Zscaler AI Guard

Adopt AI with Real-Time Guardrails

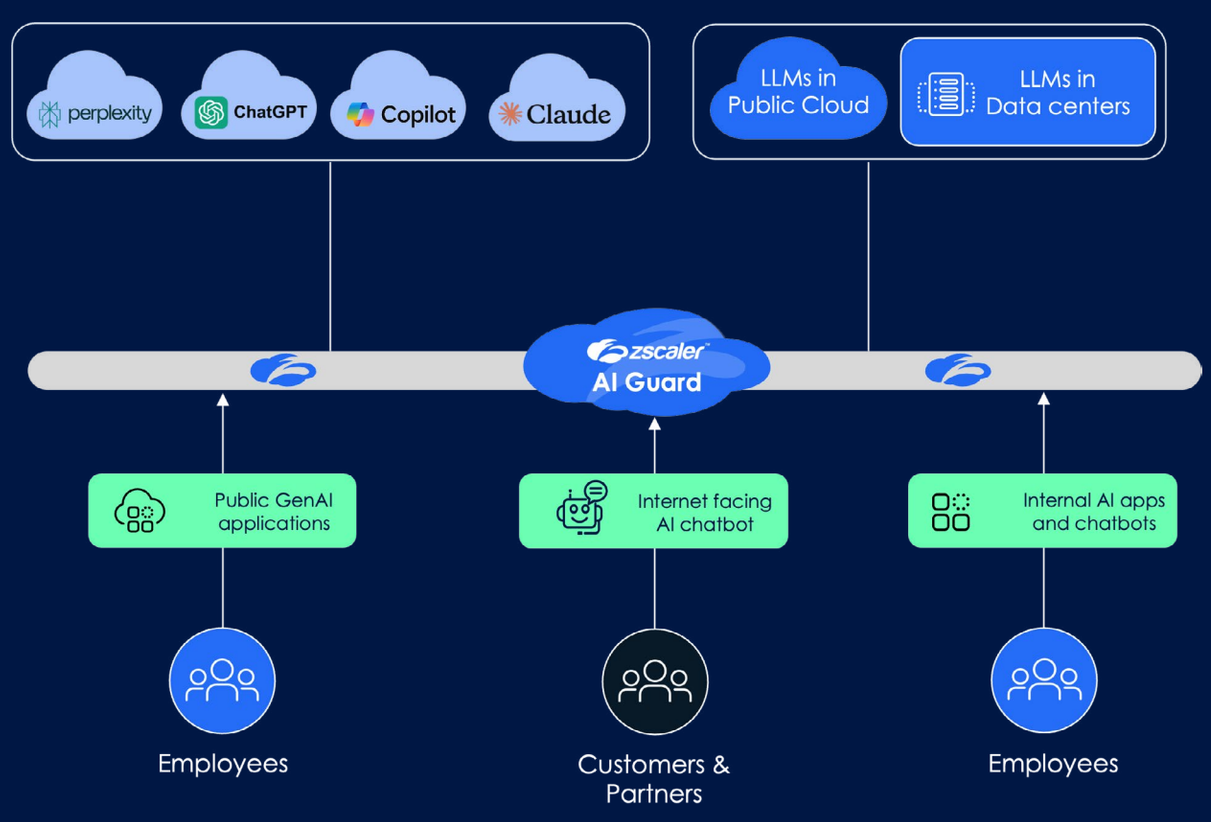
SOLUTION BRIEF

©2026 Zscaler, Inc. All rights reserved.

The proliferation of AI is creating a new set of operational, security, and compliance challenges for enterprises. Prompt Injection, jail breaking, data leakage, out-of-context responses, toxic conversations at workplace, and more are holding back enterprises from fully embracing AI. The need for robust security and visibility across these AI processes has never been greater.

Zscaler AI Guard

Zscaler AI Guard provides comprehensive runtime protection for AI user and AI applications by enforcing enterprise AI policies on the prompts and responses between end entities and LLMs. With Zscaler AI Guard, customers can confidently turn on access to popular GenAI apps and get full visibility and control into the communication between their enterprise AI apps and LLMs. From monitoring prompts and responses in real-time to blocking malicious inputs and securing sensitive data, AI Guard ensures your AI stays secure, compliant, and efficient.



Access Control

- Block
- Warn/Coach
- Isolate
- Rate Limit
- Encrypt

Data Protection

- Credentials & Secrets
- Personal Information (PII)
- Financial Data (PCI/GLBA)
- Healthcare data (PHI)
- Source Code

Content Moderation

- Context
- Toxicity
- Competition
- Brand & Reputation
- Gibberish

Threat Protection

- Prompt Injection
- Jailbreak
- Malicious URL
- Unreachable URL
- Invisible Text

©2026 Zscaler, Inc. All rights reserved.

Key Capabilities

Get Visibility & Insights

- Monitor prompts, responses, and AI model usage.
- Classify prompts and responses to uncover hidden risks or patterns.
- Get visibility into usage of popular GenAI apps such as ChatGPT, Gemini, and Microsoft Copilot and more.

Govern the use of popular AI apps

- Control who and how popular AI apps such as ChatGPT, Gemini and Microsoft Copilot are accessed and used.
- Monitor, coach, and prevent toxic, off-topic, or harmful content.
- Create and apply policies to specific users or groups of users to ensure compliance and prevent data leakage.

Secure AI apps and agents in production

- Protect your enterprise AI apps in real-time against adversarial attacks such as prompt injections, embedded code, toxicity, and malicious outputs to ensure business alignment and reduce risk.
- Integrate seamlessly with popular cloud-based AI platforms, SDKs and AI agent frameworks including AWS, Azure, Google Cloud, Langchain Agents, Palantir Webhooks, AWS Boto3, and Google Vertex SDK.
- Prevent leakage of sensitive data and information such as API keys, emails, or passwords.

Audit AI usage, ensure privacy, and enable compliance

- Maintain logs of users, apps, prompts, responses, policies, and actions.
- Enable compliance with latest standards including the NIST AI Framework and the EU AI Act.
- Zscaler does not store prompts or responses. All customer data is stored in an Amazon S3 bucket with a customer key to secure access.

Key Benefits

- **Higher accuracy and lower false positives:** AI-powered engines help block toxic or malicious inputs without overfiltering safe content. Severity settings help you set the right priorities.
- **Broad detection coverage:** More than 18 detectors with multilingual support that can be applied on prompts as well as responses to uncover a wide range of attack patterns, exposure points, and undesired content.
- **High fidelity identification for sensitive content:** Powerful safeguards that identify and prevent leakage of sensitive data – both for users accessing popular GenAI apps as well as for enterprise AI apps talking to LLMs.

©2026 Zscaler, Inc. All rights reserved.

- **Security for the complete AI lifecycle:** Zscaler AI Red Teaming and AI Guard work seamlessly to help identify vulnerabilities and secure AI apps from development through production and operation.
- **Leverage with cloud-based AI platforms and frameworks:** Easily connect to LLMs and use across popular cloud-based AI platforms and frameworks including AWS, Azure, Google Cloud, Langchain and Palantir Webhooks.
- **Compliance for AI frameworks:** Protect AI apps with compliance to latest standards and frameworks including NIST AI Framework, the EU AI Act, and OWASP Top 19 for LLM Applications.

Operating Modes

OPTION	DESCRIPTION
Proxy Mode	Intercepts and inspects LLM traffic in real-time.
Detection as a Service (DAS)	Provides detection capabilities via API for content analysis.

Detectors

DETECTOR	DESCRIPTION	PROMPT	RESPONSE
Toxicity	Detects and filters harmful language	Yes	Yes
Code	Detect and block unwanted programming languages	Yes	Yes
Prompt Injection	Detect & prevent malicious & unauthorized modifications to input prompts	Yes	N/A
Malicious URL	Identifies URLs with domains categorized as malicious	N/A	Yes
Response Tags	Filter and control prompt responses by predefined tags	N/A	Yes
Brand and reputation risk	Detects negative sentiment towards a brand	Yes	Yes
Refusal	Identifies LLM refusal		Yes
Text	Detect and block sensitive text using regex patterns	Yes	Yes
Gibberish	Identify and filter out nonsensical or meaningless text	Yes	Yes
Competition	Prevent the inclusion of competitor names in the prompts submitted by users	Yes	Yes
Language	Detect and block unwanted languages	Yes	Yes
Legal Advice	Block legal advice, interpretation, or compliance guidance	Yes	Yes

©2026 Zscaler, Inc. All rights reserved.

DETECTOR	DESCRIPTION	PROMPT	RESPONSE
Secrets	Detect and block sensitive information such as API keys	Yes	Yes
Off Topic	Filter and control content by topic description	Yes	Yes
PII	Detect and block PII entities such as email, SSN	Yes	Yes
Personal Data	Identifies sensitive personal attributes and blocks invasive questions or confirmations about identity, background, or affiliations	Yes	Yes
PII DeepScan	Detects and blocks attempts to share or solicit high-risk identifiers that directly expose financial, legal, or digital identity	Yes	Yes
Topic	Filter and control content by identifying custom topics	Yes	Yes
URL Reachability	URLs are accessible and functioning correctly by continuously testing and verifying link status in real time	N/A	Yes
Invisible Text	Identifies hidden or obscured text within digital content	Yes	Yes
Finance Advice	Blocks actionable financial guidance	Yes	Yes
Prompt Tags	Filter and control prompts by predefined tags	Yes	N/A

Prompt Privacy Options

OPTION	DESCRIPTION
No Prompt Storage	By default, prompts are inspected in real time and discarded immediately.
S3 Bucket with Customer Key	Prompts securely stored in an encrypted bucket only accessible by the customer with their own encryption keys.
Direct-to-Customer S3 Bucket	Prompts sent directly to the customer's S3 bucket for complete control.

About Zscaler

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

© 2026 Zscaler, Inc. All rights reserved. Zscaler™ and other trademarks listed at **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**

zscaler™

Zero Trust Everywhere