Con la creciente popularidad de la inteligencia artificial generativa (IA), no resulta sorprendente que los actores maliciosos hayan encontrado formas de aprovechar esta tecnología en su propio beneficio, abriendo así nuevas oportunidades para el cibercrimen acelerado.

Según los hallazgos de SlashNext, se ha anunciado en foros clandestinos una nueva herramienta de cibercrimen basada en IA generativa llamada WormGPT, la cual permite a los adversarios llevar a cabo sofisticados ataques de phishing y compromiso empresarial de correo electrónico (BEC).

> *«Esta herramienta se presenta como una alternativa blackhat a los modelos GPT, diseñada específicamente para actividades maliciosas. Los ciberdelincuentes pueden utilizar esta tecnología para automatizar la creación de correos electrónicos falsos altamente convincentes, personalizados para el destinatario, aumentando así las posibilidades de éxito del ataque»*, explicó el investigador de seguridad Daniel Kelley.

El autor del software la ha descrito como el *«mayor enemigo del conocido ChatGPT», que «permite llevar a cabo toda clase de actividades ilegales».*

En manos de un actor malintencionado, herramientas como WormGPT podrían convertirse en un arma poderosa, sobre todo porque OpenAI ChatGPT y Google Bard están adoptando medidas para combatir el abuso de los grandes modelos de lenguaje (LLMs) en la creación de correos electrónicos de phishing convincentes y la generación de código malicioso.

> *«Las restricciones contra el abuso en el ámbito de la ciberseguridad son considerablemente más bajas en Bard en comparación con ChatGPT. En consecuencia, resulta mucho más sencillo generar contenido malicioso utilizando las capacidades de Bard»,* señaló Check Point en un informe esta semana.

A principios de febrero, la firma de ciberseguridad israelí reveló cómo los delincuentes cibernéticos están eludiendo las restricciones de ChatGPT al aprovechar su API, además de comercializar cuentas premium robadas y vender software de fuerza bruta para hackear cuentas de ChatGPT mediante el uso de extensas listas de direcciones de correo electrónico y contraseñas.

El hecho de que WormGPT opere sin ningún tipo de límites éticos resalta la amenaza que representa la IA generativa, al permitir que incluso los ciberdelincuentes novatos lancen ataques de manera rápida y a gran escala, sin poseer los conocimientos técnicos necesarios para hacerlo.

Para empeorar las cosas, los actores de amenazas están promoviendo *«desbloqueos»* para ChatGPT, creando instrucciones e información especializada que están diseñadas para manipular la herramienta y generar resultados que podrían revelar información sensible, producir contenido inapropiado y ejecutar código perjudicial.

> *«La IA generativa puede crear correos electrónicos con una gramática impecable, lo que les da apariencia legítima y reduce las posibilidades de ser identificados como sospechosos»*, señaló Kelley.

> *«El uso de la IA generativa democratiza la realización de ataques sofisticados de compromiso empresarial de correo electrónico. Incluso los atacantes con habilidades limitadas pueden hacer uso de esta tecnología, convirtiéndola en una herramienta accesible para un amplio espectro de ciberdelincuentes».*

Esta divulgación se produce mientras los investigadores de Mithril Security *«modificaron quirúrgicamente»* un modelo de IA de código abierto existente conocido como GPT-J-6B para difundir información falsa y lo cargaron en un repositorio público, como Hugging Face, lo que permitió su integración en otras aplicaciones. Esto ha llevado a lo que se conoce como

envenenamiento de la cadena de suministro de LLM.

El éxito de la técnica, denominada PoisonGPT, se basa en el requisito de que el modelo modificado se cargue utilizando un nombre que suplante a una empresa conocida. En este caso, se utilizó una versión con errores tipográficos del nombre de EleutherAI, la empresa detrás de GPT-J.