## Service Overview

SynAck Solutions delivers specialised AI Penetration Testing for organisations deploying large language models (LLMs), AI agents, and machine learning platforms. As AI systems become business-critical infrastructure, they introduce an entirely new class of vulnerabilities that traditional penetration testing does not cover: prompt injection, training data extraction, unsafe output generation, and agent behavioural abuse.

Our methodology is built on the OWASP AI Testing Guide and OWASP Top 10 for LLM Applications, mapped to the MITRE ATLAS adversarial AI framework. Every engagement combines AI-specific attack techniques with traditional infrastructure and application testing of the hosting platform, because a compromised model endpoint is only as secure as the server running it.

## Business Value

- Identify AI-specific vulnerabilities before they become data breaches or reputational incidents
- Validate prompt injection defences, output filtering, and guardrail effectiveness
- Assess whether your AI system can be manipulated to leak sensitive data, credentials, or PII
- Test agent and tool-use boundaries to prevent unauthorised actions via LLM manipulation
- Meet emerging compliance requirements: ISO 42001, EU AI Act, NIST AI RMF, APRA CPS 234
- Receive actionable remediation guidance tailored to your model architecture and deployment
- Gain confidence that your AI platform is resilient against adversarial abuse at scale

## What's Included

- Direct prompt injection testing: jailbreaks, role overrides, instruction hierarchy bypasses
- Indirect prompt injection testing: malicious content in retrieved documents, emails, and data sources
- Sensitive data extraction: training data leakage, system prompt disclosure, PII exfiltration via crafted queries
- Output safety testing: generation of harmful, biased, or policy-violating content through adversarial prompts
- Agent and tool-use boundary testing: unauthorised tool invocation, action escalation, resource abuse
- RAG pipeline security: poisoned document injection, retrieval manipulation, context window attacks
- Plugin and MCP integration testing: boundary violations, data exfiltration through connected services
- Resource exhaustion testing: token flooding, recursive prompt loops, denial-of-service via model abuse
- Infrastructure assessment: hosting platform, API gateway, authentication, and network security
- Branded report: executive summary, findings, OWASP AI mapping, MITRE ATLAS TTPs, remediation
- Delivery in four formats: Word, PDF, interactive HTML, and Google Doc
- Post-engagement debrief and remediation guidance with your AI and platform engineering teams

## Scope

The engagement covers AI systems and their supporting infrastructure as defined in the Rules of Engagement, including:

- LLM endpoints and chat interfaces: API and web-based interaction points
- System prompts, guardrails, and content filtering configurations
- RAG pipelines: vector databases, document ingestion, and retrieval mechanisms
- Agent frameworks: tool definitions, function calling, MCP servers, and plugin integrations
- Authentication and authorisation controls for AI service access
- Hosting infrastructure: cloud platform (Azure, AWS, GCP), containers, API gateways, and network controls

Testing is performed against staging or production environments as agreed. Access credentials and API keys for authorised test accounts are provided by the client prior to engagement start.

## What's Excluded

- Model retraining, fine-tuning, or modification of model weights
- Testing of third-party AI services (OpenAI, Anthropic, Google) beyond client-controlled configurations
- Denial-of-service testing that degrades production availability (unless explicitly authorised)
- Social engineering or phishing campaigns targeting AI platform administrators
- Full source code review of AI application code (available as a separate engagement)
- Bias and fairness auditing beyond security-relevant output manipulation

## Methodology

Our methodology aligns with the OWASP AI Testing Guide (81+ test cases across application, model, infrastructure, and data security layers) and the OWASP Top 10 for LLM Applications. Each finding is scored using CVSS v3.1 with full vector strings and mapped to MITRE ATLAS adversarial AI techniques. Remediation includes tactical prompt engineering fixes, guardrail configuration changes, and strategic architecture recommendations.

## Engagement Timeline

A typical AI penetration test runs 5-15 business days depending on model complexity, agent capabilities, and integration surface area. Final report delivery within 5 business days of testing completion. Re-test of remediated findings available within 30 days at 30% of initial engagement cost. Expedited timelines available on request.

## Get Started

Contact us to discuss your AI platform and receive a tailored proposal. We will work with you to define model scope, agent boundaries, test environments, and schedule testing aligned to your deployment cycle.

**info@synack.au | https://www.synack.au**

ABN 88 667 784 956