PointGuard AI

# Intelligent AI Guardrails

Context-Aware Runtime Protection for Applications and Agents

## AI Guardrails Adapt to Your Needs

AI applications operate in real time, handling sensitive data and generating outputs that can directly impact users and systems. Static rules and generic filters fail to account for business context, application risk, and evolving AI behavior.

**PointGuard AI Active Guardrails** deliver effective runtime protection that adapts dynamically as your AI systems evolve. This enables your AI systems to run smoothly and securely, stopping dangerous errors without breaking legitimate AI workflows.

## Runtime Security for AI Applications

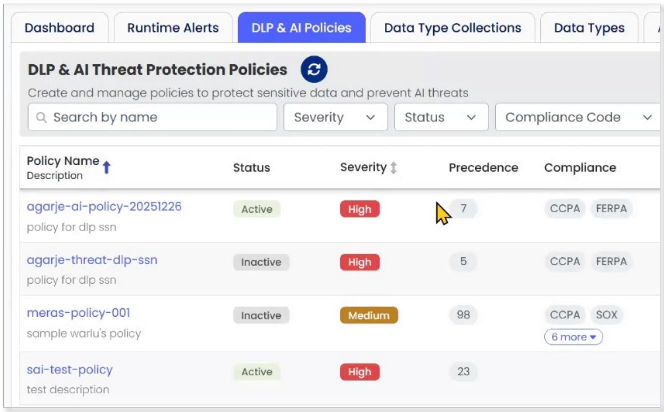
Modern AI applications rely on chatbots, models, agents, and chained interactions rather than isolated prompts. Guardrails must operate at the right points in these flows to provide effective protection for production environments. PointGuard applies guardrails to:

- **Prompt–response exchanges**
- **Chatbots and conversational interfaces**
- **AI agents and agent workflows**

## Advanced Policy-Driven Guardrails

Adaptive AI Guardrails use a flexible policy framework designed specifically for AI workloads. Policies combine context, data protection, and threat detection to deliver precise, risk-based enforcement without embedding complex security logic into applications. Policies define:

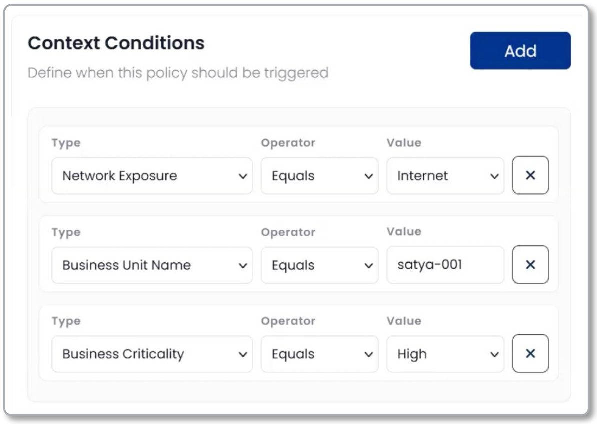
- **What to inspect**
- **Detection confidence**
- **When enforcement applies**
- **Enforcement action**



DLP & AI Threat Protection Policies

Create and manage policies to protect sensitive data and prevent AI threats

| Policy Name<br>Description | Status | Severity | Precedence | Compliance |
|---|---|---|---|---|
| agarje-ai-policy-20251226<br>policy for dlp ssn | Active | High | 7 | CCPA FERPA |
| agarje-threat-dlp-ssn<br>policy for dlp ssn | Inactive | High | 5 | CCPA FERPA |
| meras-policy-001<br>sample warlu's policy | Inactive | Medium | 98 | CCPA SOX<br>6 more |
| sai-test-policy<br>test description | Active | High | 23 | |



Context Conditions

Define when this policy should be triggered

| Type | Operator | Value |
|---|---|---|
| Network Exposure | Equals | Internet |
| Business Unit Name | Equals | satya-001 |
| Business Criticality | Equals | High |

## Protection Based on Business Context

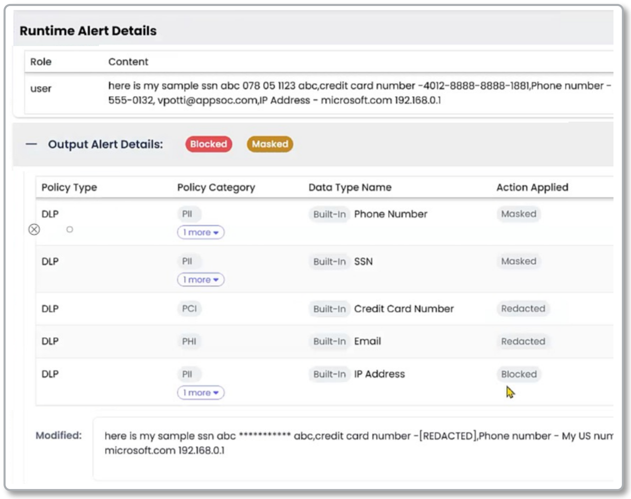
Many runtime controls treat all AI interactions but PointGuard uniquely applies business and application context to enforcement decisions, allowing organizations to tailor guardrails based on real-world risk rather than static content rules including. Context includes:

- **Application criticality**
- **Internet-facing exposure**
- **Business unit or environment**
- **Data sensitivity and compliance scope**

## Built-In and Custom Data Protection

AI systems frequently process sensitive data—often unintentionally. PointGuard provides AI-aware data protection that detects and controls sensitive information in both prompts and responses, supporting privacy and compliance requirements at runtime.

- **Built-in PII, PHI, and financial data types**
- **Custom data types and patterns**
- **Composite detection expressions**
- **Compliance mapping (e.g., HIPAA, PCI)**



**Runtime Alert Details**

| Role | Content |
|---|---|
| user | here is my sample ssn abc 078 05 1123 abc,credit card number -4012-8888-8888-1881,Phone number - 555-0132, vpotti@appsoc.com,IP Address - microsoft.com 192.168.0.1 |

— Output Alert Details: Blocked Masked

| Policy Type | Policy Category | Data Type Name | Action Applied |
|---|---|---|---|
| DLP | PII<br>1 more ▾ | Built-In Phone Number | Masked |
| DLP | PII<br>1 more ▾ | Built-In SSN | Masked |
| DLP | PCI | Built-In Credit Card Number | Redacted |
| DLP | PHI | Built-In Email | Redacted |
| DLP | PII<br>1 more ▾ | Built-In IP Address | Blocked |

Modified: here is my sample ssn abc *********** abc,credit card number -[REDACTED],Phone number - My US num microsoft.com 192.168.0.1

## Real-Time AI Threat Detection

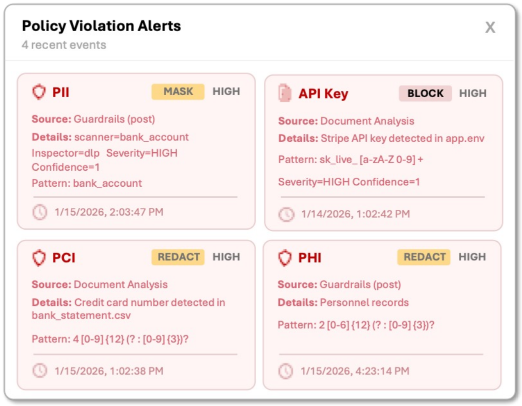
AI systems face unique threats such as prompt injection and jailbreak attempts that traditional security tools cannot detect. Adaptive AI Guardrails evaluate prompts and responses using ML-based analysis to identify malicious or unsafe behavior in real time. Threat types include:

- **Prompt injection**
- **Jailbreaking**
- **Toxic or unsafe content**
- **Gibberish and manipulation**
- **Security misuse patterns**

## Flexible Enforcement Without Disruption

Not every violation requires blocking. Adaptive AI Guardrails support multiple enforcement actions, allowing teams to balance protection and usability based on risk and context. Enforcement actions can include:

- **Block**
- **Mask / Redact**
- **Log / Alert**



**Policy Violation Alerts** X

4 recent events

**PII** MASK HIGH

**Source:** Guardrails (post)

**Details:** scanner=bank_account Inspector=dlp Severity=HIGH Confidence=1

Pattern: bank_account

1/15/2026, 2:03:47 PM

**API Key** BLOCK HIGH

**Source:** Document Analysis

**Details:** Stripe API key detected in app.env

Pattern: sk_live_ [a-zA-Z 0-9] +

Severity=HIGH Confidence=1

1/14/2026, 1:02:42 PM

**PCI** REDACT HIGH

**Source:** Document Analysis

**Details:** Credit card number detected in bank_statement.csv

Pattern: 4 [0-9] {12} (? : [0-9] {3})?

1/15/2026, 1:02:38 PM

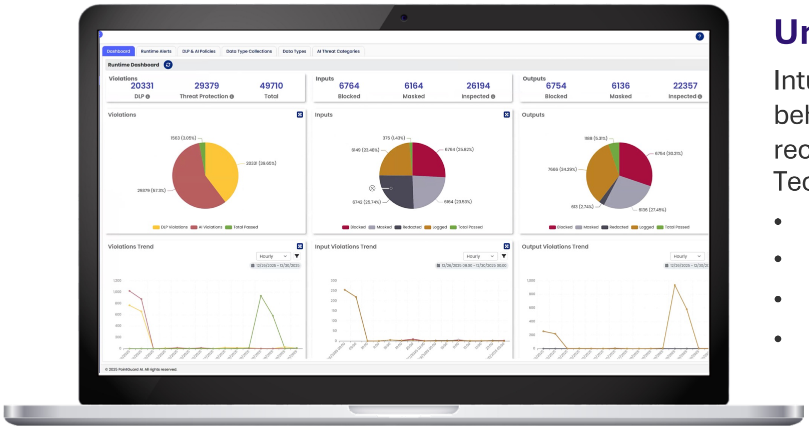
**PHI** REDACT HIGH

**Source:** Guardrails (post)

**Details:** Personnel records

Pattern: 2 [0-6] {12} (? : [0-9] {3})?

1/15/2026, 4:23:14 PM



Violations: 20331 DLP, 29379 Threat Protection, 49710 Total

Inputs: 6764 Blocked, 6164 Masked, 26194 Inspected

Outputs: 6754 Blocked, 6136 Masked, 22357 Inspected

## Unified Dashboard Reporting

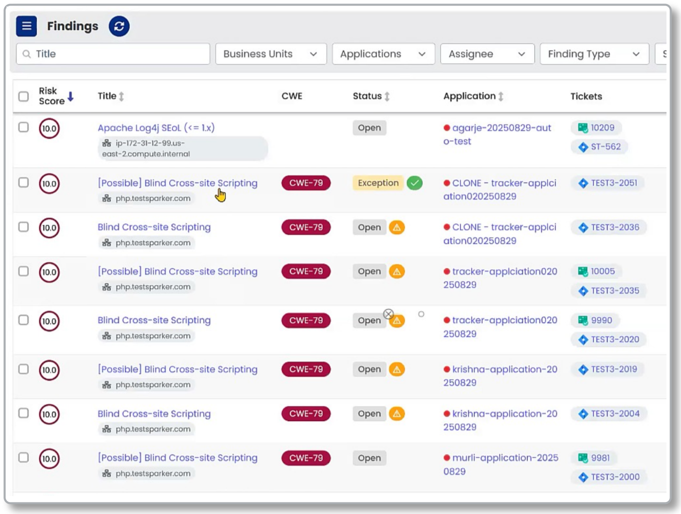
Intuitive dashboards summarize critical findings, AI behavior, and testing frequency, as well as recommendations for remediations and compliance. Technical and compliance dashboards include:

- **Prompt and response violations**
- **DLP violations and enforcement**
- **Threat types and trends**
- **Compliance & governance mapping**

## Flexible Deployment Options

Adaptive AI Guardrails are designed to integrate into production AI systems with minimal friction. Organizations can choose deployment models that align with their architecture and operational constraints, including:

- **API-based inspection of prompts and responses**
- **LLM gateway integration without code change**



Findings

Title | Business Units | Applications | Assignee | Finding Type

| ☐ | Risk Score | Title | CWE | Status | Application | Tickets |
|---|---|---|---|---|---|---|
| ☐ | 10.0 | Apache Log4j SEoL (<= 1.x)<br>ip-172-31-12-99.us-east-2.compute.internal | | Open | agarje-20250829-auto-test | 10209<br>ST-562 |
| ☐ | 10.0 | [Possible] Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Exception ✓ | CLONE - tracker-applciation020250829 | TEST3-2051 |
| ☐ | 10.0 | Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open ⚠ | CLONE - tracker-applciation020250829 | TEST3-2036 |
| ☐ | 10.0 | [Possible] Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open ⚠ | tracker-applciation020250829 | 10005<br>TEST3-2035 |
| ☐ | 10.0 | Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open ⚠ | tracker-applciation020250829 | 9990<br>TEST3-2020 |
| ☐ | 10.0 | [Possible] Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open ⚠ | krishna-application-20250829 | TEST3-2019 |
| ☐ | 10.0 | Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open ⚠ | krishna-application-20250829 | TEST3-2004 |
| ☐ | 10.0 | [Possible] Blind Cross-site Scripting<br>php.testsparker.com | CWE-79 | Open | murli-application-20250829 | 9981<br>TEST3-2000 |

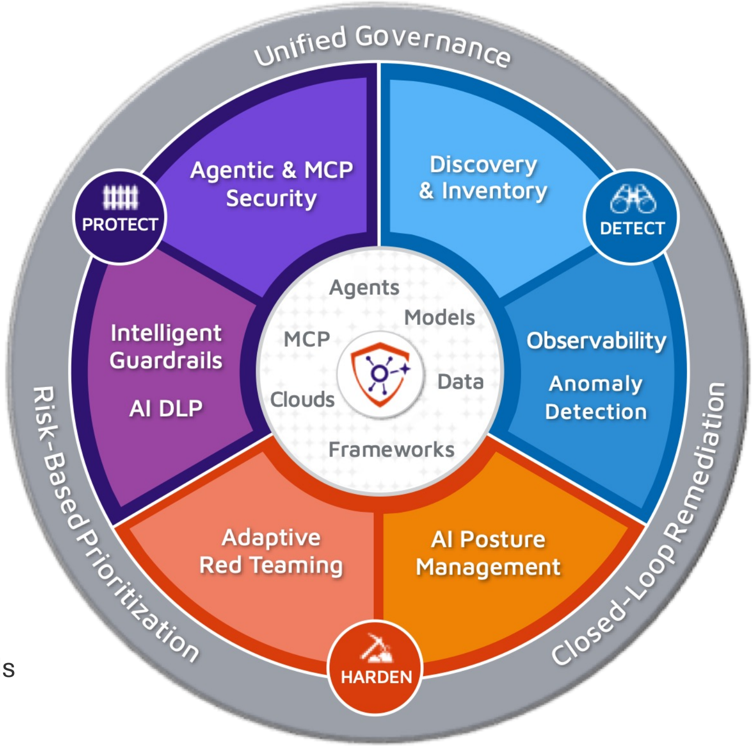
## Visibility and Operational Workflows

Runtime enforcement must be transparent and actionable. PointGuard provides detailed visibility into every guardrail decision, enabling teams to understand behavior, investigate incidents, and improve policies over time.

- **Inspect prompts and responses tied to violations**
- **Track trends across applications**
- **Manage alerts with workflows and comments**
- **Automate remediation ticketing and workflows**

## Part of the PointGuard AI Platform

PointGuard AI uniquely secures both AI systems and software applications. Through a single, unified management console all components work seamlessly together to secure the complete AI lifecycle, from discovery to data protection.



## Built for the Agentic Era

As AI systems evolve into autonomous, interconnected agents, security testing must evolve with them. PointGuard Adaptive AI Guardrails deliver context-aware, policy-driven protection designed for today's AI systems and tomorrow's agentic environments.

## Get Started

View demos and detailed technical content on our website or schedule a call to discuss your specific needs with our security experts.

www.pointguardai.com/contact

# AI Guardrails Capabilities

| Runtime Enforcement Targets | |
|---|---|
| **Prompts & Responses** | Inspects AI inputs and outputs in real time |
| **AI Agents** | Applies guardrails to agent interactions and workflows |
| **Chatbots** | Protects conversational AI interfaces |
| **Policy & Context Engine** | |
| **Application Criticality** | Differentiates enforcement by business risk |
| **Exposure Awareness** | Distinguishes internet-facing vs. internal use |
| **Environment Mapping** | Supports prod, non-prod, and business unit context |
| **AI DLP / Data Protection** | |
| **Built-In Data Types** | PII, PHI, financial, and sensitive data |
| **Custom Data Types** | User-defined patterns and classifications |
| **Composite Expressions** | Combines multiple data conditions |
| **Compliance Mapping** | Aligns data types to regulations(e.g., HIPAA, GDPR, GLBA, HITECH, PCI-DSS) |
| **AI Threat Detection** | |
| **Prompt Injection** | Detects instruction manipulation attempts |
| **Jailbreaking** | Identifies guardrail bypass techniques |
| **Toxic Content** | Flags unsafe or harmful language |
| **Gibberish & Manipulation** | Detects malformed or coercive prompts |
| **Security Misuse** | Identifies risky or malicious output patterns |
| **Enforcement Actions** | |
| **Block** | Prevents prompts or responses from proceeding |
| **Redact** | Removes sensitive content from AI traffic |
| **Mask** | Obscures specific data elements |
| **Alert** | Generates notifications for violations |
| **Deployment Options** | |
| **API-Based Enforcement** | Simple inspect call for prompts and responses with minimal code changes |
| **LLM Gateway Integration** | Works with gateways like LiteLLM |
| **Application-Centric Governance** | |
| **Application Mapping** | Associates violations to business applications |
| **Unified Risk View** | Correlates guardrails with discovery and testing |
| **Ownership & Accountability** | Aligns findings to application owners |
| **Prioritization** | Focuses remediation on highest-risk apps |