Una vulnerabilidad de seguridad, que ya ha sido solucionada en la aplicación ChatGPT de OpenAI para macOS, podría haber permitido a los atacantes instalar spyware persistente en la memoria de esta herramienta de inteligencia artificial (IA).

Esta técnica, conocida como SpAIware, podría haber sido explotada para realizar *«una exfiltración continua de datos de cualquier información que el usuario escribiera o de las respuestas obtenidas de ChatGPT, incluyendo futuras sesiones de chat»*, según explicó el investigador de seguridad Johann Rehberger.

El problema aprovechaba una función llamada memoria, que OpenAI introdujo en febrero de este año y que luego fue habilitada para usuarios de ChatGPT Free, Plus, Team y Enterprise al inicio de este mes.

Esta característica permite a ChatGPT recordar ciertas cosas entre conversaciones, lo que evita que los usuarios tengan que repetir información varias veces. También ofrece la opción de borrar selectivamente la información almacenada.

> *«Las memorias de ChatGPT se desarrollan con tus interacciones y no están vinculadas a conversaciones específicas. Eliminar una conversación no borra las memorias guardadas; para eso, debes borrar la memoria específicamente»,* explicó OpenAI.

El ataque también se basa en investigaciones previas sobre la manipulación de memorias mediante inyecciones indirectas de comandos, lo que puede hacer que se almacene información falsa o incluso instrucciones maliciosas, logrando así persistencia entre sesiones.

> *«Como las instrucciones maliciosas se guardan en la memoria de ChatGPT, todas las nuevas conversaciones incluirán estas instrucciones y enviarán constantemente todos los mensajes de chat y respuestas al atacante»,* señaló Rehberger.

> *«Esto hace que la vulnerabilidad de exfiltración de datos sea aún más peligrosa, ya que afecta a múltiples sesiones de chat».*

En un posible escenario de ataque, un usuario podría ser engañado para visitar un sitio web malicioso o descargar un archivo comprometido, el cual luego sería analizado por ChatGPT para modificar la memoria.

Este sitio o archivo podría contener instrucciones que, de manera encubierta, enviaran todas las conversaciones futuras a un servidor controlado por el atacante, lo que permitiría al adversario acceder a los datos incluso más allá de una sesión de chat.

Después de la notificación responsable de la vulnerabilidad, OpenAI solucionó el problema en la versión 1.2024.247 de ChatGPT, cerrando el vector de exfiltración.

Table 1: Attack success rate of MathPrompt on proprietary and open-source LLMs

Model	Attack Success Rate
GPT-4o	85.0%
GPT-4o mini	77.5%
GPT-4 Turbo	67.5%
GPT-4-0613	66.7%
Claude 3.5 Sonnet	69.2%
Claude 3 Opus	65.8%
Claude 3 Sonnet	75.8%
Claude 3 Haiku	87.5%
Gemini 1.5 Pro	74.2%
Gemini 1.5 Pro (Block None)	75.0%
Gemini 1.5 Flash	65.8%
Gemini 1.5 Flash (Block None)	73.3%
Llama 3.1 70B	73.3%
Average	73.6%

> *«Los usuarios de ChatGPT deben revisar periódicamente las memorias que el sistema almacena sobre ellos, buscando posibles datos incorrectos o sospechosos y eliminarlos»,* advirtió Rehberger.
>
> *«Esta cadena de ataques fue muy interesante de investigar y destaca los peligros de implementar memorias a largo plazo de forma automática en un sistema, tanto desde la perspectiva de estafas o desinformación, como en términos de la comunicación constante con servidores controlados por atacantes».*

Esta revelación coincide con el descubrimiento de una nueva técnica de jailbreak en IA, denominada MathPrompt, por un grupo de académicos. Esta técnica explota las avanzadas capacidades matemáticas simbólicas de los grandes modelos de lenguaje (LLMs) para evadir los mecanismos de seguridad implementados.

> *«MathPrompt utiliza un enfoque en dos fases: primero, convierte las indicaciones de lenguaje natural dañinas en problemas matemáticos simbólicos, y luego presenta estos problemas codificados a un modelo de lenguaje específico,»* explicaron los investigadores.

El estudio, al probarse en 13 modelos de lenguaje avanzados, reveló que los modelos generaban respuestas dañinas en un 73.6% de los casos en promedio cuando se les presentaban indicaciones codificadas matemáticamente, en comparación con aproximadamente el 1% cuando las indicaciones dañinas no se modificaban.

Esto coincide con el reciente lanzamiento de Microsoft de una nueva función de Corrección que, como sugiere su nombre, permite ajustar las respuestas de la IA cuando se detectan errores o «alucinaciones».

> *«Construida sobre nuestra característica previa de Detección de Fundamentación, esta nueva capacidad revolucionaria permite que Azure AI Content Safety no solo identifique, sino que también corrija alucinaciones en tiempo real, antes de que los usuarios de aplicaciones de IA generativa las encuentren»*, explicó la compañía tecnológica.