AI

SECURITY INNOVATION
A BUREAU VERITAS COMPANY

GenAI LLM ASSESSMENT SERVICES

Security Innovation's GenAI LLM assessment services are tailored to address the unique threats that GenAI systems face, providing tailored remediation guidance for optimal risk mitigation in AI environments. In addition, our assessments align with industry best practices, including assisting organizations in meeting the security requirements of the ISO/IEC 42001 standard, which focuses on AI system governance and risk management. By ensuring your AI solution is secure and follows these international guidelines, we help you maintain compliance while addressing critical security concerns.

Organizations such as AWS, eClinical Works, and HP have relied on Security Innovation to ensure their AI solution is secure and follows data security best practices.

ASSESSMENT SERVICES

Our assessment services reduce AI risk across multiple angles: from the process by which Gen AI is designed and implemented, deploy level exploitation and deployment review.

- **Threat Modeling** - identify potential risks and threats in AI landscape
- **Source Code Review** – identify flaws in native code
- **Architecture & Design Review** – ensure resiliency to AI threats and attacks are built in
- **Penetration Testing** – exploit vulnerabilities in the AI ecosystem
- **IT Attack simulation** – conduct targeted attacks on the operating environment and infrastructure
- **Secure SDLC assessment** – identify dangerous practices and replace with best practice
- **Cloud Configuration Review** – identify unused resources and insecure configurations

COMMON FOCUS AREAS OF THE ASSESSMENT MAY INCLUDE:

- User Injected Bias
- Sensitive Data Exposure
- Uncontrolled Repetition
- Loss of Control Context
- Insecure Plugin Design
- Template Injection
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Prompt Disclosure
- Model Theft
- Tokenization
- History Replay
- Prompt Injection

SCALABLE METHODOLOGIES FOR OPTIMIZED RESULTS

Each engineer follows the same test methodology, which includes the OWASP LLM Top 10 and over twenty-five custom test cases based upon the multitude of contexts, modalities, and content processing that Gen AI and LLM provide.

Our team is able to inspect tokenization, guardrails, and model training through static analysis, while penetration testing enables dynamic assessment of the context sensitive user-controlled prompts through interaction with the system, and validation that integrated systems and supporting environments are utilized securely.

GenAI security assessments require crafting input tailored to the User Prompt, Guardrails, Input and Output filtering, and supporting technology stack as required. Documented test cases will highlight inputs to the system, expected outputs, and analysis of resulting outputs. Real world attacks include context specific requests such as from an employee facing HR assistant, a customer facing system that blends self-help content with escalation to human agents, or a system to generate user-designed images.

FOCUS ON MITIGATING KEY AI THREATS

Emerging threats that our engineers consider while assessing AI models or systems developed with AI include:

USER INTRODUCED BIAS

AI systems are constantly learning and evolving on user input. We look for weak spots that would allow a bad actor to introduce bias or malicious instructions to the AI so that it is later giving incorrect or malicious output to future application users.

PROMPT INJECTION

This is the most common early vulnerability in LLMs and AI systems. We ensure that an attacker can't misuse the AI to give results outside of its content restrictions, abuse computational resources, steal internal information about the system, or compromise the entire system.

SENSITIVE DATA EXPOSURE

Many AI developer tools and scanners require developers to grant permission to use their code base to train future models. We ensure code that can't later be used by other customers of the AI tool and result in a "Write me an app like my competitors" scenario.

Common mitigation measures that our engineers will provide tailored remediation guidance for your AI solution/ implementation include:

- **Data Security** – how to encrypt sensitive data at rest and transit using encryption algorithms
- **Model Security –** how to protect AI models against tampering by ensuring integrity by implementing digital signatures
- **Infrastructure Security** – how to implement various cloud solutions, e.g. Security Center or AWS shield, to protect AI infrastructure against threats
- **Identity and Access Management (IAM**) – how to properly implement Role Based Access Control (RBAC) to limit exposure to resources belonging to AI
- **Multi-Factor Authentication** - how to enforce MFA for services belonging to AI workloads to add an extra layer of security