NetSPI™

AI/ML penetration testing

AI/ML adoption is skyrocketing, and practical LLM business applications are growing at an astronomical rate. Ensuring the security and robustness of these models is paramount as adversaries take advantage of the emerging attack surface. We want to be your security partner through it all – from development and training to real-world deployment and implementation.

Our **AI/ML Penetration Testing** is rooted in Adversarial Machine Learning, and focuses on identifying, analyzing, and mitigating the risks associated with adversarial attacks on your machine learning systems. If evaluating ML security, building model resiliency, preparing for real-world deployment, or improving trustworthy AI is on your to-do list, consider partnering with NetSPI.

What to expect during an AI/ML pentest

Holistic and contextual testing across your tech stack

We won't test your models in a silo. Our team collaborates closely with you to contextualize ML with your existing technology stack (cloud, web, applications). Get comprehensive testing tailored to your use cases.

Build a robust pipeline for development and training

We'll assess the security of your existing pipeline and guide you in implementing best practices for secure feature engineering, preprocessing techniques, and model training.

Evaluate your defenses against major attacks

Through a combination of attack simulations, pentesting, and advanced evaluation techniques, measure how your defenses perform against major adversarial attacks such as evasion, poisoning, inference, availability, and extraction.

Actionable reports and recommendations

Our experts provide comprehensive reports and recommendations for remediation and improving defense mechanisms, all delivered via the NetSPI platform in real time. Make informed decisions to enhance your AI/ML security posture.

SCHEDULE A MEETING WITH THE AI/ML PENTESTING TEAM

A holistic approach to securing ML models and LLM implementations

As the practical use cases of machine learning continue to diversify, real threats to these deployments emerge. Our AI/ML Penetration Testing solutions cater to a diverse range of industries and deployments — from chatbots to data analytics to text generation and everything in between. **Below are key components included in our AI/ML pentests:**

- **Model architecture and modalities:** Understand the core model design, inputs, outputs, and hyperparameters. This is crucial for identifying potential vulnerabilities. Different architectures create diverse model behaviors and affect their susceptibility to attacks.
- **Dataset security:** Review the dataset used for training the model. Ensuring data privacy, data integrity, and protecting against data poisoning attacks is vital to maintain the security of the model.
- **Model updates and patch management:** Monitor for model updates and promptly address security vulnerabilities with proper patch management.
- **Integrations:** Ensure model outputs are reliable and secure in context of their integrations. Evaluate the entire data pipeline to guarantee that security measures are not assessed in isolation.
- **Model output analysis:** Carefully analyze the model's output and predictions. This can help detect potential biases and fairness issues. It's important to ensure that the model doesn't exhibit harmful behavior or make unfair decisions.
- **Adversarial testing:** Conduct rigorous adversarial testing to identify vulnerabilities and potential attack vectors. Testing the model against various adversarial examples, perturbations, and evasion techniques can reveal weaknesses and areas for improvement.
- **Input validation and sanitization:** Verify that the model adequately validates and sanitizes its inputs. This is critical for preventing injection attacks and avoiding exploitable vulnerabilities.
- **API security:** If the model is deployed via APIs, verify that the APIs are adequately secured with authentication, rate limiting, and input validation.
- **Transfer learning and fine-tuning:** If the model is based on transfer learning or fine-tuning, understand the original pre-trained model's vulnerabilities and how they may carry over to the new task.

You deserve The NetSPI Advantage

250+ In-house security experts

Intelligent process

Advanced technology

Your proactive security partner

NetSPI is the proactive security solution used to discover, prioritize, and remediate security vulnerabilities of the highest importance. NetSPI helps its customers protect what matters most by leveraging dedicated security experts and advanced technology, including Penetration Testing as a Service (PTaaS), Attack Surface Management (ASM), and Breach and Attack Simulation (BAS).