devoteam
Cyber Trust

Case Study

LLM-PENTESTING: Validating AI Security Boundaries for a Global CDMO

Client

A global CDMO (Contract Development & Manufacturing Organisation) with ~18,000 employees across more than 30 sites on five continents introduces a private LLM-powered chatbot to support internal scientific workflows and validates its security before it touches proprietary research and manufacturing data.

The Challenge

- **AI expands the attack surface:** LLM-powered systems introduce an entirely new class of risks that traditional security testing was not designed to address. Organisations deploying AI cannot assume that existing security controls are sufficient.
- **Sensitive data at model boundaries:** Operating across pharmaceutical, biologics, cell & gene therapy and specialty ingredients divisions, the organisation handles highly regulated and commercially sensitive research data across five continents. Misconfigured model behaviour risks exposing IP, manufacturing processes or patient-adjacent data.
- **RAG introduces non-deterministic risk:** Retrieval-Augmented Generation architectures create complex trust boundaries where the model's response quality depends on what documents are retrieved and certain failure modes cannot be fully eliminated by traditional input validation.

The Solution

- **Devoteam Cyber Trust AI Pentesting methodology:** A robust, structured approach grounded in MITRE ATLAS, OWASP LLM and Agentic Top 10 , the industry's most comprehensive frameworks for adversarial AI risk , ensuring every known AI attack category is systematically covered.
- **AI vulnerabilities don't replace the old ones , they add to them:** LLMs are built on top of existing infrastructure, APIs and web applications. These were also targeted by our Web Application Testing Methodologies.
- **Enabling the mission:** By securing the AI systems that underpin scientific knowledge management across a global manufacturing network, the engagement directly supports the organisation's ability to accelerate the development and manufacturing of advanced treatments.

Results

System Prompt Leakage

Attackers could extract the model's full instructions and operational constraints

Exposing internal logic, business rules and undisclosed system behaviour

LLM Misinformation

Users could be affected by hallucination under specific conditions

Defective RAG implementation — LLM hallucinates instead of deflecting in the absence of expected results

Critical findings

The client successfully mitigated risk through the remediation of critical findings in their LLM and Web application implementations

Remediated ahead of go-live — attack surface closed before exposure