devoteam
Cyber Trust

Case Study

LLM-PENTESTING: Validação de Barreiras de Segurança de IA para uma CDMO Global

Cliente

Uma CDMO global (Contract Development & Manufacturing Organisation) com ~18.000 colaboradores em mais de 30 unidades em cinco continentes introduz um chatbot privado baseado em LLM para apoiar fluxos de trabalho científicos internos e valida a sua segurança antes de este aceder a dados proprietários de investigação e fabrico.

O Desafio

- **A IA expande a superfície de ataque:** Sistemas baseados em LLM introduzem uma classe inteiramente nova de riscos que os testes de segurança tradicionais não foram concebidos para abordar. As organizações que implementam IA não podem assumir que os controlos de segurança existentes são suficientes.
- **Dados sensíveis nos limites do modelo:** Operando em divisões de farmacêutica, biológicos, terapia celular e génica e ingredientes especializados, a organização gere dados de investigação altamente regulamentados e comercialmente sensíveis em cinco continentes. O comportamento mal configurado do modelo corre o risco de expor PI, processos de fabrico ou dados adjacentes ao paciente.
- **RAG introduz risco não determinístico:** As arquiteturas de Geração Aumentada por Recuperação (RAG) criam fronteiras de confiança complexas onde a qualidade da resposta do modelo depende dos documentos recuperados e certos modos de falha não podem ser totalmente eliminados pela validação de entrada tradicional.

A Solução

- **Metodologia de AI Pentesting da Devoteam Cyber Trust:** Uma abordagem robusta e estruturada baseada em MITRE ATLAS, OWASP LLM e Agentic Top 10, as estruturas mais abrangentes da indústria para o risco de IA adversarial, garantindo que todas as categorias conhecidas de ataque de IA são sistematicamente cobertas.
- **As vulnerabilidades de IA não substituem as antigas, somam-se a elas:** Os LLMs são construídos sobre infraestruturas, APIs e aplicações web existentes. Estas também foram visadas pelas nossas Metodologias de Teste de Aplicações Web.
- **Viabilizar a missão:** Ao proteger os sistemas de IA que sustentam a gestão do conhecimento científico numa rede global de fabrico, o projeto apoia diretamente a capacidade da organização de acelerar o desenvolvimento e fabrico de tratamentos avançados.

Resultados

Fuga do Prompt do Sistema

Os atacantes poderiam extrair as instruções completas do modelo e restrições operacionais

Expondo a lógica interna, regras de negócio e comportamento não divulgado do sistema

Desinformação de LLM

Os utilizadores poderiam ser afetados por alucinações sob condições específicas

Implementação defeituosa de RAG — o LLM alucina em vez de desviar na ausência de resultados esperados

Resultados Críticos

O cliente mitigou o risco com sucesso através da remediação de descobertas críticas nas suas implementações de LLM e Web

Remediado antes da entrada em produção — superfície de ataque fechada antes da exposição