AI Content Filtering and Governance Practices Using Zscaler

WHITEPAPER

Introduction

AI adoption inside the enterprise is no longer a question of if, but of how fast and how safely.

In most enterprises today, employees are already using AI tools in some capacity, often well ahead of formal review by security, legal, or compliance teams. Traditional access controls fall short here, as the primary risk lies in the content exchanged with these platforms, not merely access to them.

Three risks consistently surface as a result:

01

Shadow AI Usage

Employees engage public AI tools and AI-enabled SaaS features through personal accounts or unsantioned platforms— creating blind spots beyond the visibility of security teams.

02

Sensitive Data Leakage

Source code, customer records, financial data, and proprietary content flow into AI prompts and uploads, reaching third-party models with unclear retention and downstream exposure.

03

Lack of Governance & Auditability

Without visibility into who is using which AI tools, with what data, and to what end, organizations cannot enforce policy, demonstrate compliance, or respond to incidents.

Looking ahead, this challenge will only deepen with AI becoming a native capability woven into nearly every SaaS application and website rather than a discrete destination users navigate to— expanding what constitutes an “AI application” well beyond today’s standalone GenAI platforms.

Zscaler Internet Access

To stay ahead, Zscaler is advancing real-time, inline detection of embedded AI behavior, enabling customers to govern AI activity within any application without per-app development or signature updates, ensuring governance scales naturally as the landscape evolves.

Powering this is Zscaler Internet Access (ZIA), which sits inline across all internet-bound traffic, enforcing Zero Trust where no traffic is implicitly trusted.

From there, you gain production-grade control:

Discover

Every AI tool, embedded SaaS feature, and shadow app in use

Assess

Risk AI data handling, terms of service, and security posture

Restrict

Unsanctioned platforms outside approved use

Allow

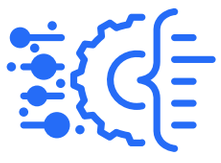
Sanctioned tools under continuous control

Govern

Approved usage with granular, context-aware policy

Protect

Sensitive data inline before exfiltration



Monitor

Telemetry, audit trails & CXO visibility

Continuously monitor and report with telemetry, audit trails, and CXO-level visibility.

The result: you don't block AI. You govern it. Discovery feeds assessment. Assessment informs policy. Policy drives enforcement. Enforcement is automated, transparent, and auditable.

The sections that follow detail the foundational requirements and how ZIA delivers each capability in practice.

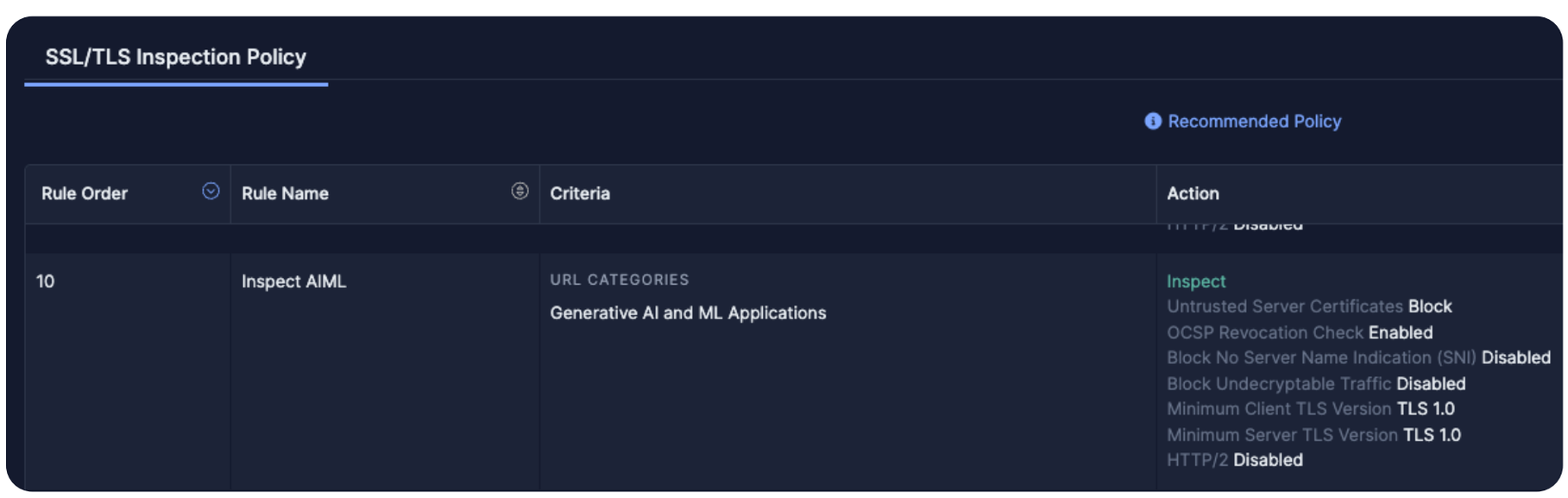
SSL/TLS Inspection: The Cornerstone of AI Governance

SSL/TLS inspection is the single most critical enabler of AI content governance. Because nearly all AI applications operate over encrypted channels, inspection is what unlocks visibility into prompts, uploads, and model responses. Without it, enforcement collapses to domain-level decisions—rendering advanced controls such as DLP, Cloud App Control, and prompt visibility ineffective.

Modern GenAI platforms increasingly rely on WebSocket-based communication for real-time interactions, where relevant payloads may not surface through traditional HTTP inspection alone. Extending inspection to WebSocket traffic ensures consistent enforcement across these protocols.

In short, SSL/TLS inspection is the foundation upon which every meaningful AI security control depends, activating it across Generative AI and ML applications is non-negotiable.

***Activation of WebSocket inspection for specific GenAI platforms (e.g., Microsoft Copilot) may require coordination with Zscaler Support to enable on the tenant.**



Policy Enforcement in Zscaler

Zscaler evaluates traffic through a structured sequence aligned with request and response flows.

During the **request phase**, traffic flows through Cloud App Control and URL Filtering for access decisions. This is followed by SSL inspection to evaluate the traffic for file type restrictions, malware, and sensitive data through DLP. For POST requests, which include prompts and uploads, DLP acts as the primary enforcement mechanism to prevent data exfiltration.

In the **response phase**, Zscaler evaluates downloaded content in reverse order to prevent threats or policy violations from reaching the user.

This model ensures that both **access and content are governed consistently across the transaction lifecycle**.

Zscaler's Approach to AI Content Filtering

With foundational visibility and traffic evaluation in place, Zscaler deploys a comprehensive defense that addresses AI security from multiple angles. Each layer builds upon the previous, creating overlapping controls that reduce risk through defense-in-depth.

1 Discovery and Classification of AI Applications

The first step in protecting against AI-related risks is understanding what AI tools exist across the organization and how they are being used. This step provides comprehensive visibility into AI adoption.

- **Cloud App Discovery:** Specifically identifies thousands of AI-related cloud apps, including those embedded within popular SaaS platforms.

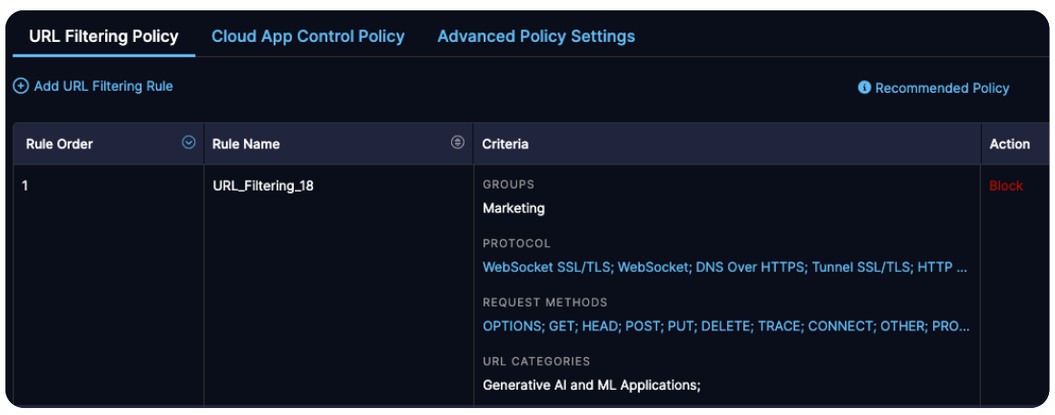
- **AI/ML URL Filtering and Categorization:** URL Filtering provides the broadest categorization layer for AI applications and forms the baseline of any AI governance strategy. Zscaler maintains dedicated categories for Generative and General AI and ML Applications, ensuring even newly emerging or uncategorized AI sites are governed by category-level posture before more granular controls are applied.

This categorization allows organizations to clearly differentiate between:

- **Generative AI platforms**, which pose higher data leakage risk

- **Non-generative AI services**, which are typically lower risk

URL Filtering enables multiple enforcement models. Organizations can completely block access to Generative AI applications or allow access while restricting interaction. A critical capability in this context is **HTTP method control**, where POST and PUT requests can be blocked to prevent users from submitting prompts or uploading files while still allowing read-only access.



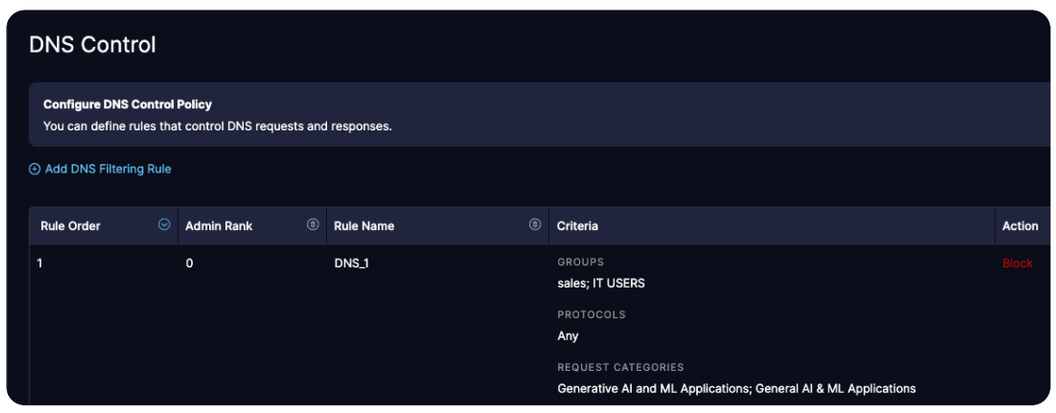
- **Shadow AI Discovery:** Provides interactive visualization of AI app usage trends, identifying which departments or users are using specific AI tools and the volume of data being sent to them, and at-risk data flows, helping uncover and remediate "shadow AI."

- **Risk Scoring:** Assigns risk levels to AI applications based on their data handling policies, terms of service, and security posture. Organizations can then make informed decisions about which tools to allow, restrict, or monitor closely.

2 Access Control and Governance for AI Usage

Once visibility is established, granular access controls determine who can use which AI platforms and what actions they can perform. This moves beyond simple 'allow/block' to context-aware governance.

- **DNS Control:** Provides an early enforcement layer before any HTTP session is established, blocking unauthorized AI domains at the resolution stage across UDP, TCP, and DNS over HTTPS (DoH). This prevents connections from being initiated, supports a deny-first posture for unvetted AI platforms, and detects DNS-based threats such as tunneling and exfiltration, offering visibility into emerging AI usage patterns before deeper inspection is required.



- **URL Filtering:** Using the AI/ML and GenAI-related URL categories, ZIA policies can allow, block, or monitor access to classes of AI sites, or specific URLs/IPs in custom categories, based on user, group, location, device posture, or risk level. Administrators can create policies that, for example, block personal GenAI tools while allowing enterprise instances, or allow only read-only access to certain AI research sites

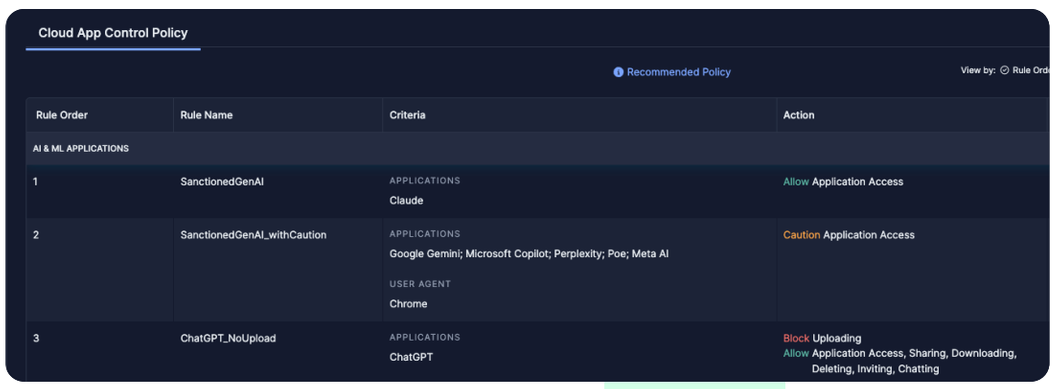
- **Cloud App Control:** Cloud App Control provides granular, application-level governance for AI platforms through Zscaler's dedicated AI/ML application database. It allows administrators to write policies that specifically target AI_ML apps with fine-grained actions:

ALLOW — Full access to the platform

CAUTION — Allow but log and alert

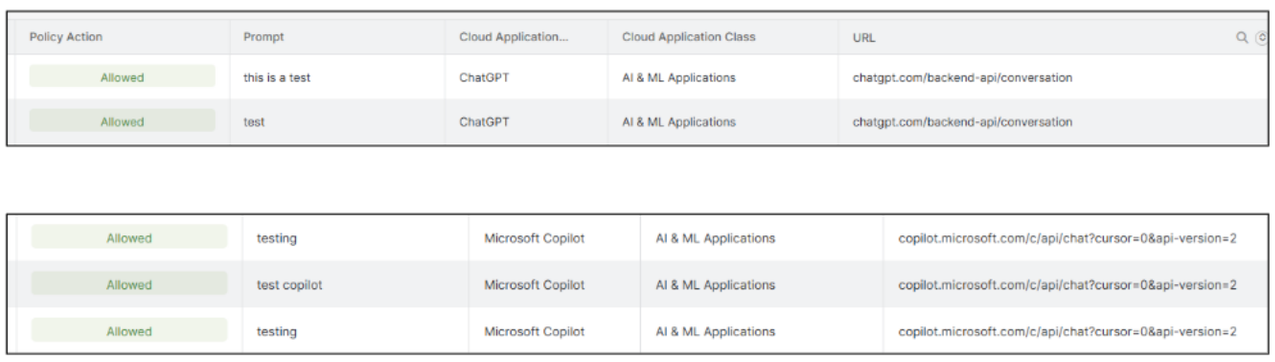
DENY — Block access

ISOLATE — Allow through browser isolation



For example, an administrator can allow a user to view ChatGPT but block them from posting content or uploading files, or allow read-only access to AI research sites.

This layer includes prompt visibility to log user inputs, providing the telemetry necessary to identify data exposure risks and analyze tool utilization.



Policy Action	Prompt	Cloud Application...	Cloud Application Class	URL
Allowed	this is a test	ChatGPT	AI & ML Applications	chatgpt.com/backend-api/conversation
Allowed	test	ChatGPT	AI & ML Applications	chatgpt.com/backend-api/conversation

Allowed	testing	Microsoft Copilot	AI & ML Applications	copilot.microsoft.com/c/api/chat?cursor=0&api-version=2
Allowed	test copilot	Microsoft Copilot	AI & ML Applications	copilot.microsoft.com/c/api/chat?cursor=0&api-version=2
Allowed	testing	Microsoft Copilot	AI & ML Applications	copilot.microsoft.com/c/api/chat?cursor=0&api-version=2

Furthermore, the Generative AI Security Report consolidates metrics on usage volume, interaction types, and behavioral risk posture, allowing organizations to refine policies based on actual adoption patterns. Ultimately, Cloud App Control serves as the primary mechanism for enforcing behavioral and contextual security across the enterprise AI ecosystem.

- User-Based Controls: Zscaler AI access security emphasizes user-based access controls, enabling organizations to allow, block, or coach access to AI applications by user, group, or department. Policies can differentiate between personas (for example, data science teams with broader access versus general employees with restricted access) and align AI usage to business need.

- Tenant Restrictions: Tenant Restrictions ensures users can only access authorized corporate AI accounts, preventing the use of personal instances on the corporate network.

  It can be enforced in two ways: for supported services, use the predefined applications available under Tenant Restrictions to ensure users can only access authorized corporate AI accounts and not personal instances. For services that aren't covered, use HTTP Header Control, especially when tenant restriction is implemented by requiring a specific (custom) HTTP header to be inserted in requests.

3 Real-Time Content Inspection & Guardrails

This section focuses on the actual content of the interaction—the prompts sent to the LLM and the responses received.

- **WebSocket Inspection for AI Chat protocols:** Since most AI assistants (like Microsoft Copilot) use persistent WebSockets, ZIA provides real-time inspection of this bidirectional traffic to enforce policies on prompts and responses.

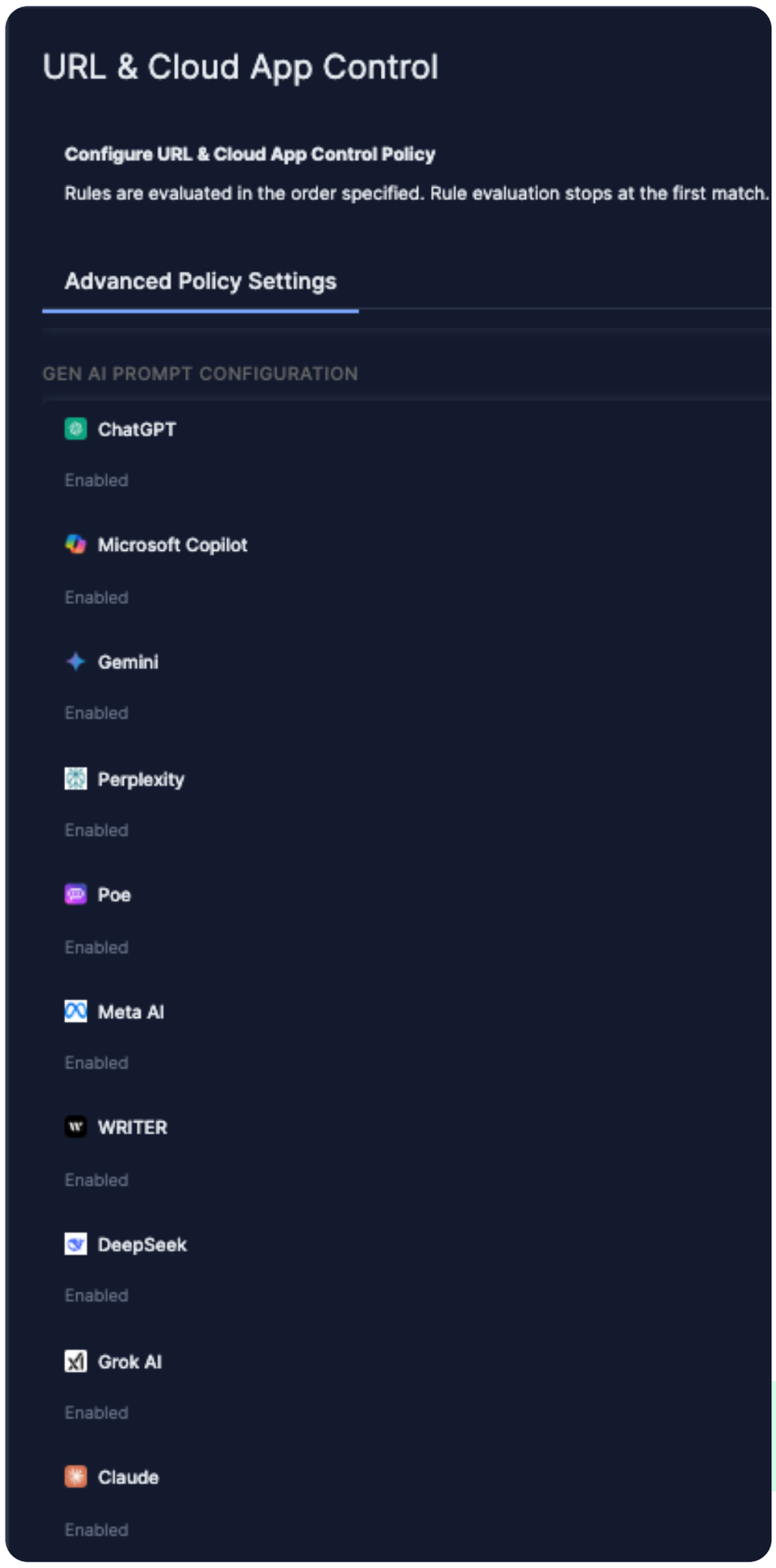
  By inspecting WebSocket traffic at scale, ZIA can identify malicious behaviors, data exfiltration attempts, or policy violations within AI interactions that would otherwise evade traditional HTTP-only controls.

- **Zscaler AI @:** A runtime guardrail solution that monitors model usage and blocks malicious or non-compliant content.

  - **Detectors:** Features multiple specialized detectors across eight languages for jailbreak attempts, prompt injections, and data poisoning.

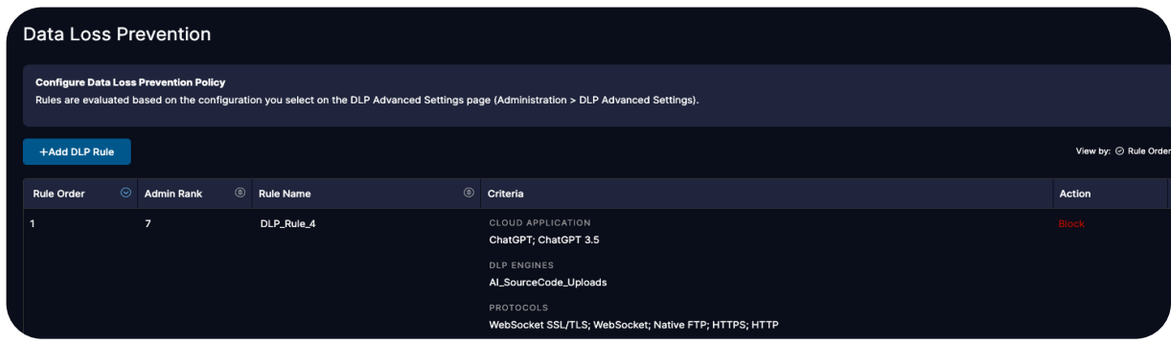
  - **Toxicity Detection:** Automatically detects and blocks toxic, restricted, or competitive topics to protect brand reputation.

- **Smart Input Prompt Visibility:** ZIA can inspect the actual text prompts sent to LLMs especially for GenAI applications such as Microsoft Copilot, providing advanced classification and inspection of prompts. These capabilities enable blocking of prompts that violate company policies, and they integrate with existing Zscaler DLP engines for additional protection of sensitive data.



4 Advanced Protection, Isolation, and Safe Access

This is the core of Zscaler's "AI Content Filtering," ensuring that sensitive organizational data does not leak into public AI models.



- **DLP for AI: Inline + OOB CASB:** DLP is the most critical control for AI content protection. It ensures that sensitive data is not exposed through user prompts or file uploads to AI platforms. Zscaler DLP uses multiple detection techniques—including dictionaries, Exact Data Match (EDM), and Indexed Document Matching (IDM)—to identify both structured and unstructured sensitive data.

  In the context of Generative AI, Inline DLP acts as the final enforcement layer in the request flow. Even when access to an application is allowed, Inline DLP prevents sensitive enterprise data from being transmitted.

  In addition, Zscaler also does OOB CASB DLP. OOB CASB DLP complements inline enforcement by providing visibility and remediation for data at rest within sanctioned SaaS/AI services (e.g., discovering sensitive data, identifying overshared content, and taking corrective actions such as quarantine, unshare, or revoke access).

  Together, Inline DLP + OOB CASB DLP provide end-to-end protection against data exposure across both in-flight AI interactions and cloud-resident content.

- **File Type Control:** File Type Control provides an additional mechanism to restrict data transfer by blocking specific file types during upload or download.This control can be used to prevent users from uploading documents such as PDFs, spreadsheets, or text files to AI platforms. While it does not provide content-level inspection, it is effective in reducing risk where DLP is not available.

- **Exact Data Match (EDM) & Index Document Match (IDM):** Advanced DLP techniques that can identify specific, unique organizational data (like a customer database or a specific design document) being sent to an AI tool.

- **Multimode CASB for SaaS data:** ZIA's data protection reference architecture supports both inline and out-of-band CASB modes:

  - **Inline CASB** for real-time app identification and control
  - **API-based CASB** for scanning data at rest in SaaS platforms

Zscaler Cloud DLP and anti-malware engines can scan files stored in cloud services that may be used for retrieval-augmented generation (RAG) or other AI workflows, enforcing consistent DLP policies across both live AI traffic and underlying data repositories.

5 Advanced Data Loss Prevention and Content Filtering

For threats that evade earlier controls and high-risk AI usage scenarios, ZIA provides advanced threat detection, zero trust egress enforcement, and safe access isolation.

- **AI-powered threat detection and intelligence:** ZIA applies AI/ML-powered threat protection, using global threat intelligence to identify behaviors tied to malicious intent within WebSocket and other encrypted AI traffic.

The platform can detect:

- **Command-and-control channels**
- **Malware delivery attempts**
- **Abuse of AI services disguised as legitimate traffic**
- **Advanced threats being delivered through AI responses**

By analyzing traffic at scale, ZIA enhances visibility into encrypted AI communications and proactively mitigates threats before they impact users or systems, without relying solely on static signatures.

- **Zero Trust Egress filtering:** Zscaler's zero trust architecture separates users and applications from the network, exposing only authorized application connections rather than broad network segments.

For AI workloads, this allows organizations to tightly control which models, APIs, and external data sources AI agents can reach, enforcing zero trust egress policies that complement AI content filtering. This model also limits lateral movement and reduces attack surface if an AI-assisted compromise occurs, as exposed applications are individually brokered and not reachable via flat networks.

Isolation and Safe Access Patterns

- **Zero Trust Browser:** For AI apps that are "borderline" or unvetted, ZIA can render the application in a secure isolated container. This allows the user to interact with the AI while preventing them from copy-pasting sensitive data into the prompt or downloading AI-generated content locally.
- **Malware Sandboxing:** Automatically quarantines unknown files or code generated by AI for dynamic analysis in an isolated environment.
- **Workflow Automation:** If a user attempts to send sensitive data to an AI app, ZIA can trigger an automated workflow to "coach" the user, requiring them to provide a justification or complete a training module before proceeding.

Recommendations and Best Practices

The practices below operationalize the controls described in this guide. Apply them to establish consistent AI governance across users, applications, and content—from baseline posture to advanced enforcement.

Enable SSL inspection for Generative AI and ML applications to ensure all content-based controls are effective.

Use Cloud App Control as the authoritative layer for sanctioned AI platforms. It evaluates before URL Filtering, making it the right place for granular, per-application governance.

Use URL Filtering for baseline AI access posture. It provides the broadest categorization across AI/ML categories, covering newly emerging or uncategorized sites. Place the governing rule at the top of the policy.

Govern by risk profile. Apply tighter controls to Generative AI and ML Applications; allow or caution General AI and ML Applications, which are typically non-generative, based on organizational need.

For full restriction of AI interaction, both categories can be denied through a single consolidated policy—reserved for regulated or high-sensitivity contexts.

Manage exceptions through Cloud App Control where available. For platforms without a corresponding cloud app, create a custom URL category, retain the parent, and permit it through a higher-priority URL Filtering rule.

Always use DLP to prevent sensitive data exfiltration through prompts and file uploads.

Configure sandbox policies to allow and scan traffic associated with Generative AI and ML applications to detect advanced threats.

For environments without DLP, **use File Type Control** to block uploads and downloads of sensitive file types for both Generative and General AI and ML categories.

To further reduce risk from interactions with Generative AI websites and AI-enabled cloud apps, we also recommend using the Zero Trust Browser. It keeps the browsing session off the endpoint and streams only a safe visual experience to the user, significantly limiting exposure to web-based threats and reducing the impact of risky interactions.

Users can also **leverage the Zero Trust Browser** to enforce read-only access for untrusted or non-sanctioned AI destinations by applying read-only isolation profiles, for example:

- Block uploads (prevent sending files to AI sites/apps)
- Block copy/paste (limit data exfiltration through prompts)
- Block downloads and printing (prevent generated content from being saved locally
- Restrict clipboard / keyboard input controls as needed

The choice isn't whether to adopt AI, it's how.

With Zscaler, enterprises adopt with confidence, governance, and full visibility.

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

zscaler™

Act Fast.
Stay Secure.™