NetSPI™

AI/ML

NetSPI AI/ML Penetration Testing

Let NetSPI help you reduce the risk of using AI in your environment. Whether you are fine tuning off-the-shelf models, building your own, leveraging large language learning model functionality in your applications, or in other processes, our security experts can help you assess and enhance the resilience of AI in your environment. Our AI/ML penetration testing solutions cater to a wide range of use cases, models, and industries. We offer LLM web application testing, as well as LLM benchmarking and jailbreaking testing. We also provide custom AI testing, an advanced evaluation process that entails a comprehensive review. This includes, but is not limited to, an analysis of data collection, the structure of training data, and the validation of the AI model.

Securely incorporate LLM capabilities into your web-facing applications

Companies are eager to add AI/LLM capabilities to their web-facing applications. However, not every company understands the associated risks, and securely testing LLMs must be integrated into the software development life cycle process. Without proper configuration, LLM-capabilities like chatbots can be abused to take malicious actions or leak private data.

NetSPI can identify vulnerabilities specific to LLM capabilities that are not included in traditional static and dynamic web application testing. NetSPI can test any LLM, including GPT, Llama, Mistral, Titan and more, in any LLM framework, including Azure OpenAI, Bedrock, and others. Our advanced NetSPI Platform scanning technology, combined with the intelligence of our security experts, allows for testing deep vulnerabilities that scanning along cannot uncover, especially the potential of adversarial attacks.

Gain detailed benchmarking and analysis of potential jailbreak consequences of your LLM

Security and privacy have become significant concerns as more applications and SaaS providers adopt LLM capabilities. Without proper evaluation, these new features may allow users to manipulate LLMs, such as chatbots and expose sensitive data, generate unauthorized content, or take actions on their behalf. Application teams regularly update code, and models change over time, so it is critical to evaluate LLM-enabled application capabilities with monthly testing, which includes reporting on security metrics and trend data.

NetSPI combines advanced technology with the intelligence of our security experts to help identify potential data leakage, adversarial attacks, and content moderation, and expand beyond traditional security, including use cases such as bias and data drift. We can help you anonymize PII, redact secrets, and counteract threats such as direct prompt injections and jailbreaks. Assess and enhance your resilience against real-world threats that may seek to abuse LLM-enabled capabilities for malicious purposes. Our security experts, combined with NetSPI's Platform technology, enable you to evaluate these risks, gain recommendations for mitigating controls, and track improvements over time as you implement controls.

Customize a deep advanced model evaluation and review of your LLM

Applications with custom model coverage beyond standard third-party, LLM-enabled web applications need custom analysis. NetSPI leverages our security experts, along with our NetSPI Platform to perform a tailored service. This includes a deep review of the data collection, training data structure and cleaning, training data validation, model algorithms and hyper-parameters, as well as advanced model extraction, member attribution, inference, inversion, and evasion attacks.

Training data collection, cleaning, selection and hyper parameter configurations can have a dramatic impact on security, usability, bias, and fairness, but are often overlooked as part of the evaluation process. NetSPI security experts conduct interviews and review the current pipeline and configurations to produce a threat model that can highlight core areas of weakness and recommend mitigating controls that can improve the overall security posture of the target model.

Our comprehensive AI/ML penetration testing service offerings:

SERVICE FEATURES	LLM WEB APPLICATION TESTING SERVICE	LLM BENCHMARKING AND JAILBREAKING SERVICE*	CUSTOM AI TESTING SERVICE
Continuous	✓	✓	●
Traditional	✓	●	✓
Jailbreak Benchmark	✓	✓	✓
LLM Capability Code Extension	●	●	✓
LLM Model: Model Theft (Extraction)	●	●	✓
LLM Model: Member Attribution (Inference)	✓	●	✓
LLM Model: Data Theft (Inversion)	●	●	✓
LLM Model: Evasion Enumeration	✓	●	✓
LLM Model: Evaluate System Prompts	●	●	✓
Evaluate Training Data Collection	●	●	●
Evaluate Training Data Structure	●	●	●
Evaluate Training Data Cleaning	●	●	●
Evaluate Training Data Validation	●	●	●
Evaluate Model Algorithms & Configurations	●	●	●

**If changes to the target application's authentication, request structure, response structure, or workflow require modifications to the test harness developed during onboarding, additional fees may apply.*

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), External Attack Surface Management (EASM), Cyber Asset Attack Surface Management (CAASM), and Breach and Attack Simulation (BAS).