applying seven distinct layers of zero trust security.

Step 5: Enforce the Zscaler DLP engine to prevent data leakages

Finally, the organization should enforce a DLP engine for the ChatGPT instance to prevent accidental leakage of critical information, including proprietary data and code, customer data, personal data, financial and legal data, and more. This ensures that any highly sensitive data will never leave the production environment.

By following this journey, enterprise users can reap the full benefits of a generative AI tool like ChatGPT while eliminating the most critical data risks of adopting an AI application.

AI best practices

In general, enterprises can adopt a few key best practices when it comes to integrating AI tools into the business.

- **Continually assess and mitigate the risks that come with AI-powered tools** to protect intellectual property, personal data, and customer information.
- **Ensure that the use of AI tools complies with relevant laws** and ethical standards, including data protection regulations and privacy laws.
- **Establish clear accountability for AI tool development and deployment,** including defined roles and responsibilities for overseeing AI projects.
- **Maintain transparency when using AI tools—**justify their use and communicate their purpose clearly to stakeholders.

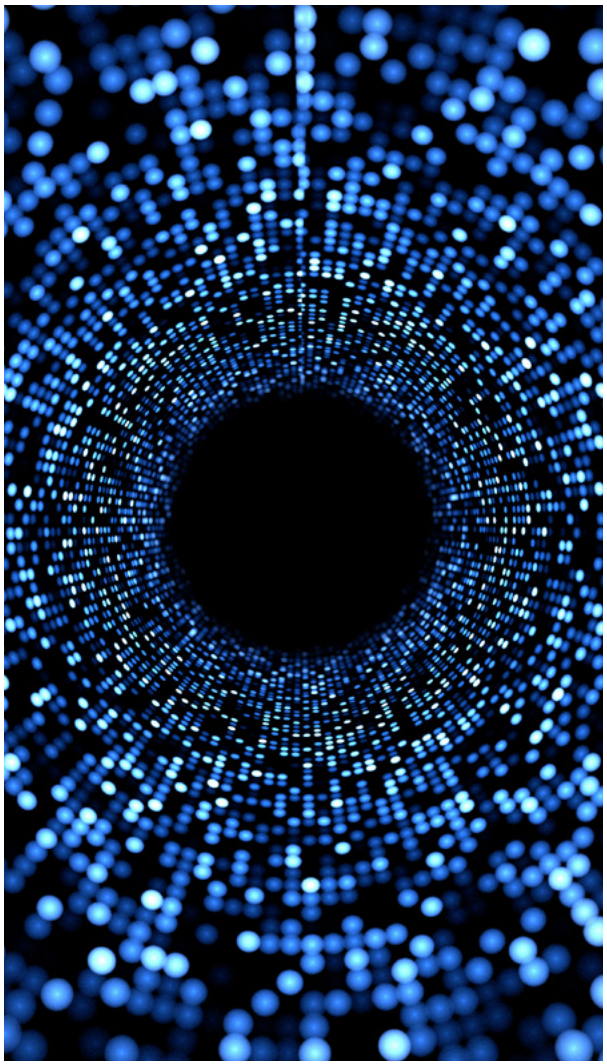
AI policy guidelines

Enterprises should go behind these best practices and establish a clear policy framework that governs enterprise-wide acceptable use, integration and product development, security and data policies, and employee best practices when using AI tools. The following best practices can form a useful starting point for establishing clear AI policies.

- **Do not provide AI models with personally identifiable information (PII)** or any non-public, proprietary, or confidential information.
- **AI cannot replace a human being,** and it should not be used to make decisions without appropriate human intervention.
- **AI-generated content should not be used without human review and approval,** especially when the content represents your organization.
- **Development and integration of AI tools should follow a Secure Product Lifecycle Framework** to guarantee the highest level of security.
- **Perform thorough product due diligence before implementing AI solutions,** making sure to evaluate their security and ethical implications.

How Zscaler Delivers AI + Zero Trust and Secures Generative AI

The transformative power of AI in cybersecurity lies in its ability to be harnessed to combat the evolving landscape of AI-driven threats. At Zscaler, we're leveraging AI to help enterprises stop attacks across all stages of the attack chain as well as easily diagnose and mitigate risk.



The key to AI-driven cybersecurity: high-quality data at scale

Enterprises generate a vast wealth of log data that can contain high-fidelity signals that may indicate likely avenues for a breach. However, signal-to-noise challenges have historically made it a challenge to isolate these signals quickly. Using generative AI, Zscaler can leverage this data to effectively enhance triage and protection measures by understanding the vulnerabilities and weaknesses attackers are likely to exploit. This not only allows Zscaler to predict breaches before they happen, but also gives executives a holistic way to visualize and quantify cyber maturity and risk while prioritizing cybersecurity remediation steps with Zscaler Risk360.

Generative AI capabilities not only extend to meta-analysis of enterprise cyber risk—they are also directly inserted into cybersecurity products to better detect and disrupt advanced threats across the attack chain. Directly integrated into the world's largest security cloud, Zscaler LLMs and AI models take advantage of a data lake that sees more than 390 billion daily transactions, with more than 9 million blocked threats and 300 trillion signals. Far from "garbage in, garbage out," this is "large-scale, high-fidelity data and threat intelligence in, finely-tuned, hyper-aware AI cybersecurity out." All of this translates to more powerful, more effective cybersecurity outcomes for IT and security practitioners.

Leveraging AI across the attack chain

We've discussed numerous ways threat actors are using AI to launch sophisticated threats at greater speed and scale. Zscaler deploys AI capabilities across the Zero Trust Exchange platform and cyber product suite to identify and stop both AI-driven and conventional attacks at each stage of the attack chain.

Stage 1: Attack surface discovery

The first stage of a cyberattack typically involves threat actors probing the internet-connected enterprise attack surface to identify exploitable weaknesses. Often, this includes things like VPN or firewall vulnerabilities and misconfigurations or unpatched servers. Generative AI has made this once-arduous task significantly easier for attackers, who can simply query a list of known vulnerabilities associated with these assets.

Leveraging AI-driven insights in Zscaler Risk360, enterprises can instantly see these discoverable (and thus risky) applications and assets—their internet-connected attack surface—and hide them from the public internet behind the Zero Trust Exchange. This instantly and dramatically reduces the enterprise attack surface while preventing attackers from ever discovering weak entry points.

Stage 2: Risk of compromise

During the compromise stage, attackers work to exploit vulnerabilities to gain unauthorized access to enterprise systems or applications. Zscaler AI innovations help reduce the risk of compromise, breaking up sophisticated attacks while prioritizing productivity.

AI-POWERED PHISHING AND C2 PREVENTION

Zscaler AI models detect known and patient-zero phishing sites to prevent credential theft and browser exploitation, as well as analyze traffic patterns, behavior, and malware to detect never-before-seen command-and-control (C2) infrastructure in real time. These models draw on a combination of threat intelligence, ThreatLabz research, and dynamic browser isolation to detect suspicious sites. As a result, enterprises are even more efficient and effective in detecting new phishing attacks, including AI-generated attacks, and C2 domains.

FILE-BASED AI SANDBOX DEFENSE

The AI-powered inline Zscaler Sandbox instantly detects malicious files while keeping employees productive. Traditional sandbox technologies make users wait while files are analyzed, or else assume patient-zero risk when files are allowed on first pass. Our AI Instant Verdict technology instantly identifies, quarantines, and prevents high-confidence malicious files—including zero-day threats—while removing the need to wait for analysis on these files. This includes threats that are delivered over encrypted channels (TLS and HTTPs) and other file transfer protocols. Meanwhile, benign files are delivered safely and instantly.

AI TO BLOCK WEB THREATS

AI-powered Zscaler Browser Isolation blocks zero-day threats while ensuring employees can access the right sites to do their jobs. In practice, enterprise URL filtering often requires more granular controls than allow/block; blocked sites are often safe and required for work, resulting in needless help desk tickets. Our AI Smart Isolation can identify when a site may be risky and open it in isolation for the user—safely streaming the site as pixels in a secure, containerized environment. This effectively stops web-based threats like malware, ransomware, phishing, and drive-by downloads, creating a strong web security posture without requiring enterprises to overblock sites as a default.

Stage 3: Lateral movement

Once attackers have a foothold inside an organization, they will try to move laterally to access sensitive data and applications. And for many organizations, user access is vastly overprovisioned to dozens of critical applications—meaning that their internal attack surface is substantial.

Zscaler AI capabilities reduce the potential blast radius of attacks by analyzing user access patterns and recommending intelligent application segmentation policies to limit lateral risk. For example, it's common to see that only 200 users out of 30,000 with access to a finance application actually need it. Zscaler can automatically create an application segment that limits access to only those 200 employees, reducing threat actors' lateral movement opportunities by more than 99%.

Stage 4: Data exfiltration

In the final stage of an attack, threat actors work to exfiltrate sensitive data. Zscaler uses AI to allow organizations to deploy data protections more quickly. AI-driven data discovery eliminates the time-consuming task of data fingerprinting and classification, which can otherwise delay or prevent deployment. Zscaler AI automatically discovers and classifies all data across an organization right out of the box, enabling enterprises to immediately classify sensitive information while configuring Data Loss Prevention (DLP) policies to prevent that data from ever leaving the organization in an attack or breach.

Summary of Zscaler's AI-infused offerings

Zscaler Internet Access™ provides AI-powered protection for enterprise users, devices, and web and SaaS applications across all locations as part of the Zero Trust Exchange, delivering:

- **AI-powered phishing and C2 detection** against never-before-seen phishing sites and C2 infrastructure, using inline AI-based detection from the Zscaler Secure Web Gateway (SWG).
- AI-powered sandboxing with comprehensive malware and zero-day threat prevention.
- **Dynamic, risk-based policy** with continuous analysis of user, device, application, and content risk to fuel dynamic security and access policy.
- **AI-powered segmentation** with Zscaler Private Access™, with automated access policy recommendations to minimize the attack surface and stop lateral movement using user context, behavior, location, location, and private app telemetry.
- AI-powered browser isolation, which creates a safe gap between users and malicious web categories, rendering content as a stream of picture-perfect images to eliminate data leaks and delivery of active threats.

MOREOVER, ZSCALER BLOCKS:

URLs and IPs observed in the Zscaler cloud as well as natively integrated open source and commercial threat intel sources. This includes policy-defined, high-risk URL categories commonly used for phishing, such as newly observed and newly activated domains.

IPS signatures developed from ThreatLabz analysis of phishing kits and pages.

Zscaler Risk360 delivers a comprehensive and actionable risk framework that helps security and business leaders to quantify and visualize cyber risk across the enterprise.

Data Protection with DLP and CASB delivers AI-powered data classification and data protection across all channels, including endpoint, email, workloads, BYOD, and cloud posture.

Advanced Threat Protection blocks all known C2 domains.

Zscaler ITDR (Identity Threat Detection and Response) mitigates the risk of identity-based attacks without ongoing visibility, risk monitoring, and threat detection.

Zscaler Firewall extends C2 protection to all ports and protocols, including emerging C2 destinations.

DNS Security defends against DNS-based attacks and exfiltration attempts.

Zscaler Private Access™ safeguards applications by limiting lateral movement with least-privileged access, user-to-app segmentation, and full inline inspection of private app traffic.

AppProtection with Zscaler Private Access provides high-performance, inline security inspection of the entire application payload to expose threats.

Zscaler Deception™ detects and contains attackers attempting to move laterally or escalate privileges by luring them with decoy servers, applications, directories, and user accounts.

Enabling the enterprise AI transition: control is in your hands

Zscaler provides a way for enterprises to foster innovation, creativity, and productivity with AI applications while keeping users and data safe among emerging channels for data exfiltration. This empowers enterprises to embrace the transformative potential of AI to accelerate their business without outright blocking AI applications and domains.

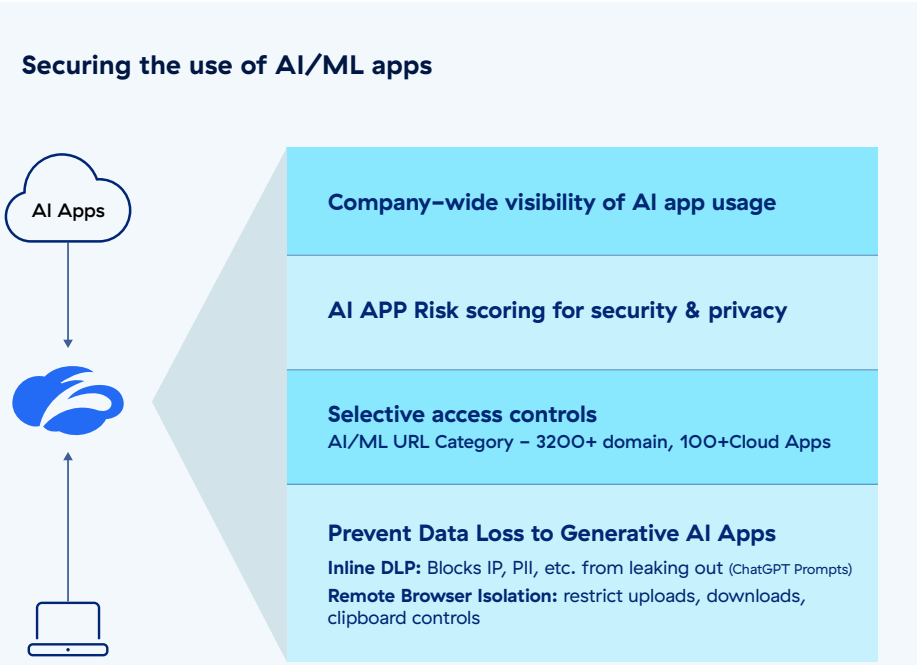


FIGURE 23 Zscaler delivers innovative, fine-tuned security controls designed for enterprise AI

ZSCALER ENABLES ENTERPRISES TO:

01 **Drive full visibility into AI tool usage**
Detailed logs provide complete visibility into how enterprise teams are using AI, including the applications and domains they're visiting as well as the data and prompts being used in tools like ChatGPT.

02 **Create flexible policies to fine-tune the use of AI**
Powerful, tailored URL filtering for AI and ML applications let enterprises easily define and enforce granular AI access controls and segmentation—blocking access when necessary, while allowing access with acceptable levels of risk using AI App Risk Scoring. Enterprises can allow access at the enterprise, department, team, and user levels as well as enable caution-based access that coaches users on the risks of generative AI tools. AI-driven segmentation makes it easy to identify appropriate user segments for access to particular AI applications while minimizing the internal attack surface associated with AI tools.

03 **Enforce granular data security for ChatGPT and other AI applications**
Enterprises can prevent the leakage of sensitive data uploaded to AI applications with granular Zscaler Cloud Application controls for generative AI. By enforcing the Zscaler DLP engine, enterprises can ensure that no data is accidentally shared when using any AI tool. Meanwhile, AI-powered data discovery and classification lets enterprises easily identify and create DLP policies around their most critical data, including their corporate code base, financial and legal documents, personal data, customer data, and more. This video demonstrates how the DLP engine prevents users from inputting credit card information into ChatGPT.

04 **Enable powerful controls using Browser Isolation**
Zscaler Browser Isolation renders AI applications in a secure environment, adding a layer of protection that allows user prompts and queries to AI tools while restricting copy/paste, uploads, and downloads. This helps mitigate the risk of sensitive data being accidentally shared with generative AI tools.

Enterprise and security leaders are at a crossroads: they must work to embrace AI to drive innovation and stay competitive, but they must also ensure that their data only powers the business, not breaches. Zscaler empowers enterprises to navigate this transition with confidence, leveraging a full-suite of AI-powered zero trust security controls that protect against AI-driven attacks while offering fine-tuned AI policies and data protections required to harness the full potential of generative AI.

Appendix

ThreatLabz research methodology

The Zscaler global security cloud processes over 300 trillion daily signals and blocks 9 billion threats and policy violations per day, with over 250,000 daily security updates. Analysis of 18.09 billion AI and ML transactions from April 2023 to January 2024 in the Zscaler cloud, the Zero Trust Exchange.

About Zscaler 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 on its portal, research.zscaler.com.

Experience your world, secured.

About Zscaler

Zscaler (NASDAQ: ZS) accelerates digital transformation so that customers can be more agile, efficient, resilient, and secure. The Zscaler Zero Trust Exchange™ 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 SASE—based Zero Trust Exchange is the world's largest inline cloud security platform. To learn more, visit **www.zscaler.com**.

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