TROJ.AI

TrojAI Defend

SOLUTION BRIEF

AI firewall for run-time security

The rapid adoption of GenAI has increased risk and expanded attack surfaces. Enterprises face new and evolving threats to their AI applications and models in run time. Unfortunately, traditional security measures can not defend against these attacks. Organizations are investing a lot of time and resources into AI technologies. Not securing these systems creates a significant liability.

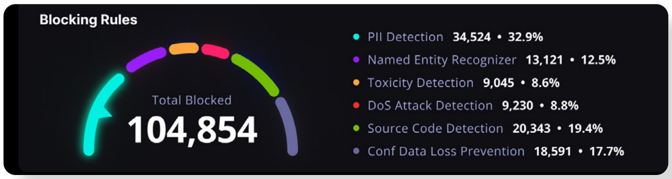
Monitoring and securing the underlying AI application and models – including the behavior of the model – is critical.

TrojAI Defend is a purpose-built AI firewall that protects against threats in run time so you can innovate without fear.

Safeguard against latest GenAI threats

Enterprises are using AI to transform critical business functions. At the same time, attacks on AI models and applications are increasing. The need for AI security has never been more urgent.

TrojAI Defend delivers real-time, multi-modal security analytics to block threats like prompt injection, data poisoning, denial of service attacks, jailbreaking, and more. TrojAI Defend leverages our extensible rules engine, adding additional detections as soon as new threats evolve.


Blocking Rules

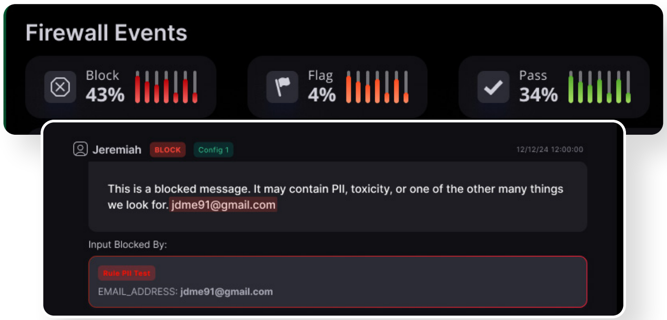
Total Blocked 104,854

- PII Detection 34,524 • 32.9%
- Named Entity Recognizer 13,121 • 12.5%
- Toxicity Detection 9,045 • 8.6%
- DoS Attack Detection 9,230 • 8.8%
- Source Code Detection 20,343 • 19.4%
- Conf Data Loss Prevention 18,591 • 17.7%

Prevent sensitive data loss

From developers to sales to HR, everyone across the enterprise is using GenAI to get their work done. The challenge now is how to prevent sensitive data from being exposed.

TrojAI Defend automatically stops data leaks and data theft at scale. It identifies and masks sensitive data within both prompts and outputs, preventing accidental exposure and unauthorized access to PII, confidential data, IP, source code, and much more.



This is a blocked message. It may contain PII, toxicity, or one of the other many things we look for. jdme91@gmail.com

Input Blocked By:

Rule PII Test

EMAIL_ADDRESS: jdme91@gmail.com

Scale AI applications securely

TrojAI Defend was built for the enterprise. It scales to filter more than 100 million tokens per second. TrojAI Defend supports the most complex use cases through a n open and extensible rules engine with deterministic and stochastic rules backed by data science and finely tuned ML and AI models.

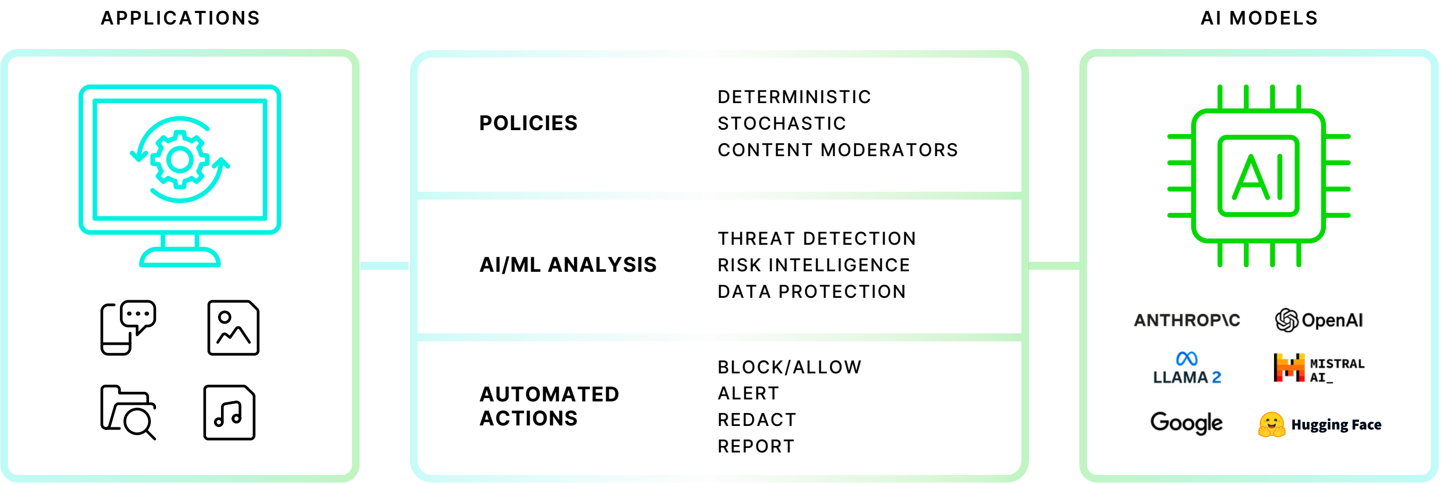
TrojAI Defend is also deployed fully on-prem so that all data and analytics performed by the TrojAI Defend AI engine remain local. This protects data privacy and maintains the overall security of the system.

Meet compliance requirements

TrojAI Defend helps enterprises meet compliance standards by enhancing security and ensuring data protection. TrojAI Defend audits third-party data storage for compliance and governance. Alerts map to both the OWASP LLM and MITRE ATLAS frameworks so you can easily provide evidence that proper security measures are in place.

TrojAI Defend

FIREWALL



APPLICATIONS

POLICIES: DETERMINISTIC, STOCHASTIC, CONTENT MODERATORS

AI/ML ANALYSIS: THREAT DETECTION, RISK INTELLIGENCE, DATA PROTECTION

AUTOMATED ACTIONS: BLOCK/ALLOW, ALERT, REDACT, REPORT

AI MODELS

TrojAI Defend key features:

- **Adversarial attack detection** - analyzes inputs and outputs of GenAI models and mitigates potential threats including prompt injection, jailbreaks, and denial-of-service attacks.
- **Continuous threat monitoring** - protects against new and evolving threats, data breaches, and PII and IP leaks.
- **AI-powered rules engine** - uses deterministic and stochastic rules that leverage AI models and also enables custom rule creation using regex, blocklists, and LLMs.
- **Enterprise-grade platform** - allows for easy integration and flexible deployment with a reverse proxy architecture; can be self-hosted or run as a cloud service.

Adaptive security for GenAI

TrojAI delivers a comprehensive AI security platform that protects AI applications and infrastructure. The best-in-class platform empowers enterprises to safeguard AI applications and models both in build time and run time. TrojAI Defend is a firewall that protects enterprises from real-time threats. TrojAI Detect pen tests AI models, safeguarding model behavior and delivering remediation guidance prior to deployment to further mitigate risk. Built by data scientists and cybersecurity experts, TrojAI secures the largest enterprises with a highly scalable, performant, and extensible solution.

Learn more at Troj.ai