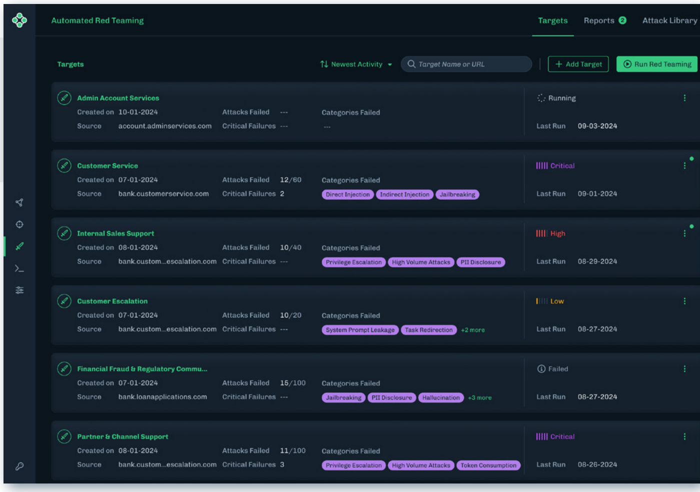
DATASHEET

HIDDENLAYER
AUTOMATED RED TEAMING

Generative AI has become a critical part of modern business, driving decision-making, automating operations, and enhancing customer experiences. But these systems also introduce new risks, from data poisoning to model tampering, that traditional security methods can't fully address.

Automated Red Teaming for AI brings the efficiency, scalability, and precision needed to identify vulnerabilities in AI systems before attackers exploit them.

Traditional red teaming is indispensable for identifying nuanced vulnerabilities and testing unique system configurations. However, it requires significant time, specialized expertise, and resources, limiting its frequency and scalability. Automated Red Teaming complements human efforts, providing continuous, repeatable testing at scale. Automated Red Teaming identifies vulnerabilities faster and adapts as systems evolve, ensuring security keeps pace with innovation.



Admin Account Services				Running
Created on 10-01-2024	Attacks Failed ---	Categories Failed		
Source account.adminservices.com	Critical Failures ---	---		Last Run 09-03-2024
Customer Service				Critical
Created on 07-01-2024	Attacks Failed 12/60	Categories Failed		
Source bank.customerservice.com	Critical Failures 2	Direct Injection Indirect Injection Jailbreaking		Last Run 09-01-2024
Internal Sales Support				High
Created on 08-01-2024	Attacks Failed 10/40	Categories Failed		
Source bank.custom...escalation.com	Critical Failures ---	Privilege Escalation High Volume Attacks PII Disclosure		Last Run 08-29-2024
Customer Escalation				Low
Created on 07-01-2024	Attacks Failed 10/20	Categories Failed		
Source bank.custom...escalation.com	Critical Failures ---	System Prompt Leakage Task Redirection	+2 more	Last Run 08-27-2024
Financial Fraud & Regulatory Commu...				Failed
Created on 07-01-2024	Attacks Failed 15/100	Categories Failed		
Source bank.loanapplications.com	Critical Failures ---	Jailbreaking PII Disclosure Hallucination	+3 more	Last Run 08-27-2024
Partner & Channel Support				Critical
Created on 08-01-2024	Attacks Failed 11/100	Categories Failed		
Source bank.custom...escalation.com	Critical Failures 3	Privilege Escalation High Volume Attacks Token Consumption		Last Run 08-26-2024

KEY PRODUCT CAPABILITIES

- **Unified Results Access** — Both the red and blue teams can access automated testing results, which provide shared visibility into vulnerabilities tied to the OWASP Top Ten framework, fostering informed and collaborative remediation efforts
- **Scalable Testing for AI Systems** — Easily scale testing as the number of AI models grows or as models increase in complexity, ensuring complete coverage across your AI infrastructure
- **Progress Tracking & Metrics** — Gain actionable insights with progress tracking and detailed metrics, allowing you to measure the effectiveness of your security posture over time
- **Prompt Injection Mitigation** — Automated tools ensure inputs to your models don't lead to unintended behaviors, protecting sensitive systems from injection-based attacks
- **Regular and Ad Hoc Scans** — Schedule scans to detect new vulnerabilities continuously or initiate ad hoc tests after significant system changes, providing real-time responsiveness to emerging threats

KEY BENEFTIS

- **Promote More Models Into Production Faster** — Accelerate model deployment via shared access to red teaming results across cross-functional teams responsible for model deployment
- **Increased Confidence In Model Resiliency** — More frequent testing identifies vulnerabilities earlier, reducing exploitation risks
- **Faster Time to Detection** — Automated scans deliver rapid insights, shortening the vulnerability remediation cycle
- **Comprehensive Scalability** — Easily adapt to expanding AI systems and evolving threats without additional overhead
- **Cost and Time Efficiency** — Save on labor costs and reduce the time to detect vulnerabilities by automating repetitive security tasks, allowing teams to focus on more sophisticated high-value analysis

Why HiddenLayer?

HiddenLayer, a Gartner recognized Cool Vendor for AI Security, creates security solutions that prevent the latest wave of cybersecurity threats against artificial intelligence assets. Using a patented approach, only HiddenLayer offers turnkey AI security without requiring increased model complexity, access to sensitive training data, or visibility into the AI assets.

Gartner
COOL VENDOR
2024

hiddenlayer.com