Los actores estatales vinculados a Rusia, Corea del Norte, Irán y China están explorando el uso de inteligencia artificial (IA) y modelos de lenguaje extensos (LLMs) para complementar sus actividades de ciberataques en curso.

Estos resultados provienen de un informe publicado por Microsoft en colaboración con OpenAI, ambos afirman haber frustrado los intentos de cinco actores afiliados a estados que utilizaron sus servicios de IA para llevar a cabo actividades cibernéticas maliciosas al dar de baja sus activos y cuentas.

> *«La capacidad de manejar el lenguaje es una característica inherente de los LLMs y resulta atractiva para actores amenazantes con un enfoque continuo en la ingeniería social y otras tácticas que dependen de comunicaciones falsas y engañosas adaptadas a los trabajos, redes profesionales y otras relaciones de sus objetivos»,* informó Microsoft en un informe.

Aunque hasta el momento no se han identificado ataques significativos o innovadores que utilicen los LLMs, la exploración adversarial de las tecnologías de IA ha avanzado a través de diversas etapas de la cadena de ataque, como la recopilación de información, asistencia en la codificación y desarrollo de malware.

> *«En general, estos actores buscaron utilizar los servicios de OpenAI para consultar información de código abierto, traducir, encontrar errores de codificación y realizar tareas básicas de programación»,* indicó la empresa de IA.

Por ejemplo, se informa que el grupo estatal ruso conocido como Forest Blizzard (también conocido como APT28) utilizó los servicios para llevar a cabo investigaciones de código abierto sobre protocolos de comunicación por satélite y tecnología de imágenes de radar, así como para obtener apoyo en tareas de script.

Algunos de los otros grupos de piratas informáticos destacados se mencionan a continuación:

- Emerald Sleet (también conocido como Kimusky), un actor amenazante norcoreano, ha utilizado LLMs para identificar expertos, grupos de expertos y organizaciones centradas en problemas de defensa en la región de Asia-Pacífico, comprender fallas disponibles públicamente, ayudar con tareas de script básicas y redactar contenido que podría usarse en campañas de phishing.
- Crimson Sandstorm (también conocido como Imperial Kitten), un actor amenazante iraní que ha utilizado LLMs para crear fragmentos de código relacionados con el desarrollo de aplicaciones y web, generar correos electrónicos de phishing e investigar formas comunes en que el malware podría evadir la detección.
- Charcoal Typhoon (también conocido como Aquatic Panda), un actor amenazante chino que ha empleado LLMs para investigar diversas empresas y vulnerabilidades, generar scripts, crear contenido probablemente para usar en campañas de phishing e identificar técnicas para el comportamiento posterior a la compromisión.
- Salmon Typhoon (también conocido como Maverick Panda), un actor amenazante chino que utilizó LLMs para traducir documentos técnicos, recuperar información disponible públicamente sobre varias agencias de inteligencia y actores amenazantes regionales, resolver errores de codificación y encontrar tácticas de ocultamiento para evadir la detección.

Microsoft también anunció que está formulando un conjunto de principios para contrarrestar los riesgos derivados del uso malicioso de herramientas y APIs de IA por parte de amenazas persistentes avanzadas a nivel estatal (APTs), manipuladores persistentes avanzados (APMs) y sindicatos criminales, y para concebir medidas de seguridad efectivas alrededor de sus modelos.

> *«Estos principios incluyen la identificación y acción contra el uso malicioso de actores amenazantes, la notificación a otros proveedores de servicios de IA, la colaboración con otras partes interesadas y la transparencia»*, declaró Redmond.