



Zscaler ThreatLabz 2025 Data@Risk Report



Table of Contents

Introduction	3
Executive Summary	4
Key Findings	5
AI apps are a major source of sensitive data leakage	5
SaaS data loss Soars: 872M violations across 3K+ applications	6
Email remains a leading source of data loss	7
File-sharing data loss spikes: 212M violations across most-popular apps	8
The United States sees the highest proportion of data loss worldwide	9
Best Practices for Data Security and Resilience	10
Data Risk Predictions	12
How Zscaler Delivers Unified Data Security	13
Research Methodology	17
About ThreatLabz	17
About Zscaler	17

Introduction

As businesses increasingly rely on cloud-driven ecosystems and AI-powered applications, the stakes for protecting sensitive data have never been higher. The Zscaler ThreatLabz 2025 Data Risk Report sheds light on the evolving landscape of data security risks, revealing how technological advancements can create vulnerabilities across enterprise environments. From AI apps to SaaS platforms, email, and file-sharing tools, data loss incidents are occurring at an unprecedented scale, threatening sensitive personal, medical, financial, and intellectual property data.

The ThreatLabz 2025 Data@Risk Report analyzes more than 1.2 billion transactions where data loss incidents were prevented by the Zscaler Zero Trust Exchange from February 2024 to December 2024. Data loss is represented here, and blocked, by Data Loss Prevention (DLP) policy violations across the Zscaler cloud. The report uncovers critical insights, including the role of AI applications like ChatGPT and Microsoft Copilot as significant sources of sensitive data leakage, the alarming volume of data loss linked to SaaS ecosystems, the persistent security risks posed by email applications, and the significant volumes of data loss across file-sharing services like Google Drive, Microsoft SharePoint, and DropBox. Our findings are clear: the path forward requires innovative, proactive, AI-driven countermeasures to secure sensitive enterprise data in a unified way.

Executive_ Summary

Our findings indicate that:

- 1. AI applications are a major source of sensitive data leakage, with 4.2 million violations.** The most leaked sensitive data: social security numbers, full names, source code, and medical and disease-related information (PHI).
- 2. Enterprises saw more than 872 million data loss violations across more than 3,000 SaaS applications,** with Datadog, Cisco Webex, Salesforce, Microsoft SharePoint/OneDrive, and Google Drive seeing the largest share of leaks. Most often, the violations related to leaked full names, social security numbers and other national identifiers, medical data (PHI), and credit card numbers.
- 3. Email remains a leading source of data loss, with data leakage occurring in nearly 104 million transactions.** The most common incidents were related to: medical data (PII), source code, social security numbers, and financial data.

- 4. File-sharing apps like Google Drive and Microsoft OneDrive saw data leakage across 212 million transactions—**with source code alone being leaked 26.6 billion times.
- 5. Top 5 countries with the most data loss violations:** United States, India, United Kingdom, Singapore, Brazil

These findings indicate data security blind spots and point to a stark reality: that enterprises leak significant volumes of sensitive data that includes source code, social security numbers (SSNs), financial data, and more, across AI applications and core business channels. Without data security solutions in place that can discover, classify, and secure data across these channels, enterprises will be at the mercy of the applications they leak data to, as well as their supply chains.





Key Findings

1. AI apps are a major source of sensitive data leakage

While AI apps are now commonplace in the enterprise, there is an elevated and pervasive risk that sensitive user, financial, medical, intellectual property (IP), and other data is being leaked continuously to many AI tools—and not just to well known AI providers.

ThreatLabz analyzed Data Loss Prevention (DLP) policy violations across enterprise environments from February to December 2024. These violations indicate when sensitive data is leaked to an external software application. The violations serve as a proxy to understand the volume and variety of sensitive data that enterprises commonly leak to software tools—including AI.

Our findings indicate that:

- AI apps like ChatGPT, Microsoft Copilot, and Anthropic's Claude have become significant data loss sinks, with **4.2 million data loss violations across all AI tools**.

- **ChatGPT and Microsoft Copilot saw nearly 3.2 million data violations alone.** The most common sensitive data leaked were: social security numbers, other national identifiers, full names, and disease-related information (PHI).
- **The data most leaked across all AI apps:** social security numbers (SSNs), full names, source code, and disease-related and medical data (PII).
- **Top AI apps data loss:** ChatGPT, Wordtune, Microsoft Copilot, DeepL, Codeium, Claude, Synthesia, Grammarly, DataRobot, QuillBot, and Google Gemini.
- **Top file types that leak data to AI:** JSON, text files, Microsoft Excel, Microsoft PowerPoint, Microsoft Word, CSV files, HTTP non-file text data, JavaScript, ZIP files.

The pervasive leakage of sensitive data to AI applications underscores a critical vulnerability for enterprises striving to balance innovation with security. As generative AI tools and platforms increasingly infiltrate everyday workflows, the volume and diversity of data breaches highlight the urgent need for robust policies and technological safeguards. Addressing these risks isn't just about mitigating violations—it's about empowering organizations to safely harness the transformative power of AI while safeguarding the integrity of their most valuable assets and maintaining compliance.

1.3 million: The number of instances that social security numbers were leaked to AI applications in 2024.

Alt: Most common data leaked to AI tools: social security numbers, source code, and medical or for disease-related information (PHI).



2. SaaS data loss Soars: 872M violations across 3K+ applications

Enterprise data loss happens over a broad constellation of channels. Looking at SaaS applications alone, enterprises leak significant volumes of sensitive data to thousands of applications—including across critical cloud infrastructure—should they not have DLP technologies in place.

Our findings indicate that:

- **Datadog, Cisco Webex, Salesforce, Altium, Microsoft SharePoint/OneDrive, and Google Drive** saw the highest volumes of data loss violations among SaaS applications—nearly 416 million violations collectively—indicating that apps related to core business functions pose some of the most pervasive security risks.
 - » **Top SaaS apps by data loss:** Datadog, Cisco Webex, Salesforce, Altium, Microsoft SharePoint/OneDrive, Google Drive, Permutive, Acronis, FullStory, Microsoft Office 365, Slack, YouTube.
- **The most-leaked sensitive data:** social security numbers and national identifiers, credit card numbers, and medical data (PHI) were among most common data security violations from these applications.
- **Compressed file leakage is pervasive in enterprise workflows.** Data leakage in compressed GZIP files accounted for over 100 million data loss violations—a significant portion of the overall total that included over 2 million violations for social security numbers. While GZIP files are widely in use, even legacy file types can lead to data leakage. For instance, Microsoft SharePoint saw 8.5 million LZH archive file violations.
- **Unstructured text poses a significant security risk in real-time collaboration tools.** Webex, for example, saw 45.4 million text file violations, most often including SSNs and other national identities. These leaks can stem from anything like meeting notes, to unencrypted logs, to casual copy-pasting of sensitive queries. Real-time collaboration platforms now rival email in facilitating risky data exchanges.
- **Structured data leaks in JSON and HTTP signal systemic misconfigurations.** Tools like Datadog saw significant JSON leakage, logging 222 million violations, while Salesforce exposed 20.6 million HTTP form inputs, as enterprises inadvertently leaked sensitive customer data including social security numbers. These trends highlight hidden risks in backend configurations and business-critical workflows.

Most-leaked sensitive data to SaaS apps: **SSNs and national identifiers, full names, credit card data, and medical data (PII)**



3. Email remains a leading source of data loss

Despite the rise of new collaboration tools, email remains a dominant channel for business communication—and a leading source of sensitive data leakage. ThreatLabz analyzed nearly 104 million email transactions that contained DLP violations from February to December 2024 using the Zscaler Cloud Access Security Broker (CASB), across the Microsoft Exchange and Gmail applications.

The findings indicated that enterprises worldwide face billions of data loss incidents across hundreds of millions of transactions via email each year, exposing critical personal, financial, and proprietary information. As the backbone of corporate workflows, email demands heightened security measures to prevent inadvertent or malicious breaches.

- **Microsoft Exchange and Gmail applications saw data loss violations across nearly 104 million transactions** from February to December of 2024.
- **The majority of data loss violations were related to sensitive data:** medical data (PHI), source code, social security numbers, and financial data.
- **Microsoft Exchange:** 97 million transactions included data loss violations, with the most common leaked data being medical data, source code, SSNs, financial data, and full names.
- **Gmail:** 6.3 million transactions included data loss violations, with the most common leaked data being drug and disease-related information (PII), SSNs, source code, and full names.

With the ubiquity of email and its ease of use, substantial data leakages can happen over relatively few emails, or even one. In other words, even relatively few transactions can result in massive data loss violations. For example, source code was leaked in 20 million email transactions in 2024—an already substantial figure—but those transactions resulted in a whopping 3.2 billion data loss violations. Without robust, inline DLP solutions to secure email data, email will remain a leading channel for data loss for many enterprises.

The scale of email data leakage: 104 million transactions with data loss violations in 2024

Data most often leaked: medical data (PII), source code, SSNs, and financial data

4. File-sharing data loss spikes: 212M violations across most-popular apps

As indispensable tools for collaboration and efficiency, file-sharing applications are deeply integrated into enterprise workflows. However, their widespread use has also uncovered a critical vulnerability: the staggering volume of sensitive data leaks. From February to December 2024, ThreatLabz analyzed file-sharing transactions that contained DLP violations using the Zscaler CASB, across popular file-sharing applications like Google Drive and Microsoft OneDrive.

These platforms saw data leakage in 212 million transactions, with hundreds of billions of individual violations, revealing how easily critical information such as personal identifiers, medical data, financial records, and intellectual property can slip through the cracks. Of course, sharing critical data across these applications is not always nefarious. However, this dual role—enabling productivity while posing security challenges—underscores the urgent need for robust data security protections within file-sharing ecosystems.

- **File-sharing apps represent a significant data loss risk**, with policy violations across 212 million transactions.

- **Sensitive data is the most-often leaked:** full names, source code, medical data (PHI), SSNs and national identifiers, and financial data.

- **The top 5 file types containing leaked data:** .xlsx (Microsoft Excel), .docx (Microsoft Word), .pptx (Microsoft Powerpoint), .pdf (PDF files), .gz (ZIP files).

- **Data types most leaked for popular file-sharing applications:**

- » **Google Drive:** SSNs, full names, source code, medical data, financial data
- » **Microsoft OneDrive:** full names, SSNs and national identifiers, source code, medical data

- » **Box:** source code, medical data, national identifiers and SSNs, financial data

- » **Dropbox:** source code, full names, financial data, medical data, SSNs

- » **Confluence:** medical data, source code, national identifiers, financial data

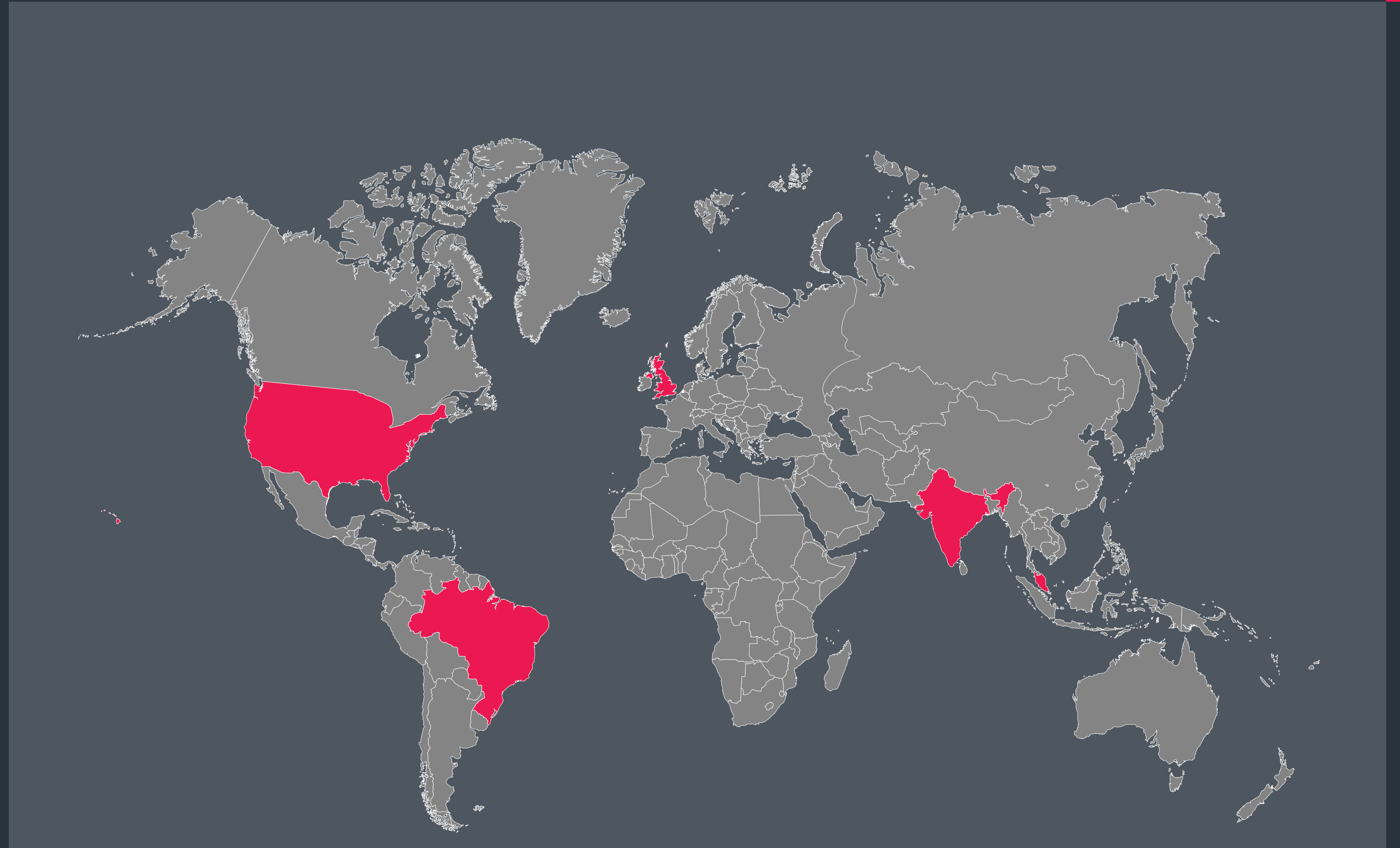
The sheer scale of data loss violations across file-sharing platforms like Google Drive and Microsoft OneDrive highlights the delicate balance between collaboration and security. As enterprises continue to rely on these tools to fuel productivity, the leakage of sensitive information—especially source code, SSNs, medical data, and financial records—serves as a stark reminder of the vulnerability within these ecosystems. Strengthening safeguards around file-sharing apps is not just a necessity; it's fundamental to ensuring that efficiency doesn't come at the cost of exposing critical data and sacrificing compliance.

Source code was leaked in 26.6 billion incidents across millions of file-sharing transactions, most often via Microsoft Excel files

5. The United States sees the highest portion of data loss worldwide

While data loss is distributed worldwide, the United States saw by far the largest portion of data loss violations across all applications.

- **Top 5 countries with the most data loss violations:** United States, India, United Kingdom, Singapore, Brazil.
- **Sensitive data most often leaked in these countries:** full names, SSNs and country identifiers, medical data (PHI), credit card data, and source code.





Best Practices for Data Security and Resilience

1 Use AI to Discover and Classify Your Data

Use automated tools like AI-powered Data Security Posture Management (DSPM) to continuously scan and classify data across endpoints, SaaS apps, and cloud services. This can help ensure patient health records, customer lists, and other sensitive data remain accurately protected, aligned with regulatory requirements like GDPR or HIPAA.

Begin by identifying and classifying sensitive data across your applications and channels. Classify data based on its location and type, such as intellectual property (source), financial data, sensitive user data, (SSNs, tax IDs), or medical records (PII). Use this classification to prioritize protection efforts and drive data loss policy creation, ensuring critical data gets the highest level of security.

2 Understand Your Data Loss Channels

Map out all the channels through which data flows within and outside your organization—email, SaaS apps, AI tools (e.g., Microsoft Copilot), BYOD, cloud storage, and physical storage devices. Each channel presents unique risks and requires tailored security controls.

3 Secure GenAI and AI Tools with Granular Controls

For generative AI tools like ChatGPT and Microsoft Copilot, enforce granular controls on user sessions, such as input or output restrictions. Block unsafe prompts that might expose sensitive data during user interactions. Additionally, monitor anomalies in user behavior (e.g., excessive queries) and flag or block activities that violate data security policies.

4 Implement a Zero Trust Architecture

Transition from a perimeter-based security model to a Zero Trust Architecture (ZTA) that enforces least-privileged access. Use identity-based access control, granular policies, and inline security services to inspect all internet traffic, segment networks, and minimize your organization's attack surface.

5 Apply In-Line Data Loss Prevention (DLP)

Integrate full-spectrum DLP to secure data across all stages—at rest, in motion, and in use. Ensure policies are consistently enforced across web interactions, enterprise applications like Salesforce and Microsoft 365, and endpoints. Regularly reevaluate and update DLP policies as new threats emerge.

6 Fix Misconfigurations and Monitor SaaS Supply Chains

Misconfigured accounts and excessive permissions in popular apps like SharePoint or cloud storage solutions are gateways for breaches. Regularly audit permissions, close oversharing loops, and monitor risks from third-party apps linked to your SaaS platforms.

7 Enforce Data Governance in Encrypted Channels

A growing number of attacks occur over encrypted channels like HTTPS. Ensure your security stack includes deep visibility to inspect and block data exfiltration even when traffic is encrypted.

8 Continuously Train Your Teams

Equip employees and contractors with data security awareness training to help them identify phishing attempts, shadow IT risks, and inadvertent data mishandling. Tailored training for roles like marketers or developers working with generative AI is also essential.

9 Proactively Plan for Supply Chain Attacks

Assume vendors in your ecosystem could expose your organization to attacks. Conduct thorough security evaluations and enforce strict contractual security requirements. Include vendors in your disaster recovery and incident response plans, and ensure Zero Trust principles are applied to external users.

Data Risk Predictions

1. **As the scale of AI data loss grows significantly, enterprises will quickly move to adopt data security controls** to prevent sensitive data leakage and understand user queries across the full gamut of AI apps.
2. **Enterprises will implement unified data security across every channel**, as the full spectrum of data loss across every channel—endpoint, inline, SaaS, file-sharing, cloud—becomes clear.
3. **The growing challenge of data loss hidden in TLS-encrypted traffic** will push enterprises to prioritize unified data security solutions that can secure encrypted data at scale..
4. **Compliance will be a major data security driver.** As many businesses face uncertainty with compliance requirements like GDPR, they will use AI-powered data discovery to automatically discover and classify sensitive data across endpoint, inline, and cloud data,
5. **Shadow AI will be a growing data security concern**, as large volumes of data loss occur through unauthorized or personal AI applications.

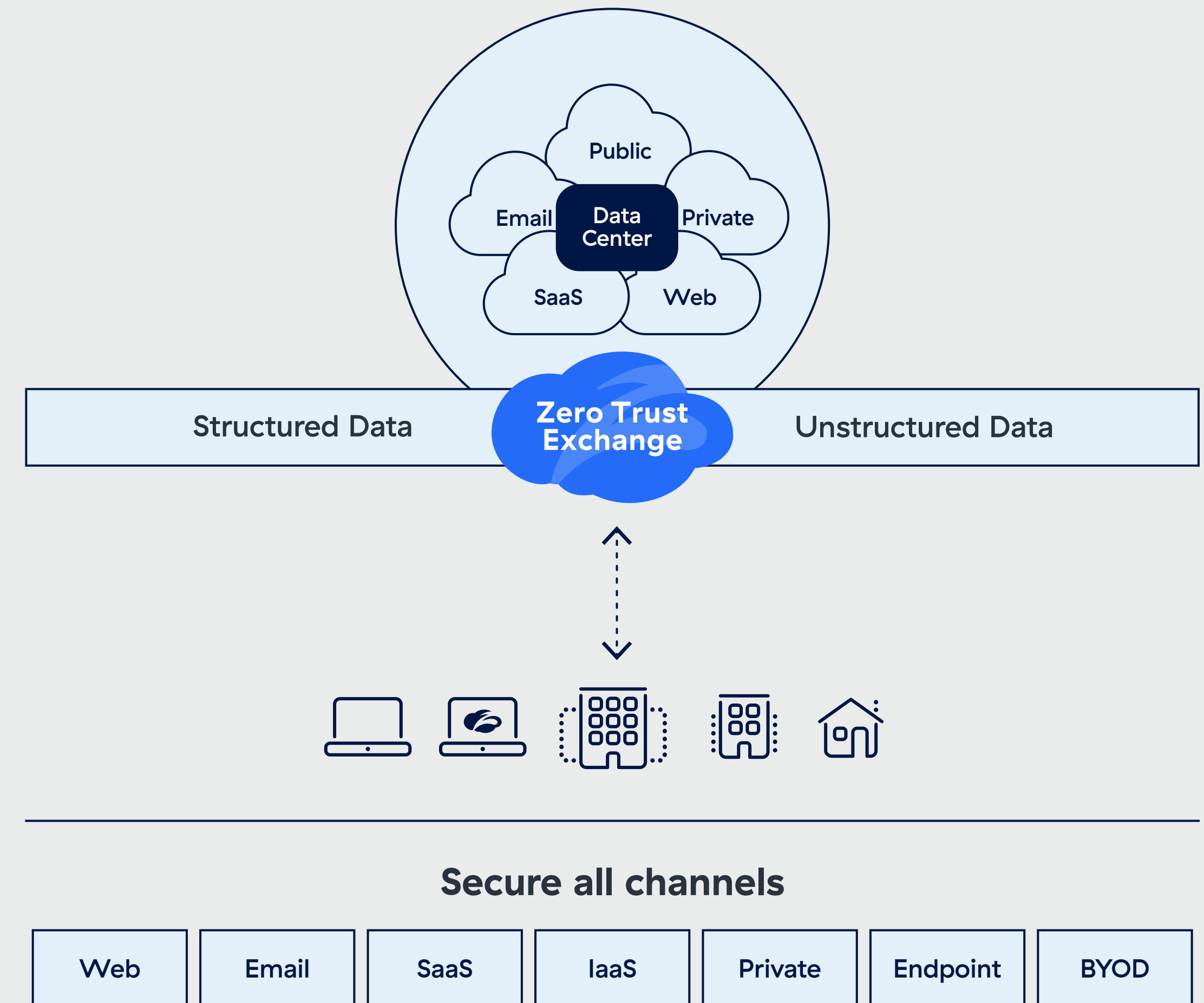
How Zscaler Delivers Unified Data Security

As enterprise AI transforms workflows and innovation, data security risks grow alongside its adoption. From sensitive prompt exposure in generative AI tools to pervasive data loss across SaaS, email, and endpoints, the modern enterprise faces unprecedented challenges. Zscaler offers best-in-class tools to secure data in this rapidly evolving landscape, providing visibility, control, and Zero Trust protection for enterprise applications worldwide.

- Find sensitive data across endpoint, inline, and cloud with AI-powered auto data discovery and classification.
- Protect data in motion with full TLS/SSL inspection and inline data loss prevention (DLP) for web, email, BYOD, and GenAI apps.
- Secure data at rest in clouds and on endpoints with unified policy, sharing controls, and device posture.

Zscaler provides unified data security for all data types, across all channels, in an AI-driven world—helping organizations strengthen their security posture wherever data resides.

The World's Most Comprehensive Data Security Platform





Protecting Enterprise AI Apps from Data Loss

AI App Visibility

As employees rapidly adopt AI tools like ChatGPT and Microsoft Copilot, Zscaler ensures enterprises never lose visibility over sensitive inputs or outputs.

- **Smart Input Prompt Blocking:** Zscaler uses AI/ML-driven URL filtering and policy enforcement to categorize AI app activity and automatically block unsafe or unapproved input prompts.
- **Deep Visibility into AI Workflows:** Innovative categorization of user prompts lets security teams track, analyze, and make educated decisions about AI application security. For instance, Zscaler policies can:
 - » Monitor for sensitive user data (e.g., SSNs) in real time.
 - » Block prompts related to intellectual property leakage.

ZIA

Dashboard

Analytics

Policies

Administration

Activation

Search

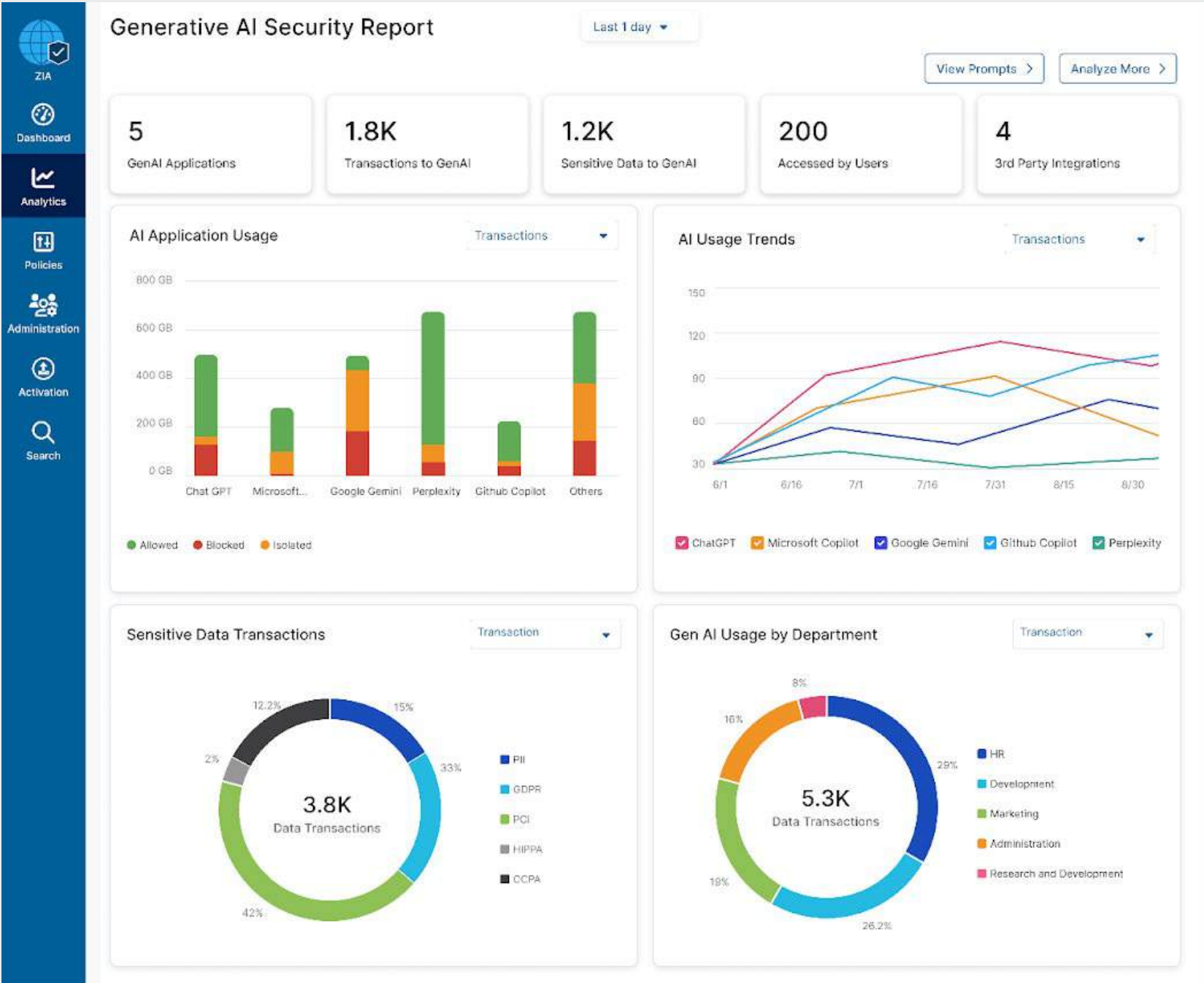
← AI Input Prompts

Department = AllApplication = AllRisk Level = AllTime Frame = Today

Q Search

User	Data Category	Risk	Department	Application	Prompt	DLP Engine
alechardy@org.	Legal	High	Product Management	ChatGPT	Write a demand letter between [party 1] and [party 2] for [consideration] for [injuries] in [jurisdiction]	PCI
john@infosys.	Business Data	Medium	Engineering	ChatGPT	Please create a customer response email to his request to bill his credit card #	PII
david@zscaler..	Financial	High	Human Resource	ChatGPT	"Analyze the performance and risk profile of my investment portfolio against the S&P 500 index."	PCI
alechardy@org.	Employee Data	Critical	Engineering	ChatGPT	Generate [number] ideas for improving an existing business in [industry]	PII
yammy@salesfo.	Real Estate	Low	Human Resource	ChatGPT	Please create a customer response email to his request to bill his credit card #	PCI

Rows per page: 1-501-50 of 236



3.8K

Data Transactions

12.2%

2%

42%

33%

15%

PII

GDPR

PCI

HIPAA

CCPA

5.3K

Data Transactions

8%

16%

10%

26.2%

29%

HR

Development

Marketing

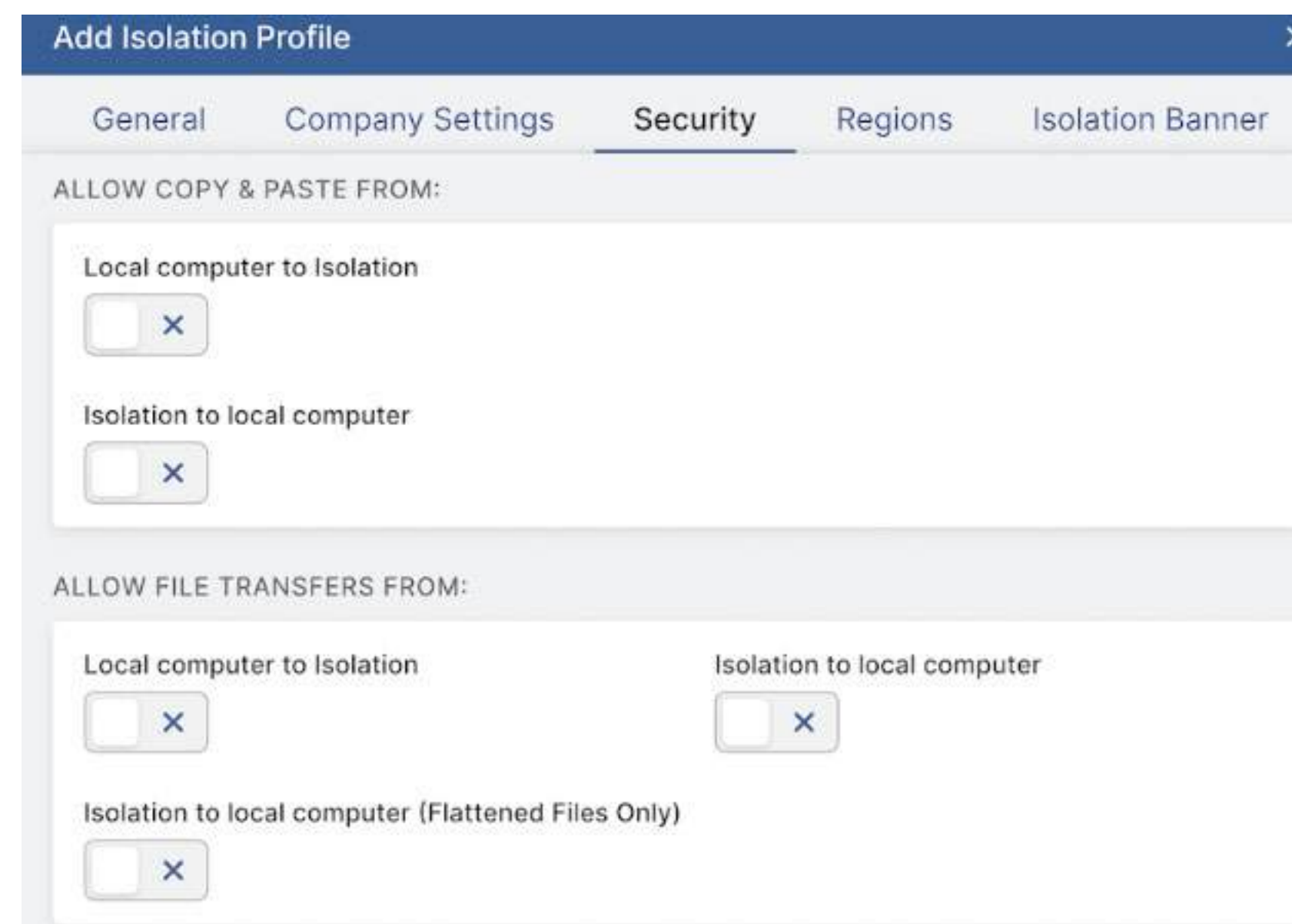
Administration

Research and Development

Secure Collaboration via Isolation

Prevent accidental data transfers in AI applications—without stifling productivity:

- **Zero Trust Browser for AI Tools:** Zscaler’s Browser Isolation technology allows employees to interact with AI tools securely by rendering applications in an isolated virtual browser.
 - » Clipboard usage, file uploads, and downloads can be restricted while still enabling prompts.
 - » Prevent accidental data exfiltration when employees interact with generative AI apps, such as ChatGPT or OpenAI-powered interfaces.
- **Safe Pixel Rendering:** By rendering applications as “pixels,” Zscaler ensures sensitive information never physically leaves the organization’s control, even during remote use.



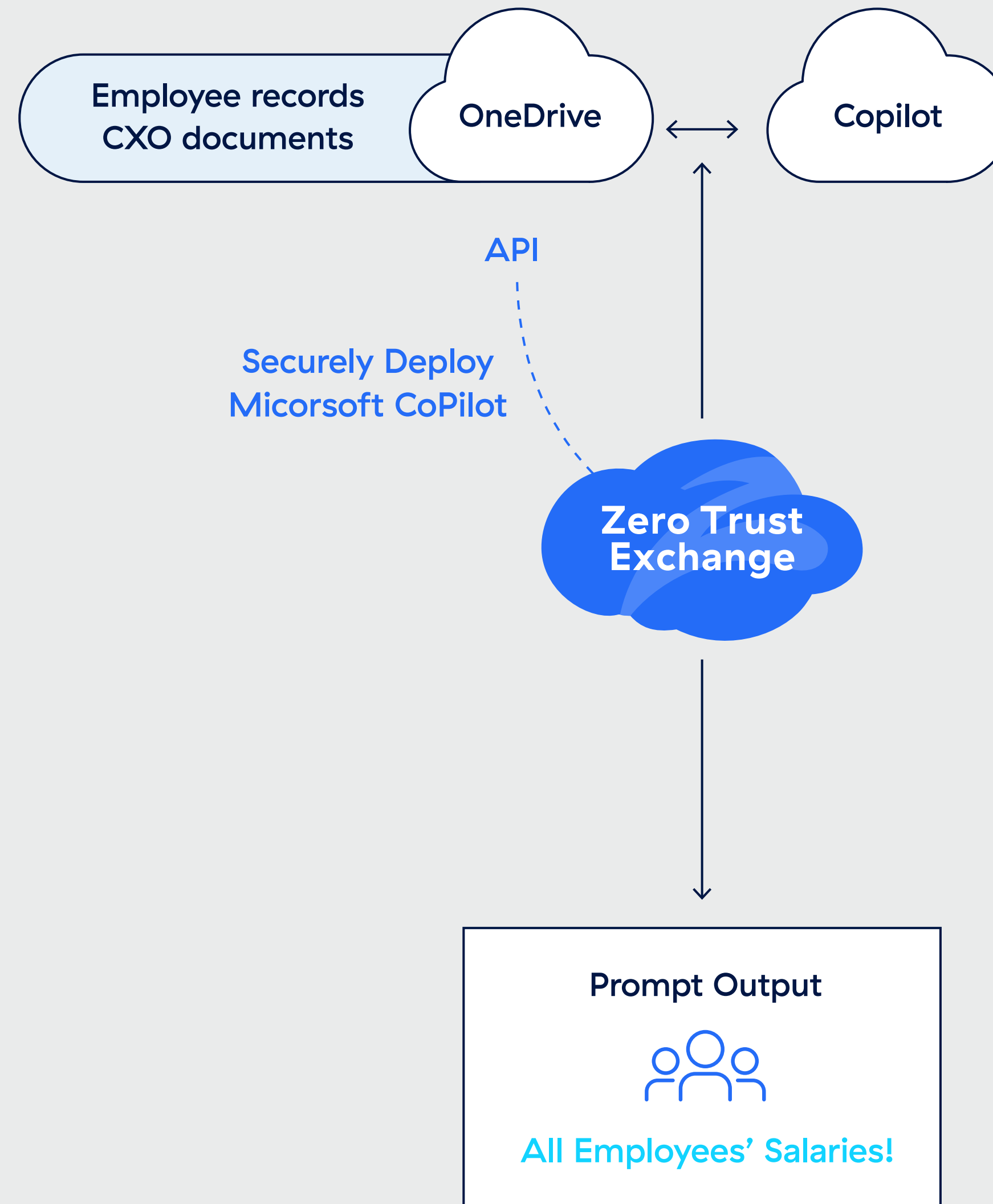
The screenshot shows the 'Add Isolation Profile' dialog box with the 'Security' tab selected. It contains two sections: 'ALLOW COPY & PASTE FROM:' and 'ALLOW FILE TRANSFERS FROM:'. Each section has three toggle switches, all of which are currently turned off. The first two switches in each section have a small 'x' icon next to them.

Section	From	To	Toggle
ALLOW COPY & PASTE FROM:	Local computer	to Isolation	<input type="checkbox"/>
	Isolation	to local computer	<input type="checkbox"/>
ALLOW FILE TRANSFERS FROM:	Local computer	to Isolation	<input type="checkbox"/>
	Isolation	to local computer	<input type="checkbox"/>
	Isolation	to local computer (Flattened Files Only)	<input type="checkbox"/>

Securing Microsoft Copilot

With Microsoft Copilot set to revolutionize enterprise productivity, Zscaler eliminates risks tied to sensitive data misuse, misconfigurations, and third-party access.

- **Inline Data Leak Prevention for Prompts:**
Zscaler scans OneDrive files and Copilot functions in real time, mapping data connections to ensure security standards. Prevent excess permissions and proactively block sensitive files from exposure.
- **Fix Misconfigurations in SaaS Settings:**
Zscaler continuously monitors configurations to resolve oversharing risks. For example, if a folder in OneDrive with client NDAs becomes write-shared through a Copilot error, Zscaler immediately flags and remediates the issue.
- **End User Behavioral Analytics (EUBA):**
Using AI-driven behavioral analytics, Zscaler identifies anomalies not only from Copilot users themselves but also any connected third-party SaaS integrations.





Research Methodology

Findings are based on an analysis of more than 1.2 billion transactions that contained Data Loss Prevention (DLP) violations in the Zscaler Zero Trust Exchange in 2024. The findings span three unique datasets, including a primary analysis of 916 million transactions that contained 4.3 billion DLP violations across the Zscaler Zero Trust Exchange from July 2024 to December 2024. In addition, ThreatLabz analyzed 104 million email transactions and 212 million file-sharing transactions that contained DLP violations using the Zscaler Multimode Cloud Access Security Broker (CASB) from February 2024 to December 2024. The Zscaler global security cloud processes more than 500 trillion daily signals and blocks 9 billion threats and policy violations per day, while 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 on its portal, research.zscaler.com.

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 160 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.



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](https://www.zscaler.com) or follow us on Twitter [@zscaler](https://twitter.com/zscaler).

© 2025 Zscaler, Inc. All rights reserved. Zscaler™ and other trademarks listed at [zscaler.com/legal/trademarks](https://www.zscaler.com/legal/trademarks) are either (i) registered trademarks or service marks or (ii) trademarks or service marks of Zscaler, Inc. in the United States and/or other countries. Any other trademarks are the properties of their respective owners.

+1 408.533.0288

Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134

[zscaler.com](https://www.zscaler.com)