DATASHEET

HIDDENLAYER

AI DETECTION & RESPONSE

HIDDENLAYER

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

WHY HIDDENLAYER?

hiddenlayer.com

HiddenLayer, a Gartner recognized AI Application Security vendor, 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.

The HiddenLayer Team consists of the world's top experts at the intersection of cybersecurity & artificial intelligence. Our collective expertise derives from previous roles at McAfee, Intel, Hewlett Packard, Dell & Cylance. Over the past decade, this team has helped usher in a new era of AI-powered cybersecurity products.

DATASHEET

HIDDENLAYER MODEL SCANNER

HIDDENLAYER

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 Artificial Intelligence 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 **static detection, dynamic analysis & artificial intelligence techniques** to identify malware, vulnerabilities, model integrity & corruption issues
- Catalog a **Known-Good State** of your AI Models as a baseline for identifying future tampering
- **Supports a variety of AI Model file types:**
  - Pickle
  - Dill
  - Joblib
  - Numpy
  - Zip
  - ONNX
  - HDF5
  - PyTorch
  - TensorFlow
  - Keras
  - Scikit-Learn
  - Safetensor
  - Cloudpickle

KEY BENEFITS

- Ensure third-party & open source AI models hosted by online communities & repositories are safe & secure to use
- 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

WHY HIDDENLAYER?

hiddenlayer.com

HiddenLayer, a Gartner recognized AI Application Security vendor, 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.

The HiddenLayer Team consists of the world's top experts at the intersection of cybersecurity & artificial intelligence. Our collective expertise derives from previous roles at McAfee, Intel, Hewlett Packard, Dell & Cylance. Over the past decade, this team has helped usher in a new era of AI-powered cybersecurity products.