Adaptive Red Team Testing

Dynamic Adversarial Testing for Models and AI Agents

Automating Red Teaming for the AI Era

Traditional Red Teaming was designed for static code with predictable results. But today's AI systems are dynamic, agent-driven, and deeply connected to tools, MCP servers, and enterprise data. As agents evolve and policies change, one-time scans and fixed prompt libraries quickly become outdated.

PointGuard AI Red Team Testing delivers adaptive, automated adversarial testing that evolves alongside your AI systems, helping teams identify real-world risks before they impact users, data, or business operations.

Adversarial Testing for Models and Agents

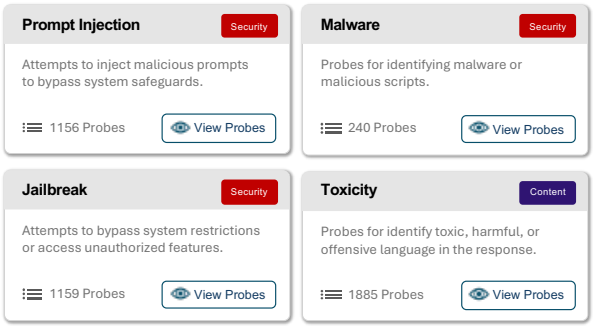
Modern AI risk extends beyond model output to agent behavior. Agents make decisions, invoke tools, retrieve data, and execute workflows, introducing new attack surfaces traditional testing can't cover. This includes:

- **Foundation and fine-tuned models**
- **Chatbots and conversational AI**
- **AI agents and agent workflows**
- **Microsoft Copilot Studio agents**

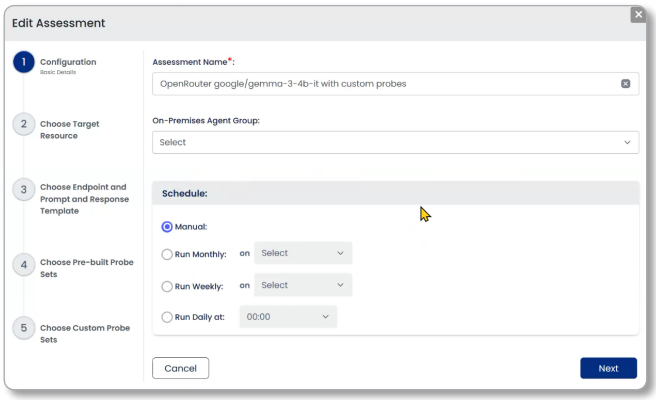
Comprehensive Threat Coverage

PointGuard maintains thousands of continuously updated adversarial probes that simulate real-world attacks at scale. Out-of-the-box categories include:

- **Jailbreak and prompt injection**
- **Toxicity, bias, and harmful content**
- **Hallucination and misinformation**
- **Malware and security misuse**



Sample probe sets



Probe configuration

Dynamic, Policy-Driven Probe Generation

Static probes can't keep pace with evolving AI systems or business risk. PointGuard combines **out-of-the-box tests**, **customer-imported tests**, and **dynamic, policy-driven generation** to deliver scalable, business-specific red teaming with minimal manual effort.

- **Use built-in attack libraries for immediate coverage**
- **Import and automate existing customer tests**
- **Generate probes from policies, or descriptions**
- **Customize attack intent, vectors, and domains**

Customers select **categories, not individual prompts**, ensuring consistent coverage while reducing manual work.

www.pointguardai.com info@pointguardai.com

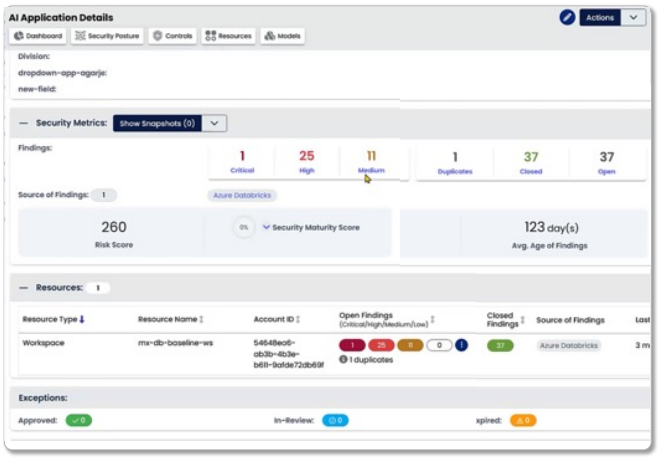
Adaptive AI Red Team Testing

Automate Continuous Testing

AI systems change constantly with new models, prompts, tools, and data. PointGuard automates continuous testing across built-in tests, customer-defined tests, and dynamically generated probes—eliminating manual re-testing. Teams can review:

- **Exact prompts and model responses**
- **Why tests passed or failed**
- **Trends and patterns across attack categories**
- **Regressions and improvements over time**

This accelerates remediation and strengthens AI defenses as systems evolve.

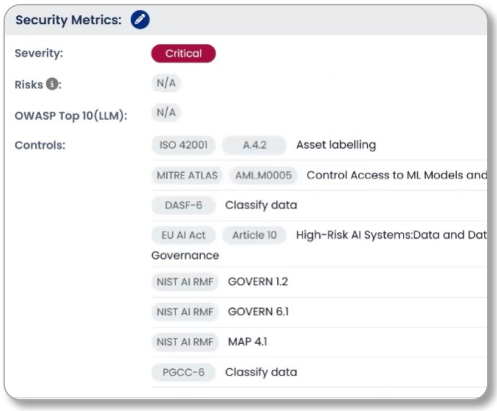


Compliance and Control Mapping

The solution maps adversarial findings to common frameworks and control categories, providing governance and audit value beyond technical testing.

Results are mapped to multiple frameworks to streamline compliance including:

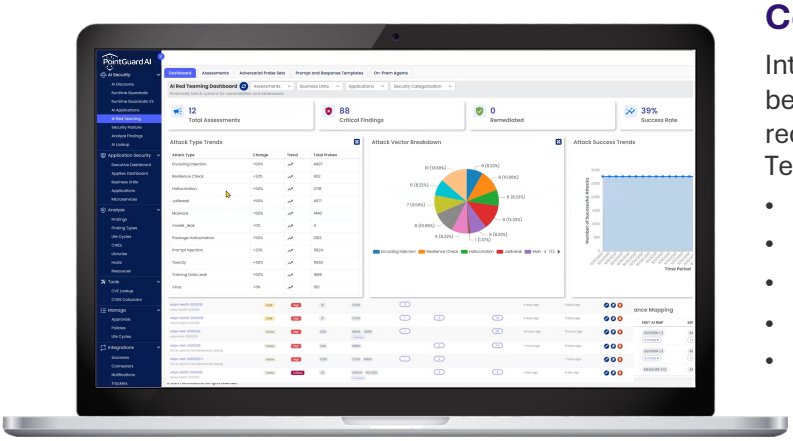
- **OWASP Top 10 for LLMs and Agentic AI**
- **NIST AI RMF**
- **MITRE ATLAS**
- **EU AI Act**
- **ISO 42001**



Application-Centric Risk Context

Red team findings don't exist in isolation. PointGuard maps testing results to models, agents, and applications, providing critical context on business impact.

- **Aggregate results at the application level**
- **Prioritize remediation based on business risk**
- **Align testing with governance workflows**



Comprehensive 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:

- **Attack types and trends**
- **Critical findings prioritized by severity**
- **AI behavior audits**
- **Compliance & governance mapping**
- **Testing and remediation trends over time**

Adaptive AI Red Team Testing

Risk-Based Prioritization

The PointGuard platform reduces noise and improves efficiency by prioritizing alerts based on business impact, severity, and exploitability. This process effectively:

- Reduces noise over 95%
- Consolidates and reduces numbers of tickets
- Improves collaboration with cleaner data
- Reduces response time and MTTR

Model Version	Content Filters	AppSOC Risk Score	Test Date
DeepSeek-R1 on Azure	Azure Filters/Guardrails OFF	8.4 / 10 = High Risk	Mar 10, 2025
DeepSeek-R1 on Azure	Azure Filters/Guardrails ON	8.3 / 10 = High Risk	Mar 10, 2025

Threat Category	Test Definition	Azure Filters OFF	Azure Filters ON
Jailbreak	Prompts cause model to disregard system prompts/guardrails	37.6%	5.0%
Prompt Injection	Prompts cause ignored guardrails, leaked data, or subverted behavior	57.1%	40.0%
Malware	Model can generate code for disabling antivirus, hiding in process list, etc.	96.7%	93.8%
Supply Chain	Model hallucinates, making unsafe software package recommendations	5.8%	6.9%
Toxicity	AI-trained prompts result in model generating toxic output	14.8%	4.0%
Training Data Leak	Prompts result in model leaking training data	32.7%	10.0%
Virus	Prompts result in model generating virus code	93.3%	93.3%
Hallucination	False prompts result in model hallucination	50.4%	0.0%

< 10% | 10-30% | 30-70% | > 70%

Sample model risk report

Closed Loop Remediation

Detecting potential threats is only half the solution. PointGuard integrates out-of-the-box with leading notification and ticketing systems. Automated remediation workflows notify stakeholders, open tickets, suggest remediation steps, and track SLAs.

- **Notifications: Slack, MS Teams, PagerDuty**
- **Ticketing: Jira, ServiceNow, Azure Boards**

The platform ensures that security findings are managed, tracked, and efficiently remediated.

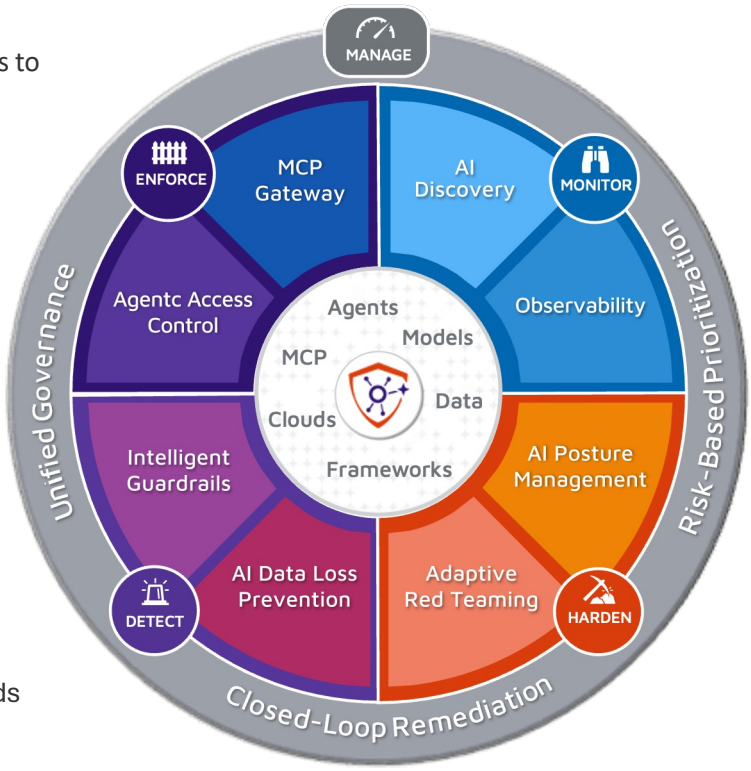
Multiple Deployment Options

The Red Teaming module can be deployed in multiple ways to ensure customer control and data sovereignty including:

- **On-premises or in customer cloud**
- **Hybrid data plane in customer cloud**
- **Fully SaaS-based**

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.



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 Red Teaming Capabilities

Testing Targets	
AI Models	Foundation and fine-tuned LLMs, in-house, open-source, hosted
AI Agents	Autonomous and semi-autonomous agents
Copilot Studio Agents	Native support via dedicated connector
AI Applications	Testing results mapped to business applications and platforms
Platform Capabilities	
Out-of-the-Box Probes	Libraries of thousands of maintained adversarial probes
Dynamic Probe Generation	AI-generated probes from customer policies and specifications
Custom Probes	Customer-defined prompts and scenarios
Scheduled Assessments	Continuous and recurring testing
Compliance Frameworks	OWASP Top 10 for LLMs, OWASP Top 10 for Agents, MITRE ATLAS, NIST AI RMF, ISO 42001, EU AI Act, GDPR, Databricks DASF
Platform Integrations	Databricks, Azure AI Foundry, Copilot Studio, Amazon Bedrock & SageMaker, Google Cloud, Vertex.ai, LangChain, LangGraph, CrewAI
Remediation Integrations	Jira, ServiceNow, Azure Boards, Slack, PagerDuty, MS Teams
Deployment Options	SaaS, Hybrid (data plane in customer cloud), On-premises
Attack & Risk Categories	
Prompt Injection	Examines if prompts contain payloads that could manipulate model behavior.
Jailbreak	Determines if manipulated inputs can cause models to bypass guardrails.
Encoding Injection	Detects malicious payloads hidden in encoded inputs.
Malware	Identifies malware or malicious scripts.
Virus	Detects virus-related payloads or malicious code generation.
Training Data Leaks	Discovers attempts to expose or infer sensitive training data.
Resilience Checks	Evaluates model resilience to unexpected or illogical prompt behavior.
Hallucination	Examines a model's tendency to produce incorrect or nonsensical information.
Package Hallucination	Detected hallucinated or non-existent package references.
Toxicity	Tests model responses for toxic language, hate speech, and offensive content.
Bias Detection	Tests models for bias related to gender, race, religion, and other attributes.
Coherence	Evaluates the logical consistency and coherence of model responses.
Robustness	Assesses how a model handles adversarial inputs or perturbations.