Cobalt

SOLUTION BRIEF

SECURE YOUR AI & LLM-ENABLED ASSETS

COMMUNITY & CONTRIBUTION

The Cobalt Core includes over 30 security experts with robust experience in LLM testing. Several Core members are also active contributors to the OWASP Top 10 Project for LLM testing.

In recent years, our society has witnessed the breakneck adoption of artificial intelligence (AI) and machine learning (ML) based technologies to innovate and accelerate business operations. The use of AI, ML, and large language models (LLMs) can drive productivity and scale in unprecedented ways, but new technologies also introduce new frontiers for cyber risk.

As organizations grow more reliant on AI, it becomes increasingly critical to stay on top of the risks to which it's uniquely susceptible. Cobalt draws on the expertise of our Core pentester community to not only assess your applications and infrastructure for prevalent AI & LLM-based threats, but also provide guidance directly from active contributors to AI security research such as OWASP.

KEY BENEFITS

DIRECT INSIGHT & GUIDANCE FROM THE AUTHORITIES

Our resident security experts and pentesters are direct contributors to the OWASP LLM Security & Governance Checklist: the first-of-its-kind framework for LLM testing.

CONSISTENT & COMPREHENSIVE COVERAGE

Gain full visibility as your pentesting team works through our comprehensive OWASP-based methodology. Uphold best practices for securing AI & LLM-enabled assets.

SIMPLIFIED SETUP, STREAMLINED TESTING

Easily schedule testing for LLM-enabled apps and APIs directly in the Cobalt platform, without the headache or back-and-forth negotiation of custom scoping and SOWs.

SOLUTION BRIEF

COMMON TEST CASES

Our experts employ their understanding of the evolving threat landscape and attacker techniques to prioritize the most critical risks to AI & LLM implementations. Common test cases we perform for AI & LLM-enabled applications include, but are not limited to:

PROMPT INJECTION Threats stemming from sensitive data extraction, the theft of application prompts, or the unauthorized use of application functions	**WHAT WE TEST FOR** • Creating a prompt to bypass LLM controls, revealing sensitive data (e.g. user credentials, product licenses, internal system details, etc.) • Bypassing content filters using language patterns or tokens
JAILBREAK The manipulation or hijacking of LLM prompts to direct them towards malicious or unintended outputs	**WHAT WE TEST FOR** • Using semantic deception & "social engineering" to generate hostile content • Using base64 encoding to bypass LLM prompts programmed not to respond to elicit or crime-related NLP requests
INSECURE OUTPUT HANDLING Capitalizing on insufficient validation or scrutinization of LLM outputs for malicious purposes	**WHAT WE TEST FOR** • Directly ingesting LLM outputs into downstream system operations, leading to remote code execution (RCE) • Enabling execution of a SQL query from an LLM • Creating vulnerable Javascript payloads from an LLM, leading to XSS

Ready to Get Started?
Request a Demo Today!

Cobalt

San Francisco | Berlin | Boston

COBALT.IO