Blueinfy

FAST-PACED ADOPTION OF GEN AI – BALANCING OPPORTUNITIES & RISKS

BACKGROUND

ACME has consistently led the way in adopting new technologies, particularly Generative AI (Gen AI) models, to enhance various business processes, including document summarization, data retrieval, customer support automation, content generation, and web search functionalities. However, the security landscape for Large Language Models (LLMs) presents unique challenges where traditional security approaches/strategies fall short. Recognizing this, ACME engaged Blueinfy to devise a tailored strategy to uncover potential vulnerabilities, such as prompt injection attacks and other contextual risks associated with Gen AI applications, along with traditional vulnerabilities.

CHALLENGE

ACME's existing security program, which includes SAST, DAST, and selected manual penetration testing, was inadequate for testing specific to LLMs. The architecture typically involves a front-end layer with a back-end API connecting to LLMs to perform various tasks. Automated scanners failed to detect even traditional attacks like Remote Code Execution (RCE) and SQL injection (SQLi) because the medium was identified through LLM prompts, which these scanners could not effectively evaluate.

SOLUTION

Blueinfy provided crucial support to ACME by implementing a comprehensive security strategy focused on the following key areas:-

AI Model Interpretation & Architecture Study:
Effective testing begins with a thorough understanding of the underlying architecture and the AI model driving the application. This involves grasping the core algorithms, input data, and expected outcomes. With this detailed knowledge, precise test scenarios were developed.

Full-Scope Penetration Testing:
Blueinfy conducted in-depth, human intelligence-driven, full-scope penetration testing of ACME's Gen AI applications. This assessment identified vulnerabilities, both traditional and specific to LLM implementations, such as prompt injection and other manipulation tactics that could compromise the AI models' integrity.

Scoring Mechanism for Risk Parameters:
To help implement guardrails and mitigate potential brand impact, Blueinfy developed a comprehensive scoring mechanism to evaluate each Gen AI application across critical parameters, including:

Fairness and Bias: Assessing the AI system for fairness across protected attributes and identifying potential biases.

Abuse and Ethics: Evaluating ethical implications, risks of misuse, and the potential for politically biased or harmful outputs.

Data Privacy: Examining the handling of personally identifiable information (PII) and ensuring data security.

Hallucination and Context: Evaluating the risk of hallucinations and out-of-context outputs that could mislead users.

Toxicity and Insults: Assessing the potential for generating insults, sexually explicit content, profanity, and severe toxicity.

Data Exfiltration: Evaluating the risk of unauthorized data extraction from AI models, ensuring that sensitive information is adequately protected.

Ongoing Risk Assessment:
Following the initial penetration testing, Blueinfy recommended an ongoing risk assessment process for identified LLM vulnerabilities. This approach allows ACME to continuously evaluate the risks associated with data and model upgrades, ensuring that security measures remain effective as the technology evolves. This also helped the ACME team to keep up with the various bypass techniques evolving continually against enhanced security measures being implemented by LLM companies.

CONCLUSION

The collaboration with Blueinfy resulted in several significant outcomes – especially uncovering vulnerabilities leading to data exfiltration, mass phishing attacks, data stealing etc. Vulnerabilities were effectively risk-rated, promptly addressed, and necessary guardrails were implemented, reducing the risks of data exfiltration and the generation of harmful or biased outputs, thereby minimizing potential brand damage. This partnership equipped ACME with the tools and strategies needed to navigate the complexities of Gen AI security, ensuring that its innovative applications remain secure against emerging threats while continuing to drive business value.

Article by
Hemil Shah & Rishita Sarabhai