Financial Strides

LLM Penetration Testing Services

With the rapid advancement of artificial intelligence, large language models are now central to many applications, but they also present unique security concerns. To keep your LLMs safe and your AI solutions secure, you need expert security services from a reliable partner

How We Can Help You Strengthen Your AI Security

We understand the complexities of securing large language models (LLMs), and our tailored penetration testing methodology is designed specifically for these advanced systems, here's how we can help you:

- Identify and prioritize the most significant risks associated with the use of GenAI and LLMs.
- Analyze potential attack vectors, data privacy concerns, and any vulnerabilities inherent to the AI/ML models used.
- Utilize experienced certified offensive security testers to simulate adversarial attacks on the AI systems, including prompt injections, data poisoning, session hijacking, and input manipulation.
- Conduct a thorough examination of the web application/service used by end users for interaction with the backend LLM, applying standard web application pen testing methodologies to ensure comprehensive security analysis.
- Design specific technical controls to safeguard against identified risks, which may include enhancing API security, strengthening authentication mechanisms, and ensuring data encryption.
- Provide a detailed report, including vulnerabilities discovered, tests conducted, and recommendations for improvement. We can help you evaluate the resilience of your LLMs and ensure the integrity of your AI-powered systems.

Why Trust us for AI Security

Our pen test methodology is developed using globally recognized security frameworks, such as MITRE's ATLAS and OWASP, ensuring our testing processes align with the highest standards in AI security.

MITRE ATLAS

OWASP LLM Top 10

NIST AI Risk Management Framework

Financial Strides

Problems We Help You Solve with AI Security and LLM Penetration Testing

Vulnerable Code and APIs

AI applications often involve complex codebases and APIs. Any vulnerabilities in this code, such as improper input validation, insecure data storage, or weak authentication mechanisms, can be exploited by attackers to gain unauthorized access or manipulate AI-driven functionalities. Penetration testing identifies these vulnerabilities to prevent such exploits.

Prompt Injections

Our experts test how the AI model responds to crafted input prompts that could lead to unintended or harmful outputs. Penetration testers might use complex or misleading prompts to see if the model can be tricked into revealing sensitive information or making incorrect decisions.

Insecure Plugin Design

Testers examine plugins or extensions used with LLMs for security vulnerabilities. Such flaws can enable malicious inputs to have harmful consequences ranging from data exfiltration, remote code execution, and privilege escalation.

Insecure Output Handling

Testers evaluate how LLMs process and display output data. Successful exploitation of an Insecure Output Handling vulnerability can result in XSS and CSRF in web browsers as well as SSRF, privilege escalation, or remote code execution on backend systems. Pen testing aims to ensure outputs are properly sanitized to prevent accidental data exposure.

Data Privacy Breaches

GenAI and LLMs handle sensitive data, which can be at risk of unauthorized access or leakage. Penetration testing ensures robust data protection measures are in place.

Reputational Risks

Inaccurate or inappropriate outputs from LLMs can lead to reputational damage, especially if the AI's responses are not aligned with the official stance of the company. Penetration testing can help prevent these scenarios by identifying vulnerabilities that might allow the AI to generate such responses.

Compliance Risks

Non-compliance with regulatory standards can lead to legal issues. Penetration testing helps maintain compliance, particularly in sectors with stringent data security regulations.

We always provide free retesting post-remediation by the client during a period of 30 days after the initial test report is produced.

Contact: info@finstrides.com
See examples of pen test reports at https://www.finstrides.com/security
Please see here some of our clients we have helped with security and compliance assignments.