TrojAI Detect

SOLUTION BRIEF

Automatically penetration test AI models

Enterprises are investing resources into AI to transform every aspect of the business. Ensuring AI models behave as expected when dealing with the new AI threat landscape is now a business imperative. Unfortunately, traditional security measures do not address these new attack vectors.

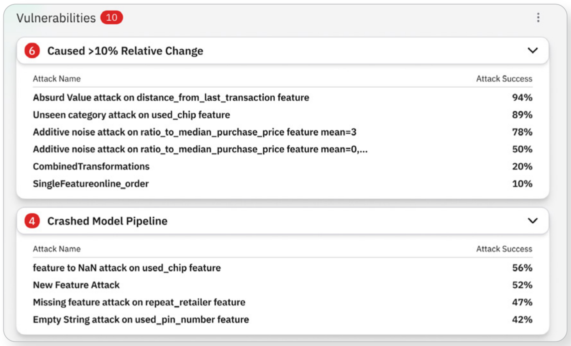
To secure AI models and applications, enterprises need a solution that addresses the unique needs of AI security.

TrojAI Detect is a penetration testing platform for AI, ML, and GenAI models. It finds risks and flaws at build time, ensuring the integrity and security of AI models and applications.

Secure AI behavior at build time

Whether using open source, commercially available, or homegrown 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 pen tests AI models to determine whether model behavior can be manipulated. It identifies both vulnerabilities and biases so that AI models are secure and operate with fairness and transparency.



Vulnerabilities 10

6 Caused >10% Relative Change

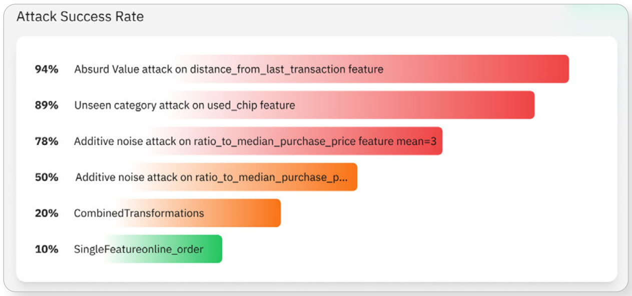
Attack Name	Attack Success
Absurd Value attack on distance_from_last_transaction feature	94%
Unseen category attack on used_chip feature	89%
Additive noise attack on ratio_to_median_purchase_price feature mean=3	78%
Additive noise attack on ratio_to_median_purchase_price feature mean=0,...	50%
CombinedTransformations	20%
SingleFeatureonline_order	10%

4 Crashed Model Pipeline

Attack Name	Attack Success
feature to NaN attack on used_chip feature	56%
New Feature Attack	52%
Missing feature attack on repeat_retailer feature	47%
Empty String attack on used_pin_number feature	42%

Protect against adversarial attacks

Hidden vulnerabilities can lurk within even the most sophisticated AI models. With more than 100 out-of-the-box security tests, TrojAI Detect delivers comprehensive reports on areas of potential risk, surfaces vulnerabilities, and ensures AI is secure.

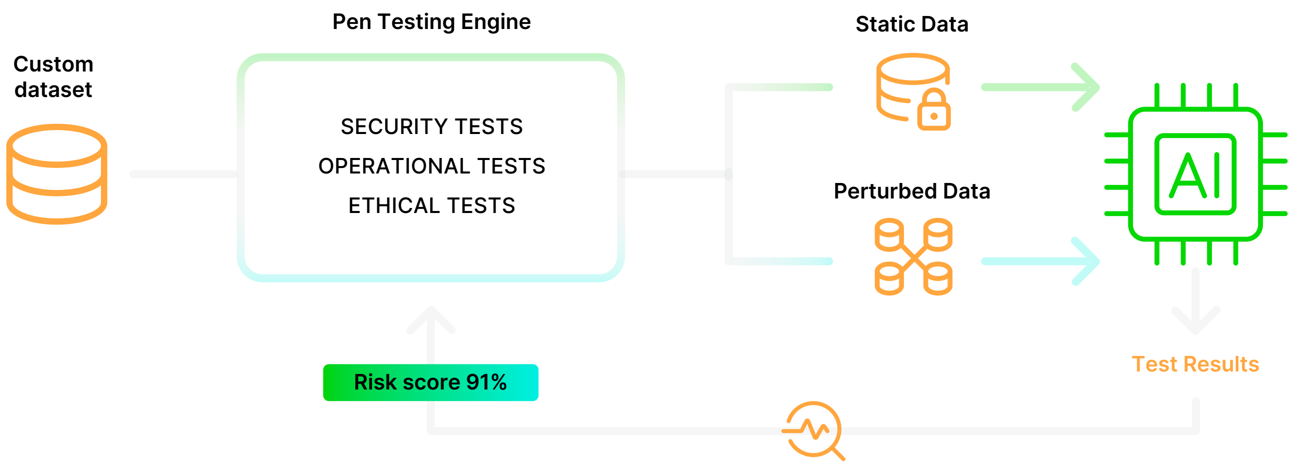


Test using advanced methodologies

With both pre-built and customizable tests tailored to specific use cases, TrojAI Detect supports a wide range of advanced testing methodologies:

- **Static** - uses established benchmark datasets
- **Perturbed** - uses manipulated inputs created by an algorithm during evaluation
- **Dynamic** - uses an LLM to attack the model, while another LLM judges the success of the attack

TrojAI Detect



Prioritize and mitigate risk

Detecting vulnerabilities in AI models is important, but the risk remains until those vulnerabilities are addressed and fixed. TrojAI Detect not only identifies potential security risks, it also helps prioritize and mitigate vulnerabilities based on severity to reduce financial and reputational risk.

Comply with industry standards

Maintaining compliance to industry regulations requires extensive time and resources. 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 100 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 to ensure models are secure.
- **Advanced reporting** - identify vulnerabilities and effective attacks with clear and concise reporting.
- **Fast and flexible deployment** - deploy with any model on any cloud; can be self-hosted or run as a cloud service.

Adaptive security for AI

TrojAI is 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 at build time and run time. TrojAI Detect pen tests AI models, safeguarding model behavior and delivering remediation guidance prior to deployment. TrojAI Defend is a firewall that protects enterprises from real-time threats. 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