

Coalfire AI/ML Offensive Security Services

Innovate With Confidence: Human-Powered Security for Emerging AI Technologies

As firms develop innovative AI/ML solutions, they must identify and remediate novel security vulnerabilities before attackers exploit them. Coalfire®'s Cyber Security Services team helps enterprises stay ahead of modern threats with specialized threat modeling and penetration testing for AI/ML systems. Our experienced hackers and technologists simulate real-world attack scenarios and leverage advanced testing methodologies to provide actionable insights to reduce risk, enhance security posture, and safeguard critical assets.

Agentic AI: Expanding the Attack Surface

Emerging AI-powered virtual agents enable threat actors to exploit novel attack vectors to inflict significant harm. Companies are increasingly expanding access privileges for AI applications to retrieve data from a complex array of company databases and applications. An AI tool with agentic capabilities, i.e., one that executes code and performs automated processes on a user's behalf, poses heightened risk of damage if taken over by malicious actors through methods such as prompt injection attacks.

Securing AI Systems with Coalfire

Many firms have continuous automated penetration testing programs. However, for nascent AI systems, threat-focused, manual testing by humans is essential to replicate the novel attacks targeting agentic AI applications. Coalfire's experienced testers bring a hacker's perspective to help clients identify real-world risks and remediation actions:

- **AI/ML Application Threat Modeling:** Evaluation of security controls, system data flows, application architecture, and overall alignment of security measures within AI and ML frameworks. Coalfire provides actionable recommendations across the software development lifecycle for comprehensive risk mitigation.
- **AI Integrated Application Penetration Testing:** Expert-powered, manual penetration testing to uncover cyber risks, with a focus on how threat actors see your solution and how those risks could impact your business. Coalfire provides technical details of vulnerabilities, actionable remediation recommendations, and examples of exploitation for further analysis.

Case Study: Coalfire AI/ML Pen Testing Revealed Critical Vulnerabilities in a RAG AI Agent

A Fortune 500 firm engaged Coalfire to evaluate and test an AI agent already in production, as well as its supporting web application infrastructure. During a two-week testing period, Coalfire's testers both validated the client's initial security concerns and successfully simulated novel prompt injection attacks that enabled remote code execution on backend infrastructure. Coalfire detailed the test's findings, rated them by criticality, and provided recommendations that subsequently guided the client's engineering team to remediate the uncovered risks.

Coalfire's Suite of AI/ML Services

In addition to threat-based testing, Coalfire has an array of complementary services for AI/ML use cases that help clients establish appropriate governance, proactively detect threats, and manage remediation actions:

- **AI Risk Advisory:** Coalfire helps clients design, implement, manage, and operationalize AI Risk Management Programs in alignment with the NIST AI Risk Management Framework (RMF).
- **Threat Hunting:** Coalfire's team of threat hunters investigates and uncovers potential threats in a client's existing security stack through an intelligence-driven, threat-focused approach.
- **Managed Services:** Coalfire's Security Managed Services include coordinating remediation actions that address vulnerabilities identified during testing.

For more information visit coalfire.com/cybersecurity-services or [contact us](#).