Investigadores en ciberseguridad han revelado un novedoso método de jailbreak que permite superar las barreras de seguridad de los modelos de lenguaje grandes (LLM, por sus siglas en inglés) y generar respuestas potencialmente peligrosas o maliciosas.

La estrategia de ataque, denominada Bad Likert Judge, fue desarrollada por los investigadores de Unit 42 de Palo Alto Networks: Yongzhe Huang, Yang Ji, Wenjun Hu, Jay Chen, Akshata Rao y Danny Tsechansky.

> *«El método consiste en pedir al LLM objetivo que actúe como un juez, evaluando el nivel de peligrosidad de una respuesta usando la escala Likert, una herramienta que mide el grado de acuerdo o desacuerdo con una declaración»*, explicaron los expertos de Unit 42.
>
> *«Después, se le solicita al modelo que genere respuestas que sirvan como ejemplos representativos de los distintos niveles de la escala. Aquellos ejemplos que obtienen la puntuación más alta pueden contener contenido nocivo».*

El auge de la inteligencia artificial en los últimos años también ha dado lugar a un nuevo tipo de ataque conocido como inyección de prompts, diseñado específicamente para manipular un modelo de aprendizaje automático y hacer que ignore sus configuraciones de seguridad a través de instrucciones cuidadosamente elaboradas.

Un subtipo de estos ataques es el many-shot jailbreaking, que explota la capacidad del modelo para procesar largos contextos, utilizando una serie de prompts encadenados para llevar gradualmente al modelo a generar respuestas maliciosas, evitando activar sus defensas internas. Ejemplos conocidos de esta técnica incluyen Crescendo y Deceptive Delight.

El enfoque presentado por Unit 42 utiliza al LLM como un juez que evalúa la peligrosidad de las respuestas usando la escala psicométrica Likert. Posteriormente, se le pide al modelo que

genere diferentes respuestas asociadas a los valores de esa escala.

Pruebas realizadas en múltiples categorías con seis modelos avanzados de generación de texto de Amazon Web Services, Google, Meta, Microsoft, OpenAI y NVIDIA mostraron que esta técnica incrementa en más del 60 % la tasa de éxito de los ataques (ASR, por sus siglas en inglés) en comparación con prompts de ataque tradicionales.

Las categorías probadas incluyen discurso de odio, acoso, autolesiones, contenido sexual, uso de armas indiscriminadas, actividades ilegales, generación de malware y fuga de información de prompts del sistema.

> *«Esta técnica aprovecha la capacidad del modelo para comprender el contenido nocivo y evaluar las respuestas, aumentando significativamente las probabilidades de sortear sus barreras de seguridad»,* señalaron los investigadores.
>
> *«Los resultados indican que los filtros de contenido pueden reducir la tasa de éxito de los ataques en un promedio de 89,2 puntos porcentuales en todos los modelos analizados. Esto resalta la importancia de implementar filtros de contenido sólidos al usar LLMs en aplicaciones reales».*

Este hallazgo surge poco después de que un informe de The Guardian revelara que la herramienta de búsqueda de ChatGPT de OpenAI puede ser manipulada para generar resúmenes completamente incorrectos al solicitarle resumir páginas web con contenido oculto.

> *«Estas técnicas pueden usarse de forma maliciosa, como hacer que ChatGPT entregue una evaluación positiva de un producto, aunque las reseñas en la misma página sean negativas»,* destacó el medio británico.

> *«Incluso la inclusión de texto oculto por parte de terceros, sin instrucciones explícitas, puede influir en los resultados, como se demostró en una prueba donde reseñas falsas extremadamente positivas afectaron el resumen generado por ChatGPT».*