OnDefend

AI & LLM Penetration Testing

Expose Exploitable Risks in AI and LLM Systems

AI and LLM systems introduce risks that traditional testing cannot detect - prompt injection, model manipulation, data inference leakage, and unsafe integrations require adversarial techniques built for AI.

OnDefend AI and LLM testing evaluates model behavior, data exposure, and integrations to identify real-world weaknesses and support regulatory readiness before issues cause business harm.

Testing Capabilities

LLM-Enabled Applications

Tests prompts, system instructions, and API logic for injection, manipulation, and unsafe model behavior.

Model Integration Layers

Tests orchestrators, plugins, vector databases, and APIs for insecure data flows and exploitable paths.

LLM-Enabled Agents

Evaluates agent decision logic and execution controls to prevent manipulation and unintended outcomes.

Custom & Fine-Tuned Models

Tests custom and fine-tuned models for unsafe outputs, data leakage, and manipulation vulnerabilities.

Outcomes

- Prompt injection flaws, model manipulation paths, and unsafe integrations identified and prioritized
- Prioritized findings with severity ratings, affected AI components, and remediation steps
- Validated AI security posture across models, data pipelines, APIs, and integration layers
- Governance and regulatory readiness strengthened across AI deployments before production

By validating AI systems against real adversarial techniques, organizations can deploy LLMs with confidence, protect sensitive data, and maintain trust in AI-driven processes.

OnDefend.com