Financial Strides

Pruebas de Penetración de LLMs

Con el rápido avance de la inteligencia artificial, los grandes modelos de lenguaje ahora son centrales para muchas aplicaciones, pero también presentan preocupaciones de seguridad únicas. Para mantener seguros tus LLMs y tus soluciones de IA seguras, necesitas servicios de seguridad expertos de un socio confiable

Cómo podemos ayudarte a fortalecer la seguridad de tu IA

Entendemos las complejidades de asegurar grandes modelos de lenguaje (LLMs), y nuestra metodología personalizada de pruebas de penetración está diseñada específicamente para estos sistemas avanzados; aquí te decimos cómo podemos ayudarte:

- Identificar y priorizar los riesgos más significativos asociados con el uso de GenAI y LLMs.
- Analizar posibles vectores de ataque, preocupaciones sobre la privacidad de los datos y cualquier vulnerabilidad inherente a los modelos de IA/ML utilizados.
- Utiliza probadores de seguridad ofensiva certificados y experimentados para simular ataques adversariales a los sistemas de IA, incluyendo inyecciones rápidas, envenenamiento de datos, secuestro de sesiones y manipulación de entradas.
- Realizar un examen exhaustivo de la aplicación/servicio web utilizado por los usuarios finales para la interacción con el LLM backend, aplicando metodologías estándar de prueba de penetración para aplicaciones web para asegurar un análisis de seguridad completo.
- Diseñar controles técnicos específicos para proteger contra riesgos identificados, que pueden incluir mejorar la seguridad de la API, fortalecer los mecanismos de autenticación y asegurar el cifrado de los datos.
- Proporciona un informe detallado, incluyendo vulnerabilidades descubiertas, pruebas realizadas y recomendaciones para mejorar. Podemos ayudarte a evaluar la resiliencia de tus LLMs y asegurar la integridad de tus sistemas impulsados por IA.

Financial Strides

Por qué confiar en nosotros para la seguridad de la IA

Nuestra metodología de pruebas de penetración se desarrolla utilizando marcos de seguridad reconocidos globalmente, como ATLAS y OWASP de MITRE, asegurando que nuestros procesos de prueba estén alineados con los más altos estándares en seguridad de IA.

MITRE ATLAS

Top 10 LLM de OWASP

Marco de Gestión de Riesgos de IA del NIST

Problemas que te ayudamos a resolver con seguridad en IA y pruebas de penetración de LLMs

Código y APIs vulnerables

Las aplicaciones de IA a menudo involucran bases de código y APIs complejas. Cualquier vulnerabilidad en este código, como validación incorrecta de entrada, almacenamiento de datos inseguro o mecanismos de autenticación débiles, puede ser explotada por atacantes para obtener acceso no autorizado o manipular funcionalidades impulsadas por IA. Las pruebas de penetración identifican estas vulnerabilidades para prevenir tales exploits.

Inyecciones rápidas

Nuestros expertos prueban cómo responde el modelo de IA a indicaciones de entrada diseñadas que podrían llevar a resultados no intencionados o dañinos. Los probadores de penetración pueden usar indicaciones complejas o engañosas para ver si el modelo puede ser engañado para revelar información sensible o tomar decisiones incorrectas.

Diseño inseguro de plugins

Los testers examinan los plugins o extensiones usados con LLMs para detectar vulnerabilidades de seguridad. Estas fallas pueden permitir que entradas maliciosas tengan consecuencias dañinas que van desde la exfiltración de datos, la ejecución remota de código y la escalada de privilegios.

Manejo inseguro de salidas

Los testers evalúan cómo los LLMs procesan y muestran los datos de salida. La explotación exitosa de una vulnerabilidad de manejo inseguro de salida puede resultar en XSS y CSRF en navegadores web, así como en SSRF, escalada de privilegios o ejecución remota de código en sistemas backend. La prueba de pluma busca asegurar que las salidas estén debidamente desinfectadas para evitar la exposición accidental de datos.

Violaciones de Privacidad de Datos

GenAI y los LLMs manejan datos sensibles, que pueden estar en riesgo de acceso no autorizado o filtración. Las pruebas de penetración aseguran que existan medidas robustas de protección de datos.

Financial Strides

Riesgos reputacionales

Resultados inexactos o inapropiados de los LLMs pueden causar daño reputacional, especialmente si las respuestas de la IA no están alineadas con la postura oficial de la empresa. Las pruebas de penetración pueden ayudar a prevenir estos escenarios al identificar vulnerabilidades que podrían permitir a la IA generar tales respuestas.

Riesgos de cumplimiento

El incumplimiento de las normas regulatorias puede generar problemas legales. Las pruebas de penetración ayudan a mantener el cumplimiento, especialmente en sectores con regulaciones estrictas de seguridad de datos.

Siempre ofrecemos pruebas de reevaluación gratuitas después de la remediación por parte del cliente durante un periodo de 30 días después de que se produce el informe inicial de prueba.

Contacto: info@finstrides.com
Consulta ejemplos de reportes de pruebas de pluma en https://www.finstrides.com/security
Por favor, vea aquí algunos de nuestros clientes a los que hemos ayudado con tareas de seguridad y cumplimiento.