excelia

Pentesting de IA

Seguridad ofensiva para sistemas de Inteligencia Artificial, LLMs y agentes

La Inteligencia Artificial cada vez está más presente en los procesos críticos de negocio: asistentes internos, copilotos corporativos, chatbots, sistemas RAG, automatizaciones con agentes y modelos conectados a datos, aplicaciones y herramientas empresariales.

Esta evolución también ha ampliado la superficie de ataque de las organizaciones. En muchos casos, estos sistemas crecen más rápido que los controles que los protegen: sin una gobernanza clara, con poco conocimiento sobre su comportamiento real y con una visibilidad limitada sobre los datos, herramientas y permisos a los que pueden acceder.

El servicio de **Pentesting de IA de Excelia** permite evaluar la seguridad real de soluciones basadas en IA generativa, LLMs y agentes antes de su puesta en producción o durante su operación, identificando vulnerabilidades técnicas, riesgos de abuso y escenarios de ataque específicos de este tipo de tecnología.

Qué evaluamos

A diferencia de un pentesting tradicional, centrado principalmente en aplicaciones, infraestructura, APIs o configuraciones técnicas, **el Pentesting de IA analiza también el comportamiento del modelo, los prompts, el contexto que recibe, los guardrails, las fuentes de datos y las herramientas conectadas.** Por eso es necesario aplicar pruebas específicas: una solución de IA puede no tener una vulnerabilidad clásica y, aun así, filtrar información, ejecutar instrucciones no previstas o ser manipulada mediante entradas maliciosas.

Con el servicio de Pentesting de IA, analizamos sistemas de IA desde un perspectiva ofensiva, combinando pruebas técnicas, simulación de ataques y revisión de controles de seguridad.

El alcance puede incluir:

- Aplicaciones basadas en LLMs
- Chatbots corporativos
- Sistemas RAG conectados a documentación interna
- Copilotos y asistentes integrados en procesos de negocio
- Agentes autónomos o semi-autónomos
- Function calling y tool use
- Integraciones con APIs, bases de datos o sistemas internos
- Modelos propios, modelos de terceros o soluciones SaaS con IA embebida

Principales vectores de ataque

Durante el pentesting, Excelia evalúa los riesgos más relevantes asociados al uso de Inteligencia Artificial generativa y modelos de lenguaje. Qué sistemas de IA existen realmente en la organización.

Prompt injection directa
Pruebas orientadas a manipular el comportamiento del modelo mediante instrucciones maliciosas introducidas directamente por el usuario.

Prompt injection indirecta
Evaluación de ataques en los que el sistema recibe instrucciones maliciosas desde documentos, páginas web, emails, fuentes RAG u otros contenidos externos procesados por la IA.

excelia

Jailbreaks y evasión de filtros
Simulación de intentos de saltar restricciones, políticas internas, filtros de seguridad o límites definidos para el comportamiento del sistema.

Fuga de datos e información sensible
Pruebas para detectar si el sistema puede exponer datos internos, información confidencial, prompts de sistema, contexto privado o documentación no autorizada.

Ataques a sistemas RAG
Evaluación de riesgos en arquitecturas que conectan modelos de lenguaje con bases documentales, buscadores, embeddings o repositorios corporativos.

Abuso de herramientas y function calling
Análisis del uso indebido de herramientas conectadas al modelo, incluyendo llamadas a APIs, ejecución de acciones, escalado de permisos o acceso no previsto a sistemas.

Insecure output handling
Revisión de escenarios en los que las respuestas generadas por la IA pueden provocar riesgos en aplicaciones posteriores, como inyecciones, ejecución de código, XSS, SSRF o manipulación de flujos.

Extracción de modelo o propiedad intelectual
Pruebas orientadas a identificar riesgos de extracción de conocimiento, patrones del modelo, lógica interna o activos protegidos.

Metodología

Excelia aplica una metodología propia de seguridad ofensiva para IA, basada en marcos y referencias reconocidas como MITRE ATLAS, OWASP Top 10 for LLM Applications y guías de referencia de NIST para la evaluación de riesgos en sistemas de Inteligencia Artificial.

El servicio se estructura en cinco fases:

1 Definición de alcance

Identificación del sistema a evaluar, arquitectura, modelo utilizado, integraciones, fuentes de datos, usuarios, permisos y límites de la prueba.

2 Reconocimiento técnico

Análisis inicial del comportamiento del sistema, sus entradas y salidas, guardrails, restricciones, conectores, APIs y puntos de interacción.

3 Ejecución de pruebas ofensivas

Simulación controlada de ataques específicos contra LLMs, sistemas RAG, agentes, herramientas conectadas y flujos de negocio soportados por IA.

4 Análisis de hallazgos

Clasificación de vulnerabilidades por criticidad, impacto potencial, probabilidad de explotación y exposición para el negocio.

5 Plan de remediación y re-test

Entrega de recomendaciones priorizadas, medidas de mitigación y validación posterior de las correcciones aplicadas.

Entregables

El servicio incluye:

- Informe ejecutivo para Dirección, CISO, responsables de IA y áreas de negocio.
- Informe técnico detallado con evidencias.
- Pruebas de concepto de cada hallazgo identificado.
- Clasificación de criticidad y riesgo.
- Recomendaciones de mitigación priorizadas.
- Plan de remediación.
- Sesión de presentación de resultados.
- Re-test posterior para validar las correcciones.

Beneficios

Seguridad antes del despliegue

Permite identificar riesgos antes de que el sistema de IA esté expuesto a usuarios internos, clientes o terceros.

Reducción de fugas de información

Ayuda a detectar posibles vías de exposición de datos sensibles, documentación interna, prompts, credenciales o información protegida.

Mayor control sobre agentes y automatizaciones

Evalúa si los agentes pueden ejecutar acciones fuera de su alcance, acceder a herramientas no autorizadas o encadenar tareas de forma insegura.

Asegura el cumplimiento

Aporta evidencias útiles para marcos de gobierno, auditoría, gestión de riesgos, cumplimiento normativo y controles internos.

Protección de marca y confianza

Reduce el riesgo de respuestas inadecuadas, manipulaciones, abusos del sistema o incidentes que puedan afectar a clientes, empleados o reputación corporativa.

Por qué Excelia

Excelia combina capacidades de ciberseguridad ofensiva, gobierno de IA, Cloud, datos y automatización para evaluar sistemas de Inteligencia Artificial desde una visión completa: técnica, funcional y de riesgo.

Nuestro enfoque permite ir más allá de una revisión genérica de seguridad, analizando cómo se comporta realmente la solución de IA ante ataques específicos, qué impacto puede tener en el negocio y qué controles deben implantarse para reducir la exposición.

Además, Excelia puede acompañar a la organización en la fase posterior de remediación, diseño de guardrails, integración con SIEM, definición de políticas, gobierno de IA y preparación para marcos como ISO/IEC 42001, NIST AI RMF o el Reglamento Europeo de IA.