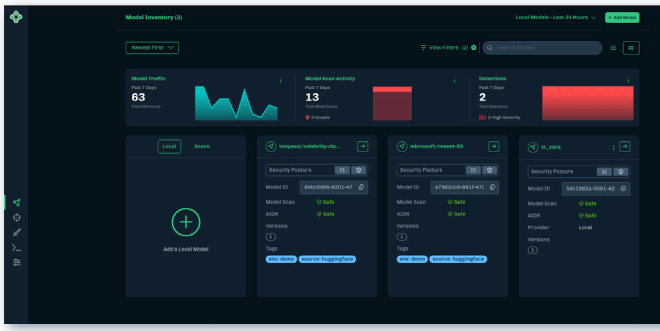
HIDDENLAYER AISEC PLATFORM

HiddenLayer's AISec Platform is a Protection Suite designed to secure AI models across the MLOps pipeline.

It proactively detects and mitigates risks from adversarial AI attacks, prompt injection, IP theft, PII leakage, and supply chain vulnerabilities — ensuring the integrity of your AI ecosystem without accessing private data or models.



AI Detection & Response

Automate and scale the protection of AI models, ensuring their security in real-time. With AIDR integrated into your environment, you can proactively defend against threats to AI unobtrusively.

Model Scanner

Scan AI models to identify hidden cybersecurity risks and threats such as malware, vulnerabilities, and integrity issues. Secure your entire AI lifecycle by protecting training, build, and production files.

Automated Red Teaming

Simulate expert attacks with zero lead time, delivering comprehensive reports to identify, prioritize, remediate, and document security risks—ensuring AI projects stay on track and compliant with security standards.

KEY PRODUCT CAPABILITIES

- **Model Genealogy & Integrity Protection** — Tracks model lineage from training through fine-tuning, identifying unauthorized changes, tampering, or corruption to ensure traceability and compliance
- **AI Bill of Materials (AIBOM)** — Automatically generates a detailed inventory of model components, datasets, and dependencies. Exportable in standard formats to support supply chain audits and licensing enforcement
- **Enhanced Threat Intelligence & Community Insights** — Combines data from public sources like Hugging Face, with expert analysis to surface actionable intelligence on emerging AI threats
- **Adversarial & Prompt Injection Defense** — Detects and mitigates adversarial attacks, model theft, and prompt injection using a blend of behavioral analysis, static inspection, and anomaly detection
- **Telemetry Dashboards & Red Teaming** — Offers advanced dashboards and simulated attack playbooks to visualize misuse patterns, agentic behaviors, and runtime anomalies
- **Security Framework Alignment** — Integrates with MITRE ATLAS and OWASP LLM, mapping over 64 tactics for unified governance and team collaboration

KEY BENEFITS

- **Supply Chain Transparency** — AIBOM ensures full visibility into model components, enabling licensing checks and regulatory compliance
- **Real-Time Threat Detection** — Continuously scans for malware, model tampering, and CVEs across the AI pipeline, ensuring rapid response to threats
- **Protection Against IP Theft & Misuse** — Stops model extraction and unauthorized agentic behavior, safeguarding IP and backend access
- **Operational Visibility & Control** — Telemetry dashboards and white-glove policy enforcement improve incident response and governance
- **Faster, Safer AI Deployment** — Cuts deployment timelines while embedding security unobtrusively—so you can ship AI products confidently

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.

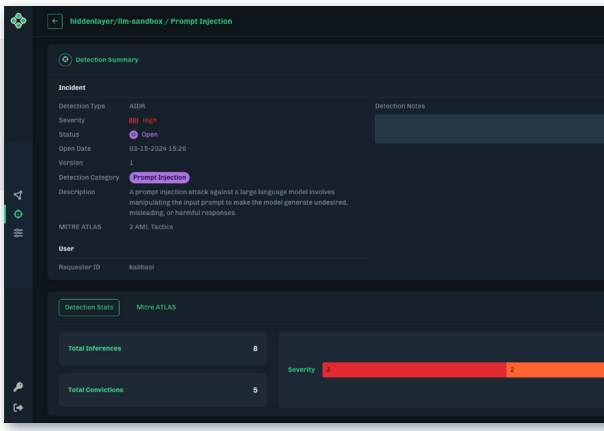
Gartner COOL VENDOR 2024

hiddenlayer.com

DATASHEET

HIDDENLAYER
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

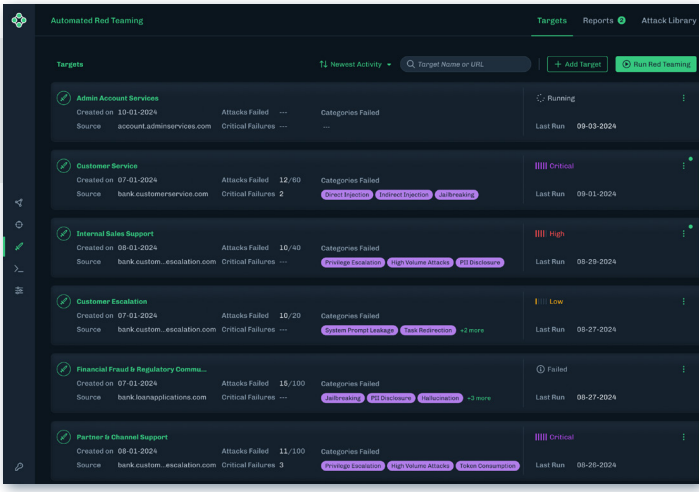
hiddenlayer.com

DATASHEET

HIDDENLAYER
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.

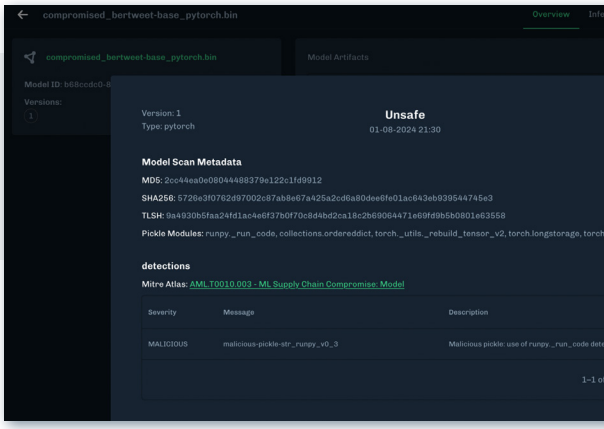
Gartner COOL VENDOR 2024

hiddenlayer.com

DATASHEET

HIDDENLAYER
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.

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
- **Model Genealogy** — Identifies model lineage, architecture & task alignment to uncover inherited risks
- **AI Bill of Materials** — Generates a standards-based inventory of model components & dependencies for compliance
- **Adversarial Detection** — Uses multiple AI techniques to detect adversarial attacks targeting machine learning models
- **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 or reused model components
- Gain visibility into a model's lineage, architecture & intended task to validate trustworthiness, support explainability & meet regulatory requirements
- Improve the security & integrity of proprietary models, protect your company's intellectual property & ensure models are production-ready
- Maintain compliance & audit readiness by generating a standards-based inventory of components, datasets & third-party dependencies

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