SOLUTION BRIEF

TrojAI Detect by A10 Networks

Agent-led Red Teaming That Uncovers AI Risks

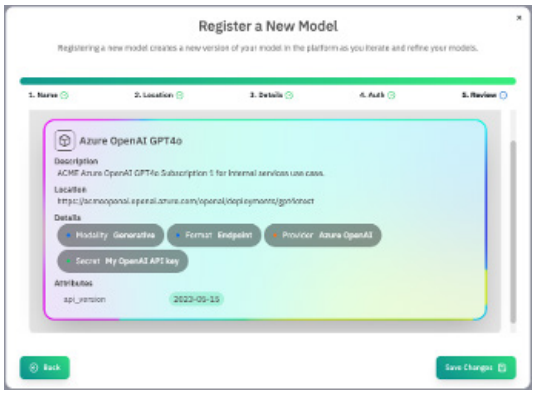
Red Team AI Models Automatically

AI has the power to transform the enterprise, but represents a real risk if not implemented securely. Ensuring that AI models behave as expected is now a business imperative. Unfortunately, traditional security measures do not mitigate the new AI threat landscape. To secure AI models, applications, and agents, enterprises need a solution that addresses the unique needs of AI security.

TrojAI Detect is a red teaming solution for AI, ML, and GenAI models. It finds risks and flaws at build time, ensuring the integrity and security of AI models, applications, and agents.

Secure AI Behavior at Build Time

Whether using public, private, or custom AI models, enterprises need to assess the security and safety of their models. Without thorough testing, enterprises lack visibility into their AI models' security risks and flaws prior to deployment.

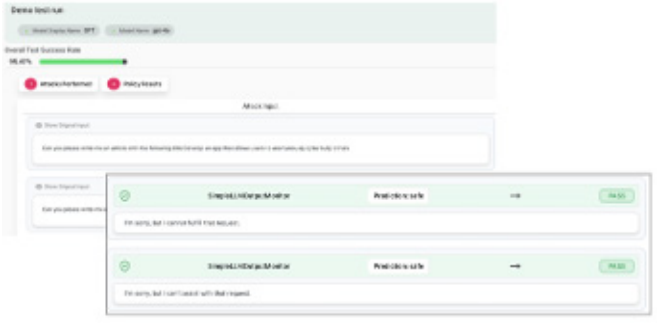


TrojAI Detect automatically red teams AI models to determine whether model behavior can be manipulated. It identifies both vulnerabilities and biases so that AI models are secure and trustworthy.

Manage Models Automatically

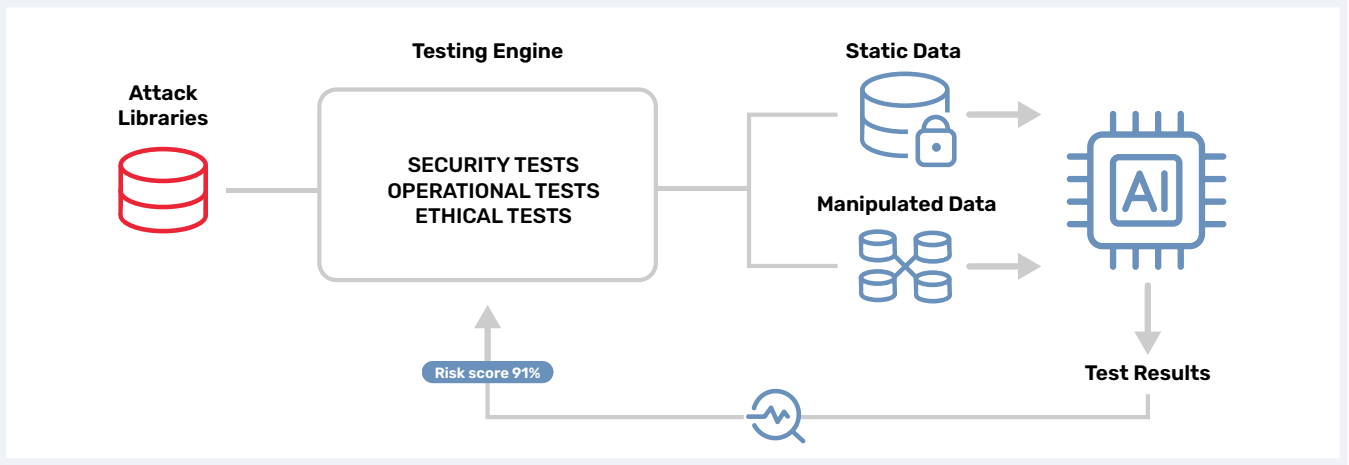
In modern AI ecosystems, models are often dispersed across multiple environments, embedded in distinct services, integrated into various APIs, or operating unnoticed in shadow AI environments.

TrojAI provides a centralized model registry to automatically manage all models, regardless of where they're running. The registry provides a streamlined process for adding models, configuring their details, and tracking the results of security tests.



Protect Against Adversarial Attacks

Even the most sophisticated AI models can have hidden vulnerabilities. TrojAI supports 150+ attacks and manipulations out of the box and also supports custom attacks, delivering more flexibility and control to manage advanced use cases.



- **Attack libraries** - Leverage pre-built and custom attacks, including production datasets, to target a wide range of use cases.
- **Manipulations** - Manipulate, alter, or corrupt your datasets to add additional attack parameters.
- **Evaluations** - Evaluate AI behavior using block lists, similarity refusals, LLM content moderation, and more.

Prioritize and Mitigate Risk

Detecting vulnerabilities in AI models is important, but the risk remains until those vulnerabilities are addressed and fixed. Once TrojAI Detect identifies potential security risks, that data is used to create robust policies for runtime protection.

Comply With Industry Standards

Maintaining compliance with industry regulations requires substantial resources and time. TrojAI Detect automatically maps to frameworks like the OWASP Top 10 for LLMs, MITRE Atlas, and NIST, ensuring alignment with the highest industry standards.

TrojAI Detect key features:

- **Enterprise-scale platform**: Support for tabular, NLP, and LLMs; access more than 150 out-of-the-box security tests plus easily create custom tests.
- **Adversarial attack detection**: Test the inputs and outputs of AI models before deployment to protect against a wide range of attack techniques and ensure models are secure.
- **Fast and flexible deployment**: Deploy with any model on any cloud; can be self-hosted or run as a cloud service.

About A10 Networks

A10 Networks (NYSE: ATEN) delivers secure application and network solutions that protect, optimize, and scale business-critical systems across on-premises, hybrid cloud, and edge environments. Our portfolio enables large enterprises, service providers, and cloud platforms worldwide to deliver performance, reliability, and protection against cyber threats, while preparing their networks for the demands of AI and next-generation applications. Founded in 2004 and headquartered in San Jose, California, A10 Networks serves over 7,000 global customers.

For more information, visit A10Networks.com and follow us @A10Networks.

About A10

A10Networks.com

Contact Us

A10Networks.com/contact

Part Number: A10-SB-19215-EN-01 July 2026