Executive Summary

As artificial intelligence (AI) systems, particularly Large Language Models (LLMs), become integral to business operations, they also introduce unique security challenges. The Open Web Application Security Project (OWASP) has identified the top 10 vulnerabilities specific to AI/LLMs, which can lead to data breaches, misuse, and reputational damage. As generative AI systems evolve, so do the security challenges they present. Organizations integrating AI solutions face new vulnerabilities that demand advanced testing and protection mechanisms.

Our penetration testing leverages advanced artificial intelligence to simulate cyberattacks, uncovering hidden vulnerabilities across systems, networks, and applications. This cutting-edge approach boosts security by automating threat detection, streamlining processes, and delivering sharper, more comprehensive insights into potential risks lurking within digital environments. RakFort offers cutting-edge AI penetration testing (pen testing) services designed to proactively detect and mitigate vulnerabilities specific to generative AI models and data pipelines. Our services are rooted in advanced red-teaming methodologies tailored for AI systems, ensuring the security and robustness of AI deployments.

RakFort specializes in penetration testing services tailored to AI systems, ensuring your organization is protected against these emerging threats. This document outlines our services, the top 10 OWASP LLM vulnerabilities, and how we help you mitigate them.

RakFort employs multiple persons to perform pen testing. Their names are Shiva, Rahul, Suresh and Jordan.

Security Challenges

Large Language Models (LLMs) like GPT, Bard, and others are revolutionizing industries. However, their complexity and integration into critical systems make them prime targets for attackers. The OWASP Top 10 for LLMs highlights the most critical risks, including prompt injection, data leakage, and model poisoning.

Our penetration testing services are designed to identify, exploit, and remediate these vulnerabilities, ensuring your AI systems are secure, compliant, and resilient. Traditional cybersecurity measures often fall short when addressing these AI-specific threats.

RakFort's AI pen testing services provide:

Vulnerability Assessment

Continuous identification and prioritization of vulnerabilities across AI assets.

- **Adversarial Testing:** Simulating real-world attacks using both automated tools and manual red-teaming techniques.
- **Data Integrity Assurance:** Protecting AI systems from data poisoning and manipulation.
- **Compliance Validation:** Ensuring adherence to industry-specific regulations and security best practices.

Unique Features

We offer comprehensive penetration testing services tailored to AI systems, including:

Automated Red Teaming

Utilize the comprehensive attack library to run detailed, automated LLM attacks based on categorized threat profiles, including jailbreaks, prompt injection attacks, and input manipulations—essential for preserving the integrity and security of AI systems.

Human Augmented

Develop targeted attack objectives tailored to your LLMs, incorporating business-specific goals for sectors like finance, healthcare, customer service, and beyond. This approach ensures a sharp, focused simulation that delivers attack insights most relevant to your business needs.

Customized attack libraries

Recon comes pre-loaded with a library of over 20,000 known vulnerabilities specifically designed to target GenAI systems, enhancing their safety and security. This attack library is updated weekly with the latest techniques and tactics and also offers the ability to integrate your own threat research, ensuring your systems are fortified against known risks.

No-Code Integration, Base model agnostic

In under 5 minutes, begin scanning your custom endpoints for vulnerabilities using any base model. Automated scans run asynchronously, and your team will be notified once the scan is finished.

Security testing for Generative AI-enabled applications and APIs

Proactively safeguard your apps by incorporating pen-testing into your application development lifecycle.

- Secure both applications and networks
- Defend against common Generative AI exploits, including prompt injection attacks, jailbreaks, and insecure output handling
- Gain insights into the unique risks associated with generative AI
- Collaborate with seasoned pen-testers, including over 3 dozen RakFort Core members specializing in LLM testing

Our Methodology

RakFort employs a multi-layered red-teaming approach to AI pen testing, combining several strategies:

1. **Prompt Brute Forcing**
    - Manual testing of AI system inputs and outputs by specialized security teams.
    - Development and use of Generative AI checklists targeting common vulnerabilities, including top OWASP LLM attack vectors.
2. **Adversarial Knowledge Base**

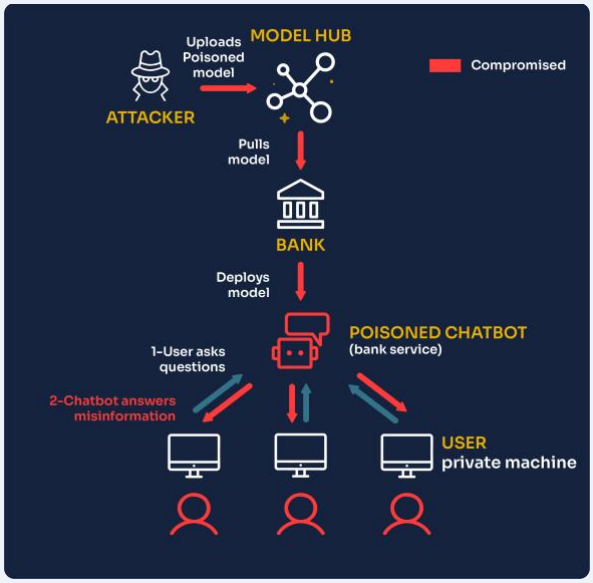
- A curated database of adversarial prompts sourced from trusted internal and external sources.
- Regular updates to reflect the evolving threat landscape, supporting black-box testing through automated scripts.

3. **AI-Assisted Adversarial Testing**
   - Leveraging AI models to generate new adversarial prompts and test cases.
   - Automated identification of vulnerabilities through sophisticated AI-driven simulations.
4. **Model Scanning**
   - **Serialization Model Attacks:** Detecting vulnerabilities introduced during the AI model serialization process.
   - **Foundational Model Scanning:** Using tools to scan foundational LLM models for inherent security weaknesses.
5. **White-Box Model Testing**
   - In-depth testing of model architecture, training data, and integrations.
   - Specialized vulnerability assessment for Retrieval-Augmented Generation (RAG) systems, ensuring contextual data integrity.

Case Studies (CS) for AI Penetration Testing Services

CS1: Preventing Prompt Injection in a Financial Chatbot

Challenge: A leading financial institution discovered that its AI-powered customer service chatbot was susceptible to prompt injection attacks. Malicious users could



manipulate the chatbot's input prompts to bypass content filters, potentially leading to unauthorized disclosure of sensitive financial information.

Solution: Our AI penetration testing team conducted a comprehensive assessment, employing both manual red-teaming techniques and automated adversarial testing tools. By simulating real-world attack scenarios through white-box and black-box testing, we identified critical vulnerabilities within the chatbot's natural language processing framework. We implemented input sanitization protocols and enhanced prompt validation mechanisms. Additionally, continuous monitoring tools were integrated to detect nd prevent future exploitation attempts in real-time.

Result: Following our intervention, the financial institution reported zero incidents of prompt injection attacks. The chatbot's interactions became significantly more secure, restoring client confidence and ensuring compliance with stringent financial regulatory standards. Regular security assessments were scheduled to maintain long-term resilience.

CS2: Securing Training Data for a Healthcare Large Language Model

Challenge: A prominent healthcare provider developed an LLM to assist with medical queries and patient documentation. However, the model faced risks of data poisoning, where malicious actors could manipulate training data, potentially leading to biased or harmful outputs. This threat jeopardized patient privacy, data integrity, and compliance with healthcare regulations such as HIPAA and GDPR.

Solution: Our team performed a thorough audit of the entire training data pipeline, combining AI-assisted testing with manual vulnerability assessments. We introduced advanced data validation mechanisms and implemented differential privacy techniques to protect sensitive patient information.

To ensure model robustness, anomaly detection systems were integrated to identify and address unusual patterns in training data inputs.

Result: The security enhancements led to a notable improvement in the model's accuracy and reliability. The LLM now operates in full compliance with relevant healthcare data protection regulations, significantly reducing the risk of future data poisoning attacks. The client observed increased trust among medical professionals using the system, along with enhanced decision-support capabilities.

RakFort's Sample Report

Below is a detailed breakdown of the top 10 OWASP LLM vulnerabilities and how our services address them. Proactively safeguard your apps by incorporating pen-testing into your application development lifecycle.

1. Summary

This penetration testing engagement was conducted to assess the security posture of *[Client Name]*'s AI systems, focusing on potential vulnerabilities within large language models (LLMs). The testing targeted the top 10 OWASP LLM threats, including prompt injection, data poisoning, and model inversion attacks.

Key Findings:

- Critical vulnerabilities identified: [Number]
- High-risk issues resolved: [Number]
- Compliance concerns: [Compliance standards affected, e.g., GDPR, HIPAA]

Recommendations:

- Implement prompt validation filters.
- Strengthen data input sanitation procedures.
- Regular vulnerability assessments and AI model monitoring

2. Methodology

Scope of Testing:

- Model Types Tested: [LLM Name, e.g., GPT-4, BERT]
- Testing Layers: White-box, Black-box, Manual, Automated
- Attack Vectors: Based on OWASP top 10 LLM attacks

Testing Techniques:

- Manual red-teaming
- AI-assisted adversarial testing
- Model scanning using tools like Garak and ModelScan
- Prompt brute forcing and adversarial prompt generation

3. Detailed Findings

Vulnerability ID	OWASP Category	Severity	Description	Potential Impact	Status
VULN-001	Prompt Injection	Critical	Allows attackers to manipulate AI outputs.	Disclosure of sensitive data	Unresolved
VULN-002	Data Poisoning	High	Insertion of malicious data into training datasets.	Compromised model integrity	Mitigated
VULN-003	Model Inversion	Medium	Sensitive data extraction from model outputs.	Privacy violations	In Progress

4. Remediation Recommendations

- **Prompt Injection:** Implement stricter prompt validation and sanitization.
- **Data Poisoning:** Regularly audit and cleanse training datasets.
- **Model Inversion:** Restrict model output verbosity and apply differential privacy techniques.

Conclusion

The integration of LLMs into your business operations offers immense potential but also introduces significant security risks. RakFort is your trusted partner in securing AI systems against the top 10 OWASP LLM vulnerabilities. Our tailored penetration testing services ensure your systems are resilient, compliant, and secure.

❖ *Contact us today to schedule a consultation and protect your AI investments.*

✉ : info@rakfort.com