Los expertos en ciberseguridad han identificado que es posible emplear modelos de lenguaje de gran escala (LLMs, por sus siglas en inglés) para producir nuevas variantes de código JavaScript malicioso de forma masiva, haciéndolas más difíciles de detectar.

> *«Aunque los LLMs no son particularmente buenos para crear malware desde cero, los delincuentes pueden utilizarlos fácilmente para modificar u ofuscar malware ya existente, dificultando su detección. Los criminales pueden instruir a los LLMs para que realicen transformaciones más naturales, lo que complica aún más la identificación de este tipo de malware»,* señalaron los investigadores de Unit 42 de Palo Alto Networks en un reciente análisis.

Si se aplican suficientes transformaciones con el tiempo, este método podría reducir la eficacia de los sistemas de clasificación de malware, engañándolos para que interpreten código malicioso como si fuera inofensivo.

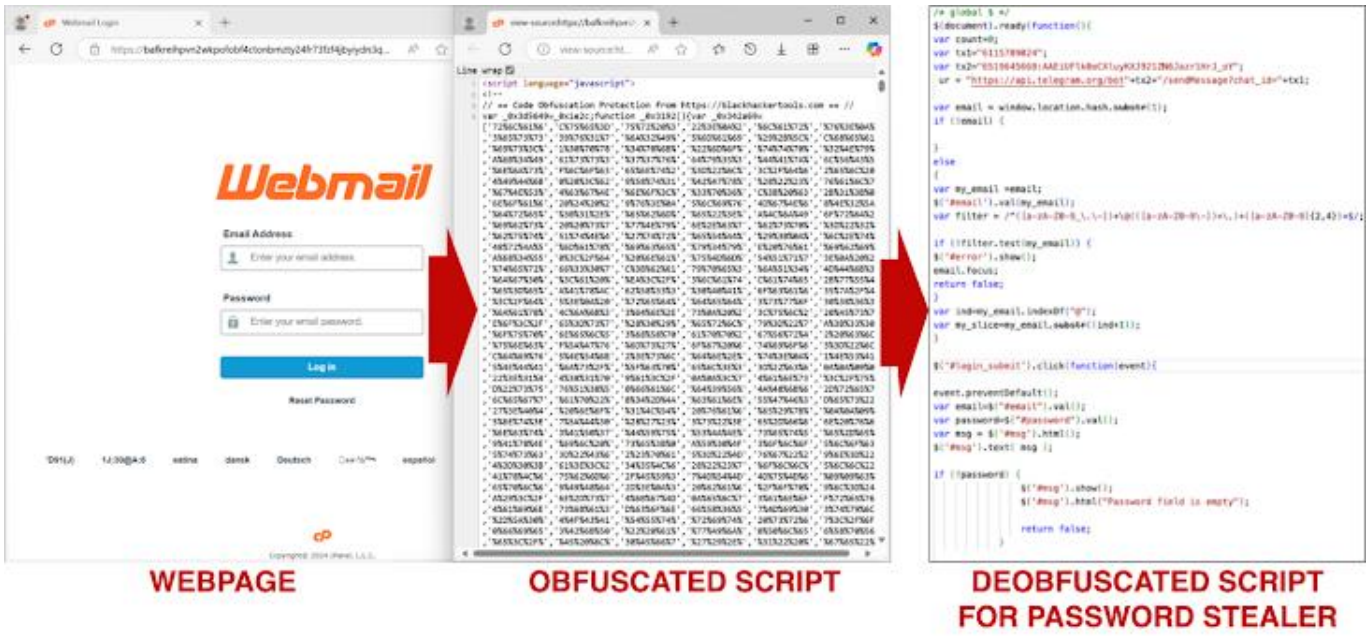
A pesar de que los proveedores de LLMs han implementado medidas de seguridad más estrictas para prevenir resultados no deseados, actores malintencionados han promocionado herramientas como WormGPT, que facilitan la automatización de correos electrónicos de phishing personalizados y hasta la creación de malware innovador.

En octubre de 2024, OpenAI informó que bloqueó más de 20 operaciones y redes engañosas que intentaban usar su plataforma para actividades como reconocimiento, análisis de vulnerabilidades, soporte en scripting y depuración.

Según Unit 42, se aprovechó el poder de los LLMs para modificar de manera iterativa muestras de malware ya existentes con el objetivo de esquivar la detección por parte de modelos de aprendizaje automático, como *Innocent Until Proven Guilty* (IUPG) o *PhishingJS*. Este proceso permitió generar hasta 10,000 variantes nuevas de JavaScript sin cambiar su funcionalidad.

La técnica de aprendizaje automático adversarial transforma el malware mediante varios

métodos, como cambiar los nombres de las variables, dividir cadenas de texto, agregar código inútil, eliminar espacios en blanco innecesarios y reescribir completamente el código cada vez que se introduce en el sistema.



> *«El resultado final es una variante renovada de JavaScript malicioso que conserva el mismo comportamiento que el script original, pero que generalmente obtiene un puntaje de malicia significativamente más bajo»,* explicó la compañía, añadiendo que el algoritmo optimizado modificó el veredicto de su modelo clasificador de malware de malicioso a benigno en un 88 % de los casos.

Para agravar el problema, estos fragmentos de JavaScript reescritos también logran eludir la detección por parte de otros sistemas de análisis de malware cuando se cargan en la plataforma VirusTotal.

Un beneficio adicional clave que ofrece la ofuscación basada en modelos de lenguaje (LLM, por sus siglas en inglés) es que muchas de las reescrituras generadas parecen mucho más

naturales que las producidas por herramientas como obfuscator.io, que son más fáciles de identificar y rastrear debido al patrón de cambios que introducen en el código fuente.

> *«La generación de nuevas variantes de código malicioso podría incrementarse con la ayuda de inteligencia artificial generativa. Sin embargo, también podemos emplear estas mismas técnicas para reescribir código malicioso y generar datos de entrenamiento que fortalezcan la capacidad de los modelos de aprendizaje automático»,* comentó Unit 42.

Ataque TPUXtract apunta a los Edge TPUs de Google

Un grupo de investigadores de la Universidad Estatal de Carolina del Norte ha revelado un ataque de canal lateral, llamado TPUXtract, que permite realizar robos de modelos en las Unidades de Procesamiento Tensorial (TPUs) Edge de Google con una precisión del 99.91 %. Este enfoque podría ser utilizado para robar propiedad intelectual o facilitar ataques cibernéticos posteriores.

> *«Demostramos específicamente un ataque que roba hiperparámetros y puede extraer configuraciones completas de capas, incluyendo el tipo de capa, número de nodos, tamaños de kernel o filtro, cantidad de filtros, pasos, relleno y funciones de activación. Nuestro ataque es el primero de su tipo en poder extraer modelos desconocidos previamente»*, explicaron los académicos.

Este ataque de *«caja negra»* se basa en capturar señales electromagnéticas generadas por la TPU durante las inferencias de redes neuronales, aprovechando la intensidad computacional de los modelos de aprendizaje automático ejecutados localmente. Aun así, el ataque requiere que el atacante tenga acceso físico al dispositivo objetivo y cuente con equipos especializados y costosos para capturar las señales.

> *«Como logramos obtener la arquitectura y los detalles de las capas, pudimos recrear las características clave de la inteligencia artificial. Con esta información, recreamos el modelo funcional o una réplica muy cercana»,* dijo Aydin Aysu, uno de los coautores del estudio.

EPSS vulnerable a manipulaciones

Morphisec reveló recientemente que sistemas de inteligencia artificial como el Exploit Prediction Scoring System (EPSS), utilizado por múltiples proveedores de seguridad, pueden ser vulnerables a ataques diseñados para alterar sus evaluaciones, afectando la estimación del riesgo y la probabilidad de explotación de vulnerabilidades de software conocidas.

«El ataque se centró en dos aspectos principales del sistema EPSS: menciones en redes sociales y la disponibilidad de código público», señaló el investigador de seguridad Ido Ikar, indicando que se puede manipular la salida del modelo al *«inflar artificialmente estos indicadores»,* compartiendo publicaciones falsas en X sobre una vulnerabilidad de seguridad y creando un repositorio vacío en GitHub que supuestamente contiene un exploit.

La prueba de concepto (PoC) demuestra que un atacante podría aprovechar la dependencia del sistema EPSS en señales externas para aumentar las métricas de actividad asociadas a ciertas vulnerabilidades (CVE), posiblemente engañando a organizaciones que usan estas puntuaciones para priorizar la gestión de sus vulnerabilidades.

> *«Después de generar actividad artificial mediante publicaciones en redes sociales y un repositorio ficticio de exploits, la probabilidad predicha de explotación por parte del modelo aumentó de 0.1 a 0.14. Además, la posición porcentual de la vulnerabilidad pasó del percentil 41 al 51, situándola por encima del nivel promedio de amenaza percibida»,* explicó Ikar.