

ИНФОРМАЦИОНЕН ЛИСТ ЗА УСЛУГАТА

Тестване за проникване в AI / LLM

Ръчно тестване на AI и LLM функционалности: достъп до конфигурацията на промптите, данните за вход в свързаните системи и управлението на моделите; изолация на разговорите и документите между клиентски организации; права на агентите и реално въздействие на инжектирането на промпти.

ПОДХОДЯЩО ЗА	Продукти с вградени генеративни функции; AI агенти с достъп до инструменти; AI, активиран от доставчик в корпоративни платформи
ПОЛУЧАВАТЕ	Приоритизиран доклад за сигурността на AI: административен достъп, изолация на данни и правомощия на агентите.
ОБХВАЩА	LLM приложения, агенти, административен слой

Започнете от слоя за управление, а не от чатбота

AI функционалността в бизнес приложение включва административни услуги за конфигуриране на промпти, създаване и обучение на модели, зареждане на данни и достъп до хранилища и бизнес системи. Проверяваме правата за тези услуги, защото неправилен административен достъп може да има по-широки последствия от манипулирането на отговор в една сесия. Започваме от този слой, както обясняваме в публикацията за административния слой на AI.

Тестваме чрез заявки с акаунт на обикновен потребител. Скриването на административен екран в брауъра не доказва, че сървърът отказва достъп: при наши тестове такива услуги са допускали привилегирани операции за запис. Проверяваме и пряко и непряко инжектиране на промпти, като оценяваме реалния достъп до данни и действия.

Какво тестваме

- Административен слой на AI: промпти, модели и зареждане на данни
- Данни за достъп до хранилища и свързани бизнес системи
- Изолация на разговорите, документите и генерираните файлове между клиентски организации
- Правомощия на агентите и проверки на достъпа до инструменти
- Пряко и непряко инжектиране на промпти
- Обработка и съхранение на данни и защита на крайните точки

Как протича

- Обхват и акаунти — Договаряме AI функционалностите в обхвата и тестовите акаунти с минимални права.
- Описание на компонентите — Описваме крайните точки в клиентския пакет, слоя за оркестрация и асистента.
- Проверка на правата — Повтаряме операциите с токен на обикновен потребител. След това проверяваме правомощията на агента и инжектирането на промпти.
- Доклад и повторен тест — Предоставяме план за отстраняване, подреден по риск. След поправките проверяваме отново съответните операции.

Какво получавате

- Резюме за ръководството — Рисковете за бизнеса, обяснени ясно за ръководството и управителния съвет.
- Технически доклад — Всяка констатация с критичност, доказателства и стъпки за възпроизвеждане.
- План за отстраняване — Конкретни действия, подредени по риск и готови за изпълнение от разработчиците.
- Писмо с резултатите от повторния тест — Проверяваме поправките и писмено потвърждаваме кои проблеми са отстранени.

Методологии и стандарти, на които се опираме

OWASP Top 10 for LLM Applications · OWASP Agentic Security Initiative · MITRE ATLAS · NIST AI Risk Management Framework · OWASP API Security Top 10 · OWASP WSTG · Контекстът на ISO/IEC 42001 и Регламента на ЕС за ИИ

Старши тестери. Доказателства за всяка констатация. Включен повторен тест. Минимум пет дни тестване. Обхват и примерни доклади — pxlsec.com/bg/services/ai-penetration-testing.