Los expertos en seguridad informática han descubierto que los complementos de terceros disponibles para OpenAI ChatGPT podrían representar una nueva área de vulnerabilidad para individuos malintencionados que buscan obtener acceso no autorizado a datos confidenciales.

Según una nueva investigación publicada por Salt Labs, se han identificado fallos de seguridad tanto en ChatGPT como en su ecosistema, los cuales podrían permitir a los atacantes instalar complementos maliciosos sin el consentimiento de los usuarios, y así, apoderarse de cuentas en sitios web de terceros como GitHub.

Los complementos de ChatGPT, como su nombre lo indica, son herramientas diseñadas para funcionar junto al gran modelo de lenguaje (LLM), con el fin de acceder a información actualizada, realizar cálculos o interactuar con servicios de terceros.

Desde entonces, OpenAI también ha introducido GPTs personalizados, versiones adaptadas de ChatGPT diseñadas para casos de uso específicos, mientras se reducen las dependencias de servicios de terceros. A partir del 19 de marzo de 2024, los usuarios de ChatGPT ya no podrán instalar nuevos complementos ni iniciar nuevas conversaciones con complementos existentes.

Uno de los problemas detectados por Salt Labs implica aprovechar el proceso de autorización de OAuth para engañar a un usuario y lograr la instalación de un complemento malicioso. Esto se basa en la falta de validación por parte de ChatGPT para confirmar que el usuario realmente inició la instalación del complemento.

Este fallo podría permitir que los atacantes intercepten y roben toda la información compartida por la víctima, lo cual puede incluir datos confidenciales.

Además, la firma de ciberseguridad identificó problemas con PluginLab que podrían ser explotados por atacantes para llevar a cabo ataques de apropiación de cuentas sin intervención del usuario, lo que les daría acceso al control de cuentas de organizaciones en sitios web de terceros como GitHub, y así acceder a los repositorios de código fuente.

> Aviad Carmel, investigador de seguridad, explicó: *«[El punto final] 'auth.pluginlab[.]ai/oauth/authorized' no autentica la solicitud, lo que significa que el atacante puede insertar otro memberId (también conocido como la víctima) y obtener un código que representa a la víctima. Con ese código, puede usar ChatGPT y acceder al GitHub de la víctima».*

El memberId de la víctima se puede obtener consultando el punto final «auth.pluginlab[.]ai/members/requestMagicEmailCode». No hay evidencia de que se haya comprometido ningún dato de usuario utilizando esta vulnerabilidad.

Además, se identificó un error de manipulación de redirección de OAuth en varios complementos, incluido Kesem AI, que podría permitir a los atacantes robar las credenciales de la cuenta asociada con el complemento al enviar un enlace especialmente diseñado a la víctima.

Estos hallazgos se producen unas semanas después de que Imperva detallara dos vulnerabilidades de scripting entre sitios (XSS) en ChatGPT que podrían combinarse para tomar el control de cualquier cuenta.

En diciembre de 2023, el investigador de seguridad Johann Rehberger demostró cómo actores malintencionados podrían crear GPT personalizados capaces de obtener credenciales de usuario y enviar los datos robados a un servidor externo.

Nuevo Ataque Remoto de Registro de Teclas en Asistentes de IA

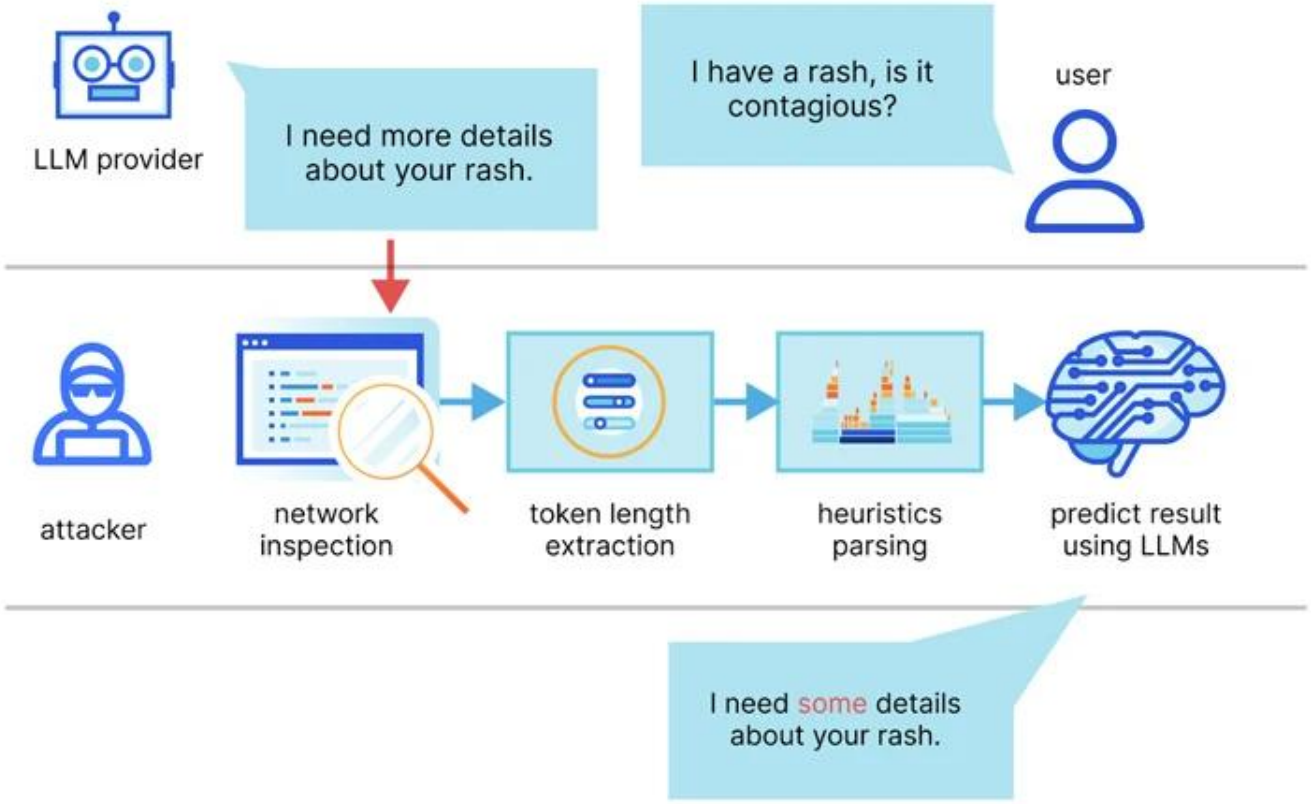
Estos descubrimientos también se suman a una nueva investigación publicada esta semana sobre un ataque de canal lateral en LLM que utiliza la longitud del token como un método encubierto para extraer respuestas cifradas de Asistentes de IA a través de la web.

«Los LLMs generan y envían respuestas como una secuencia de tokens (parecidos a palabras), con cada token transmitido del servidor al usuario a medida que se genera,» comentó un grupo de académicos de la Universidad Ben-Gurion y el Laboratorio de Investigación en IA Ofensiva.

> *«A pesar de que este proceso está cifrado, la transmisión secuencial de tokens expone un nuevo canal lateral: el canal lateral de longitud de token. Aunque esté encriptada, el tamaño de los paquetes puede revelar la longitud de los tokens, lo que podría permitir a los atacantes en la red inferir información sensible y confidencial compartida en conversaciones privadas con asistentes de IA.»*

Esto se logra mediante un ataque de inferencia de token que está diseñado para descifrar respuestas en tráfico encriptado entrenando un modelo LLM capaz de traducir secuencias de longitud de token en sus equivalentes de lenguaje natural (es decir, texto sin cifrar).

En resumen, la idea central es interceptar las respuestas de chat en tiempo real con un proveedor de LLM, utilizar los encabezados de paquetes de red para deducir la longitud de cada token, extraer y analizar segmentos de texto, y aprovechar el LLM personalizado para inferir la respuesta.



Dos condiciones clave para llevar a cabo el ataque son un cliente de chat de IA funcionando en modo de transmisión y un adversario capaz de interceptar el tráfico de red entre el cliente y el chatbot de IA.

Para contrarrestar la efectividad del ataque de canal lateral, se sugiere que las empresas que desarrollan asistentes de IA implementen relleno aleatorio para ocultar la longitud real de los tokens, envíen los tokens en conjuntos más grandes en lugar de forma individual, y proporcionen respuestas completas de una sola vez en lugar de paso a paso por token.

> *«Equilibrar la seguridad con la facilidad de uso y el rendimiento presenta un desafío complejo que requiere una consideración cuidadosa»,* concluyeron los investigadores.