Zscaler AI Guard Detectors

Real-Time Protection Through Intent-Based AI Controls

WHITEPAPER

Table of Contents

Executive Summary

Artificial intelligence (AI) has rapidly become a primary interface for work as it is embedded across widely used enterprise tools, delivering on the promise to boost employee productivity. The accelerating use of AI has shifted the risk model from deterministic to probabilistic, and expands the channels through which sensitive content can be leaked or exfiltrated. Data exposure and misuse are no longer limited to structured inputs, such as files and forms. It now extends to unstructured prompts, multi-turn conversations, tool calls, model-generated outputs, and more.

As a result, new security challenges emerge that pattern-only controls struggle to reliably address. The essence of AI behavior and risk revolves around intent, context, and nondeterministic behavior.

Organizations need AI-aware guardrails capable of continuously evaluating both prompts and responses for security and policy violations with high accuracy and low latency.

Zscaler AI Guard meets this need by providing real-time, inline protection for AI interactions through specialized, intent-based detectors. Powered by task-specific AI models trained on proprietary training sets, AI Guard detectors identify and mitigate risks as they occur, ensuring safe and confident adoption of AI technologies to drive enterprise innovation.

The Evolving Risk Landscape in the Age of AI

Generative AI has become a default tool in knowledge work. Employees use it to draft code, summarize documents, analyze customer data, and automate internal processes. The productivity gains are real, but so is the exposure: every AI interaction is a data movement event, and most enterprises have limited visibility into what is being sent to which model, through which application, by whom, and for what purpose. The initial popular use of consumer-facing AI chatbots has rapidly been surpassed by enterprise AI adoption for productivity via designated apps, embedded AI feature sets, and increasingly agents. According to Zscaler ThreatLabz research, enterprise AI/ML transactions grew 147x between April 2023 and December 2025, reflecting accelerating adoption since the introduction of ChatGPT in 2022.

This growth is driven by AI adoption that falls into three categories:

- **Employees accessing public GenAI applications**, including chat interfaces, coding assistants, and productivity copilots.
- **Employees accessing SaaS applications with embedded AI**, including CRMs, productivity tools, and many more.
- **Organizations building AI-enabled applications** that call large language models (LLMs) and connect to internal tools, APIs, and data sources.

Unlike traditional SaaS applications, AI apps are designed to dynamically react to queries from users and other agents. While this dynamic behavior enables powerful AI-driven use cases and productivity gains, it also introduces fundamentally different risks that enterprises must manage.

For example, a conventional SaaS application developed for handling payroll would only handle payroll-related tasks within pre-determined workflows. An AI-enabled/embedded version of the same app could process a wide range of prompts, potentially allowing interactions that produce unintended or policy-violating outputs without proper guardrails.

These prompt-driven behaviors expand the attack surface and create new failure modes, including adversarial manipulation (e.g., jailbreaking), context leakage, and incorrect outputs (hallucinations). In pre-AI applications, the risk level of an application was relatively static and could be well understood. The risk profile of AI applications, however, changes dynamically as they are exposed to new and emerging risks through changes in inputs, integrations, and usage patterns over time.

AI-powered development adds another layer of risk. Code generated using vibe coding and other AI-driven techniques is often less rigorously vetted than traditional software code, increasing the likelihood of vulnerabilities reaching production environments.

Several AI-specific regulations have emerged to help enterprises address the growing risks posed by AI, including the National Institute of Standards and Technology (NIST) AI framework, the EU AI Act, and global standards from the International Organization for Standardization (ISO). These frameworks emphasize the need for visibility and control over AI systems and their outputs.

Security teams need controls that work for AI application and AI model behavior that is probabilistic rather than deterministic. AI guardrails play a critical role in mitigating risk by enforcing proper intent-based and behavioral controls on AI interactions in real time.

Why Intent and Context Matter for AI Security

Pattern-based inspection, such as regex, can reliably address defined risks associated with structured data inputs (e.g., credit card patterns or known secret formats) and stable, predictable data flows (e.g., files, fields, endpoints, and known apps). AI interactions are far more nuanced and unpredictable.

AI INSPECTION BLIND SPOTS

LLM interactions are dynamic semanticly rich exchanges that require inline context-aware inspection. A prompt that embeds propriety code or customer PII inside a natural-language question rarely trips a static regex-based policy.

Two prompts can contain the same words but represent entirely different intent. Conversely, the same intent can be expressed in countless different ways. This phenomenon highlights how static approaches struggle to accommodate the dynamic and variable nature of AI interactions. A single message may look harmless, with risk materializing only after several conversational turns, or depend on subtle contextual shifts, such as instruction overrides or progressive data disclosures. Even with the same prompt, a model can return different outputs depending on sampling parameters, hidden context, retrieved documents, and upstream model changes.

Under these conditions, intent and context are critical to understanding risk. Variability complicates “test once, enforce forever” approaches and renders pattern-based enforcement ineffective against advanced AI threats. Guardrails must be designed for continuous evaluation, monitoring, and adaptation, not one-time tuning.

Zscaler AI Guard

AI Guard is a product of the Zscaler AI Protect security suite, designed as an inline control plane for AI traffic. Deploy as a stand-alone solution or as an add-on for current Zscaler platform secure internet and private access customers and leverage existing identity context, device posture, and DLP classifications as inputs to AI policy without adding new infrastructure.

AI Guard runs on Zscaler’s existing globally distributed infrastructure routing detection and enforcement to the nearest edge for low latency and high availability. Policy updates propagate instantly to all enforcement points, so that new rules take effect for every user at the same time.

INLINE AI POLICY ENFORCEMENT

Detection on its own is not governance. AI Guard couples detectors to a policy engine that supports the kinds of controls AI traffic actually requires, including graduated actions, bidirectional inspection, and policy per application, identity, and group. This feeds both real-time alerting and longer-running analytics on how AI is actually being used across the organization.

Real-Time Protection for AI Interactions

AI Guard detectors are purpose-built controls that ensure robust protection for AI interactions by constantly evaluating prompts and responses for specific security and policy risks. AI Guard detectors are designed to operate in real time and support policy actions such as allow, detect/log, block, redact, or coach, based on organizational policy configurations.

Unlike traditional pattern-based approaches, AI Guard detectors leverage advanced AI models to assess the context and intent of both prompts and responses. By using AI to secure AI, the detectors more effectively mitigate the inherent risks in AI interactions.

FOCUSED DETECTION WITH SPECIALIZED MODELS

Instead of routing detection through a single general-purpose model, AI Guard deploys specialized Small Language Model detectors, each trained for a distinct detection task (e.g., PII exposure or prompt injection). This narrower scope improves signal fidelity for individual detector classes and reduces inference latency, enabling faster detection. As threats evolve, detectors can be added or refined independently, minimizing reliance on retraining a monolithic model.

Effective guardrails require high accuracy and low latency at the scale of enterprise AI traffic. Built on Small Language Models (SLMs), AI Guard detectors are fast, lightweight, and engineered for maximum availability across globally-distributed infrastructure—delivering consistent enforcement at scale with no performance tradeoffs. These capabilities make Zscaler AI Guard a cornerstone for secure, scalable AI adoption in real-world enterprise workflows.

AI Guard Detectors

The detector set maps directly to the risk surfaces that matter for enterprise AI.
The tables below summarizes the core detector classes and the enforcement actions available for each.
Please refer to the public product documentation for the full catalog of available AI Guard detectors.

Zscaler provides detectors for many use cases across data protection, content moderation, and threat protection. These detectors can be applied to:

- Prompts (User/App → LLM)
- Responses (LLM → User/App)
- Both directions (User/App → LLM → User/App)

Core AI Detectors

DETECTOR CLASS	RISK SURFACE	USE CASE	ENFORCEMENT ACTION
Prompt Injection/Jailbreak	Attempts to override system rules, extract hidden instructions, or manipulate tools. It also flags toxicity and prompts leading to malicious outcomes.	Protect internal copilots and agentic apps from jailbreaking.	Block, redact, or flag before prompt reaches the model.
Toxicity/Harassment	Detects and filters harmful language related to Hate Speech, Offensive Language, Threats of Violence, Self-Harm/Suicide, Drug Use/Abuse, Sexually Explicit, Figurative/Poetic, Psychological Abuse, Coded Toxicity, Noisy Context, Misleading Benign, Toxic Denouncing, Disability Hate, Toxic Humor/Sarcasm, Mass Violence/Genocide, Toxic Mental Health Contexts, Social Media Attacks, Multilingual Toxicity.	HR/legal/comms safety, acceptable-use enforcement, customer-facing chatbots.	Suppress response or return policy-compliant substitute.
Intellectual Property (IP)	Safeguards private and proprietary information (IP) by automatically scanning prompts against sensitive data, flagging and blocking attempts to extract, paraphrase, or directly leak source material.	Protect sensitive content.	Redact inline or block transaction based on policy.
PII	Personal data entities (varies by region/policy).	Prevent employee/ customer PII exposure; reduce compliance risk.	Redact inline or block transaction based on policy.
Topic/Off-topic Policy	Detects and filters content by identifying custom topics. Questions that are identified as related or highly similar to the topic definition will trigger the detector.	Keep enterprise copilots on-task; reduce risky use.	Enforce intended-use boundaries per application.
Secrets	API keys, tokens, credentials, private keys.	Prevent credential leakage into prompts or back to users.	Redact inline or block transaction based on policy.

DETECTOR CLASS	RISK SURFACE	USE CASE	ENFORCEMENT ACTION
Malicious/Suspicious URLs	URLs that indicate phishing or malware risk.	Reduce risk when AI returns links.	Suppress response or return policy-compliant substitute.
Gibberish/Spam /Token-burn	Nonsense prompts, automation abuse, brute-force probing.	Reduce cost and abuse; identify scripted attacks.	Enforce intended-use boundaries per application.
Code	Detects and filters code in various programming languages, including C, C#, C++, CLI commands, Go, Java, and more.	Protect IP; control code generation in regulated contexts.	Redact inline, enforce intended-use, or block transaction based on policy.

Industry-Specific AI Detectors

DETECTOR CLASS	WHAT IT HELPS DETECT	USE CASE	ENFORCEMENT ACTION
Regulated Advice (Legal/Financial)	Requests for professional advice that can create liability.	Financial services, healthcare, HR/legal-sensitive apps.	Redact inline, enforce intended-use, or block transaction based on policy.
Custom Patterns (Regex/Text)	Organization-defined patterns not covered by AI detectors.	Legacy identifiers, internal project codes.	Enforce policy or intended-use boundaries.

Detectors run in parallel on each transaction. Policy then composes their outputs.
For example, a prompt containing PII that also exhibits injection patterns can be blocked outright rather than redacted, based on the combined signal.

Detector Deployment Options

Zscaler supports two primary deployment patterns for AI Guard detectors:

1. **Users → Public GenAI applications:** inline inspection of prompts/responses at the edge/security cloud.

2. **Applications → LLMs:** API-based inspection using detections endpoints for prompts/responses and AI application traffic.

Best Practices for Deploying AI Guardrails

For teams evaluating the operational impacts of applying AI guardrails to an existing program, a few practical points tend to matter most.

1. **Start with: Detect first, then block (rollout in monitoring mode)**
   AI intent-based detectors are a relatively new introduction within the security industry, and understanding their functionality is key to effective deployment. For most organizations, the fastest path to understanding and operationalizing detectors is to:

   - Enable detectors in **detect mode** for representative users and applications to identify potential risks and establish a usage baseline.
   - **Review detection events** to understand detection patterns and make data-backed decisions about where and what type of policies to enforce first.
   - Gradually move targeted detectors or scoped user populations to stronger enforcement actions, such as **block/redact**.

2. **Apply guardrails to both prompts and responses**
   Protecting AI interactions comprehensively requires dual actions:

   - **Prompt controls** reduce the likelihood of successful attacks and prevent sensitive data from being submitted.
   - **Response controls** minimize unsafe outputs, preventing data leakage or exposure of potentially harmful responses (e.g., hallucinated or manipulated data).

3. **Minimize PII exposure through masking (prompt protection)**
   Reducing personally identifiable information (PII) exposure is critical when using public GenAI services. In addition to detecting and blocking sensitive data, Zscaler AI Guard can **mask certain PII entities in user prompts** to reduce exposure when interacting with public GenAI services. Masking replaces the original PII values with placeholders (for example, LOCATION-1) before the prompt is stored for review and/or forwarded to the LLM, depending on policy.

   AI Guard supports two independent masking controls:

   - **Dashboard masking:** PII values in prompts are concealed in the AI Guard user interface to reduce the risk of exposure during review and troubleshooting.
   - **Send masked content to the LLM:** PII values are replaced with placeholders before being forwarded to external GenAI/LLM services, reducing exposure to third-party systems.

EXPECTED BEHAVIOR SUMMARY:

ALLOW MASKING (UI)	SEND MASKED CONTENT TO LLM	WHAT APPEARS IN DASHBOARD	WHAT IS SENT TO LLM
Disabled	Disabled	Actual PII	Actual PII
Enabled	Disabled	Masked PII	Actual PII
Enabled	Enabled	Masked PII	Masked PII

4. **Scope guardrails by risk, not just by app**
   Start by segmenting policy based on risk factors rather than solely by app type:

   - **User groups:** Differentiate based on roles, such as engineering versus general users.
   - **Application sensitivity:** Tailor for public GenAI tools, internal copilots, and customer-facing assistants/agents.
   - **Data access level:** Restrict permissions based on the sensitivity of accessible data, such as internal documents, tickets, customer records, or source code.

5. **Treat RAG and tool access as security boundaries**
   Retrieval-Augmented Generation (RAG) and tool integration create new access vectors that must be tightly governed. If your AI application retrieves internal documents or calls tools:

   - **Enforce least-privilege access** for data sources.
   - **Propagate identity and policy metadata** into retrieval and prompting.
   - **Log the retrieval set** and key policy decisions for compliance and audit purposes.

6. **Use a defense-in-depth framework**
   Align AI guardrails with established risk taxonomies (e.g., prompt injection, data exposure, and insecure tool use). A layered model typically includes:

   - **Access controls:** Define who can use which AI apps/models.
   - **Prompt/response inspection (detectors):** Guard against malicious prompts and outputs.
   - **Data governance:** Apply strict controls to data retrieved by AI applications (RAG sources).
   - **Continuous monitoring:** Regularly monitor and evaluate based on usage patterns and emerging threats.

7. **Operationalize continuous improvement**
   Model behavior, usage patterns, and attack techniques evolve, requiring ongoing refinement of controls, including:

   - **Regular reviews of detection patterns** by detector, app, user group for timely refinements.
   - **Structured testing and red teaming** for new detector versions and policy updates.
   - **Cross-functional feedback loops** between security, app owners, and governance teams.

Conclusion: Putting Detectors in the Path of AI Interactions

AI changes how data moves and how users interact with systems. The attack surface is wider, and risk shows up faster. That risk now plays out in real-time interactions, where outcomes depend on context, inputs, and how systems respond over time. Protecting those AI interactions takes more than pattern matching. It requires reflexive controls that can evaluate intent and context, handle multi-turn conversations, and operate effectively even when model outputs vary.

Security has to meet AI where it is, and Zscaler AI Guard detectors make that possible. These specialized detectors enforce policy in real time, combining context- and intent-aware analysis with deterministic matchers for clearly defined data. As usage evolves and new risks emerge, they improve instead of falling behind.

For enterprises, this is what it takes to move from isolated AI use to secure adoption at scale.

ABOUT ZSCALER AI GUARD

Zscaler AI Guard is part of the Zscaler AI Protect security suite, delivering real-time governance and control across enterprise AI usage.

Learn more at zscaler.com/products-and-solutions/ai-security

About Zscaler

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

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

zscaler™

Act Fast.
Stay Secure.