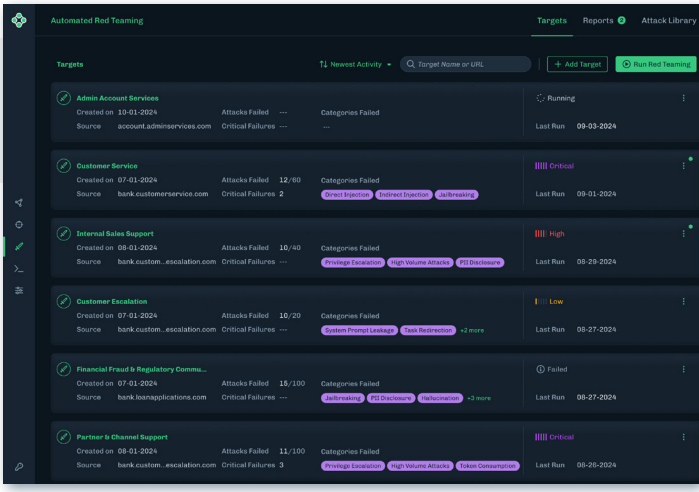
DATASHEET

# 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.

## 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 BENEFITS

- **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

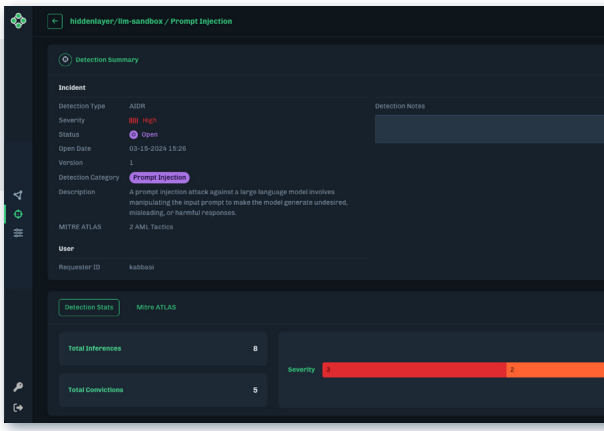
hiddenlayer.com

DATASHEET

# HIDDENLAYER
## AI DETECTION & RESPONSE

GenAI & traditional models generate immense value & aid companies in creating a competitive advantage within their market. Unfortunately, LLMs are open threat vectors to the same organizations. AI models are being attacked by ransomware, prompt injections & data exfiltration to name just a few relevant threats.

**HiddenLayer AI Detection and Response (AIDR) is the first of its kind cybersecurity solution that monitors, detects, & responds to Adversarial Artificial Intelligence attacks targeted at GenAI & traditional ML models.**




HiddenLayer's technology is non-invasive & does not inject additional data or performance overhead into your AI Models. By only observing the vectorized inputs of AI models, HiddenLayer does not need access to AI data or features, preserving the privacy & security of your company's intellectual property

Safeguard against prompt injection, PII leakage, inference attacks, evasion, and model theft while providing real-time cyber protection for AI models.

## KEY PRODUCT CAPABILITIES

- **Prompt Injection** — Ensure models can't be manipulated causing unintended consequences
- **PII Leakage** — Protect against confidential data being revealed
- **MITRE ATLAS & OWASP LLM Integration** — MITRE ATLAS & OWASP LLM integration maps to 64+ Adversarial AI attack tactics & techniques
- Protects against **Model Tampering** — know where the model is weak & tamper with the input of the model (change the sample)
- Protects against **Data Poisoning/Model Injection** — Changing the model by deliberately curating its inputs or feedback
- Protects against **Model Extraction/Theft** — stopping reconnaissance attempts through inference attacks which could result in your model intellectual property being stolen
- Uses a combination of **Supervised Learning, Unsupervised Learning, Dynamic/Behavioral Analysis & Static Analysis** to deliver detection for a library of adversarial machine AI attacks

## KEY BENEFITS

- Empower your organization to safely and securely embrace the transformative capabilities of GenAI
- Ensure security & integrity of ML Operations Pipeline
- Visibility into the risks & attacks that threaten your LLMs
- Insight into where an attack on your ML Ops & Models would most likely occur
- Detect Adversarial Artificial Intelligence attacks mapped to MITRE ATLAS tactics & techniques
- Increase return on AI projects & convert more models into production

## AVAILABLE ON

Azure Marketplace

AWS Marketplace

Google Cloud Platform

## 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.

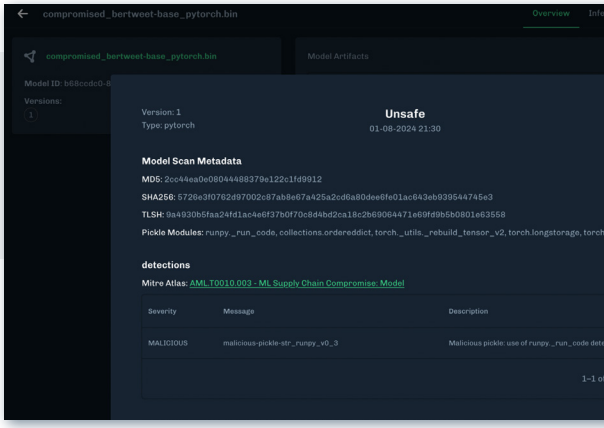
Gartner COOL VENDOR 2024

hiddenlayer.com

DATASHEET

# HIDDENLAYER
## MODEL SCANNER

Third-party and open-source model-sharing repositories have been born out of inherent data science complexity, practitioner shortage & the limitless potential & value they provide to organizations – dramatically reducing the time & effort required for AI adoption. However, such repositories often lack comprehensive security controls, which ultimately passes the risk on to the end user – & attackers are counting on it. The scarcity of security around AI models, coupled with the increasingly sensitive data that AI models are exposed to, means that model hijacking attacks evade traditional security solutions & have a high propensity for damage.




**HiddenLayer Model Scanner analyzes models to identify hidden cybersecurity risks & threats such as malware, vulnerabilities & integrity issues. Its advanced scanning engine is built to analyze your artificial intelligence models, meticulously inspecting each layer & components to detect possible signs of malicious activity, including malware, tampering & backdoors.**

HiddenLayer Model Scanner is easy to use by simply uploading your model to the Web-based Product Interface or HiddenLayer APIs will automatically analyze it for any security risks. It provides detailed reports on the findings, including recommendations on how to fix any issues & improve the model's security posture.

With HiddenLayer Model Scanner, you can ensure the integrity & safety of your artificial intelligence models, protecting them from any potential cyber threats. Whether you're a data scientist, artificial intelligence engineer, or a business leader, Model Scanner is the essential tool for securing your artificial intelligence assets.

## KEY PRODUCT CAPABILITIES

- **Malware Analysis** — Scans AI Models for embedded malicious code that could serve as an infection vector & launchpad for malware
- **Vulnerability Assessment** — Scans for known CVEs & zero-day vulnerabilities targeting AI Models
- **Model Integrity** — Analysis of AI Model's layers, components & tensors to detect tampering or corruption.
- Uses a combination of **supervised learning, unsupervised learning, dynamic/behavioral analysis & static analysis** to deliver detection for a library of adversarial machine learning attacks
- Catalog a **Known-Good State** of your AI Models as a baseline for identifying future tampering
- **Supports a variety of AI Model file types:**
  - Cloudpickle
  - Dill
  - GGUF
  - HDF5
  - Joblib
  - Keras
  - NeMo
  - Numpy
  - ONNX
  - Pickle
  - PyTorch
  - R
  - Safetensor
  - Skops
  - TensorFlow
  - Zip

## KEY BENEFITS

- Ensure third-party & open source AI models hosted by online communities & repositories are safe & secure to use by scanning the URL
- Prevent inheritance of cybersecurity vulnerabilities, malware & corruption via transfer learning of open-source AI Models
- Ensure AI Models are free of vulnerabilities & malware before deploying to production
- Improve the security & integrity of proprietary models & protect your company's intellectual property
- Prevent AI Models from being a launch pad for malware

## AVAILABLE ON

Azure Marketplace

AWS Marketplace

Google Cloud Platform

## 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