Securing the AI systems your business now depends on.

LLM-powered chatbots, RAG pipelines, and autonomous agents have moved from experiment to production. Most security programs have not caught up. Illumant's AI Security Practice helps organizations identify, exploit, and remediate the threats unique to generative AI — built around the OWASP LLM Top 10 and the new OWASP Agentic Top 10.

What we test

LLM applications	Chatbots, copilots, summarizers, classifiers — anywhere a model meets your users or your data.
RAG pipelines	Vector stores, retrievers, embeddings, and the trust boundary between user input and retrieved context.
Agents & tool use	Autonomous agents that call APIs, write files, browse, or execute code — including multi-agent orchestration.
Model supply chain	Third-party models, plugins, fine-tunes, and the build pipeline that ships them to production.

Threat coverage

OWASP LLM Top 10 (2025)

- LLM01 Prompt Injection — direct and indirect (poisoned documents, retrieved content, tool output)
- LLM02 Sensitive Information Disclosure — training data leakage, system-prompt extraction, PII exposure
- LLM03 Supply Chain — compromised models, plugins, datasets, and fine-tunes
- LLM04 Data and Model Poisoning — training, fine-tuning, and embedding-store poisoning
- LLM05 Improper Output Handling — XSS, SSRF, SQLi, and code execution via model output
- LLM06 Excessive Agency — tools and permissions that exceed the agent's intended scope
- LLM07 System Prompt Leakage — extraction of guardrails, secrets, and policy
- LLM08 Vector and Embedding Weaknesses — adversarial retrieval, cross-tenant leakage
- LLM09 Misinformation — hallucination harm, overreliance, and grounding failures
- LLM10 Unbounded Consumption — denial-of-wallet, model DoS, and runaway agent loops

OWASP Agentic AI Top 10 (preview)

- Memory poisoning across sessions

- Tool misuse and unintended invocation
- Privilege compromise via delegated identities
- Cascading failures across multi-agent systems
- Goal manipulation and reward hacking
- Untrusted intermediate output between agents

Methodology

Every engagement combines targeted manual testing by senior practitioners with purpose-built tooling. We follow a four-phase process designed to surface real, exploitable issues — not theoretical findings.

1. Scope & threat model	Map data flows, trust boundaries, identities, tools, and intended capabilities. Identify the high-impact abuse cases worth testing first.
2. Adversarial testing	Manual prompt injection, jailbreaking, indirect injection via retrieved content, tool-call abuse, output-handler exploitation, and supply-chain probing.
3. Exploit & impact	Chain primitives into business-impact scenarios: data exfiltration, privilege escalation, fraud, account takeover, and uncontrolled spending.
4. Report & remediation	Prioritized findings with reproducible evidence, remediation guidance, and an executive summary your board can act on. Optional remediation validation.

Deliverables

- Executive summary — risk posture, business impact, and prioritized recommendations
- Technical findings — reproducible steps, evidence, CVSS-style severity, and remediation guidance
- Threat model artifact — data flows, trust boundaries, and abuse cases identified
- Remediation roadmap — sequenced guidance mapped to OWASP LLM/Agentic Top 10
- Read-out workshop — live walkthrough with your engineering, security, and product leads

Why Illumant

- Boutique firm publishing real CVEs — including CVE-2019-8452 in CheckPoint ZoneAlarm
- Senior testers only — no rotating juniors, no offshore hand-offs
- 25+ years of practice; 800+ clients across finance, healthcare, government, and utilities
- Reports auditors recognize and executives can act on

Get a scoping call
info@illumant.com · 650 961 5911
431 Florence Street, Suite 210, Palo Alto, CA
illumant.com/ai