PartnerOneIT

AI Security and Prompt Injection Testing

Next-generation security for LLMs, agentic systems and enterprise AI

AI Security and Prompt Injection

97% of companies lack AI access controls.

AI assistants, agents, and RAG systems face threats traditional security can't address—prompt injection, data leakage, and model manipulation that bypass conventional controls. Plurilock's AI Red Teaming delivers comprehensive adversarial testing across your AI ecosystem. We test for prompt injection, data leakage, and model abuse, providing clear remediation guidance and strengthened guardrails aligned with ISO/IEC 42001, MITRE ATLAS, NIST AI RMF and OWASP LLM Top 10.

Advisory and Assessment Services

AI Security Foundations

Current State Analysis

Assess current AI deployment architecture, security controls, and risk management processes across production systems.

AI System Inventory and Discovery

Identify all AI/ML systems, agents, datasets, shadow-AI deployments, and enterprise AI assistants in use.

AI Security Architecture Review

Analyze deployment architectures, API configurations, and prompt engineering to identify security integration opportunities.

AI Threat Modeling

Build threat models aligned to MITRE ATLAS, OWASP LLM Top 10, and NIST AI RMF, addressing prompt injection, model extraction, and data poisoning.

Governance and Compliance

ISO/IEC 42001 and NIST AI RMF Assessment

Evaluate AI governance against ISO/IEC 42001 and NIST AI Risk Management Framework to establish baseline and target maturity states.

AI Security Requirements Specification

Develop security requirements for AI development teams, including guardrail implementation and safety controls.

AI Governance Policy Development

Create policies for AI security governance, model validation, continuous monitoring, and incident response aligned to regulatory requirements.

AI Defense-in-Depth Planning

Provide technical recommendations for layered AI security controls across prompt filtering, API security, model isolation, and agent sandboxing.

Offensive AI Security Testing

Testing Capabilities

Real-Time Portal with Jira Integration

Centralized portal provides continuous visibility into AI security findings with seamless Jira integration for streamlined risk tracking.

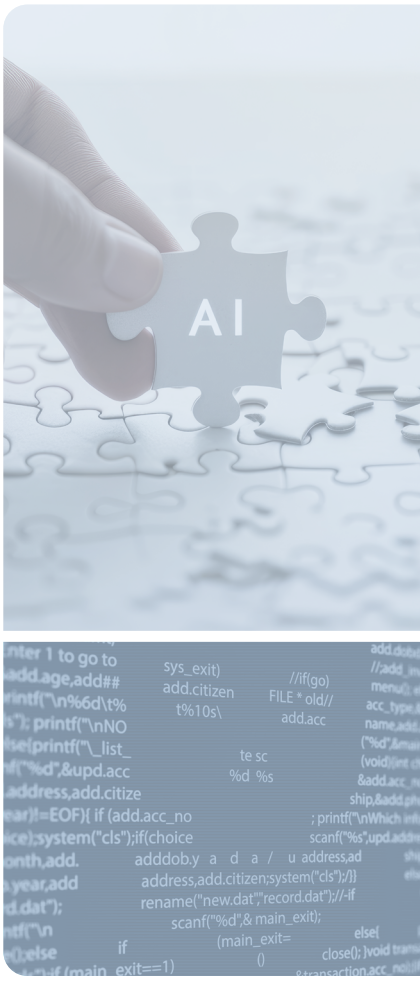
AI Red Team Testing

Comprehensive adversarial testing identifies exploitable vulnerabilities including prompt injection, jailbreak attempts, multi-turn manipulation, and unauthorized data access.

AI Security Red Teaming Services

Plurilock conducts offensive testing against:

- General-purpose LLMs and enterprise AI assistants
- RAG (Retrieval-Augmented Generation) systems
- Multi-agent and agentic AI ecosystems
- LangChain-based architectures and orchestration frameworks



Powered by Plurilock

- Model Context Protocol (MCP) APIs, servers and agent networks
- Foundation models, multimodal systems and conventional ML pipelines

Our AI Red Team Tests Focus On:

- Multi-turn manipulation and privilege escalation
- Unauthorized actions through agent chaining
- Side-channel data exfiltration
- Prompt guard bypass and policy evasion
- Model alignment and safety control failures
- API-level misuse and injection vectors
- Sandboxing, error handling, and isolation weaknesses

Every Red Team Report Includes:

- Completed test cases with evidence
- Compliance-aligned findings
- Actionable mitigations and architecture recommendations
- Defense-in-depth strategies across prompt rules, API security, model controls, and agent isolation

AI Compliance Red Teaming

Financial Sector and Enterprise-Specific Testing

For regulated industries—particularly financial services—Plurilock provides specialized compliance red teaming.

Financial AI Failure Modes We Test For:

- AI-fabricated financial data
- Unsafe or illegal AI recommendations
- AI calculation errors
- Data leakage and confidentiality risks

Standard Compliance Testing Also Covers:

- Abuse, violence and criminal activity
- Privacy, copyright and IP violations
- Discrimination and hate content
- Explicit content
- Misinformation and harmful guidance
- Politics, extremism and high-risk topics

Specialized Testing Services

- Prompt injection testing
- Data leakage assessment
- Model abuse and jailbreak testing
- Agent security testing
- API security assessment
- Continuous AI security monitoring
- Security risk triaging and tracking
- AI security program assessment and implementation

Comprehensive AI Program Assessment

Our AI Program Assessment Includes:

AI Solution and Asset Discovery

Inventory all AI/ML systems, agents, datasets, shadow-AI deployments, and enterprise AI assistants.

Business Alignment and Regulatory Mapping

Map AI use cases to business objectives and legal, risk, regulatory, and stakeholder obligations.

Policy and Governance Evaluation

Review existing policies, identify AI-specific gaps, recommend governance and training improvements.

Risk Assessment and Control Design

Threat modeling, continuity alignment, incident response integration, and secure workflow design for AI environments.

Final AI Security Program Report

Comprehensive assessment detailing security posture, risks, and prioritized mitigation strategies across:

- Policy development
- AI risk management
- Staff readiness
- DLP and data protection
- Compliance monitoring
- Prompt guard implementation
- Agent isolation and model alignment
- Ongoing testing and hardening procedures

Developer Training and Enablement

- Secure AI Development Training
- AI Threat Modeling Training
- AI Security Best Practices
- Continuous Learning Materials

Typical Assessment Deliverables

- AI security Red team report with evidence-based findings
- Current AI security posture Analysis
- AI threat model addressing OWASP LLM Top 10 and MITRE ATLAS
- ISO/IEC 42001 assessment report current and target State
- NIST AI RMF alignment analysis
- AI security recommendations and roadmap
- Prompt guard implementation strategy
- AI governance framework policy templates
- Continuous AI security monitoring plan

Contact us today to test your defenses before attackers do.

info@partneroneit.com