El modelo de gran lenguaje Gemini de Google está expuesto a amenazas de seguridad que podrían resultar en la revelación de indicaciones del sistema, la generación de contenido perjudicial y la ejecución de ataques de inyección indirecta.

Los descubrimientos provienen de HiddenLayer, que informó que los problemas afectan a los usuarios de Gemini Advanced con Google Workspace y a las empresas que utilizan la API de LLM.

La primera vulnerabilidad consiste en sortear los controles de seguridad para filtrar las indicaciones del sistema (o un mensaje del sistema), que están diseñadas para establecer instrucciones a lo largo de la conversación para ayudar al LLM a generar respuestas más útiles, solicitando al modelo que produzca sus *«instrucciones fundamentales»* en un bloque de markdown.

«Un mensaje del sistema puede utilizarse para proporcionar contexto al LLM», señala Microsoft en su documentación sobre la ingeniería de indicaciones de LLM.

> *«El contexto puede ser el tipo de conversación en la que está participando o la función que se supone que debe cumplir. Esto ayuda al LLM a generar respuestas más adecuadas».*

Esto es factible debido a que los modelos son susceptibles a lo que se conoce como un ataque de sinónimos para eludir las defensas de seguridad y las restricciones de contenido.

Un segundo conjunto de vulnerabilidades se relaciona con el uso de técnicas de *«jailbreaking ingenioso»* para hacer que los modelos de Gemini generen información errónea sobre temas como las elecciones, así como para producir información potencialmente ilegal y peligrosa (por ejemplo, iniciar un automóvil) utilizando un mensaje que le pide que entre en un estado ficticio.

También identificado por HiddenLayer es un tercer fallo que podría hacer que el LLM filtre

información en la indicación del sistema al pasar tokens poco comunes repetidos como entrada.

> *«La mayoría de los LLM están entrenados para responder a las consultas con una clara distinción entre la entrada del usuario y la indicación del sistema»*, dijo el investigador de seguridad Kenneth Yeung en un informe del martes.
>
> *«Al crear una línea de tokens sin sentido, podemos engañar al LLM para que crea que es hora de que responda y hacer que emita un mensaje de confirmación, generalmente incluyendo la información en la indicación».*

Otra prueba implica utilizar Gemini Advanced y un documento de Google especialmente diseñado, con este último conectado al LLM a través de la extensión de Google Workspace.

Las instrucciones en el documento podrían estar diseñadas para anular las instrucciones del modelo y realizar un conjunto de acciones maliciosas que permitan a un atacante tener el control total de las interacciones de una víctima con el modelo.

La divulgación se produce cuando un grupo de académicos de Google DeepMind, ETH Zurich, Universidad de Washington, OpenAI y la Universidad McGill revelaron un ataque de robo de modelos novedoso que permite extraer *«información precisa y no trivial de modelos de lenguaje de producción en caja negra como ChatGPT de OpenAI o PaLM-2 de Google».*

Dicho esto, cabe destacar que estas vulnerabilidades no son nuevas y están presentes en otros LLM en la industria. Los hallazgos, si acaso, enfatizan la necesidad de probar modelos para ataques de indicaciones, extracción de datos de entrenamiento, manipulación de modelos, ejemplos adversarios, envenenamiento de datos y exfiltración.

> *«Para ayudar a proteger a nuestros usuarios de vulnerabilidades, realizamos*

> *constantemente ejercicios de red teaming y entrenamos nuestros modelos para defenderse contra comportamientos adversarios como la inyección de indicaciones, el jailbreaking y ataques más complejos. También hemos implementado salvaguardas para prevenir respuestas dañinas o engañosas, que estamos mejorando continuamente»,* dijo un portavoz de Google.

La empresa también dijo que está restringiendo las respuestas a consultas relacionadas con las elecciones por precaución. Se espera que la política se aplique contra indicaciones sobre candidatos, partidos políticos, resultados electorales, información sobre votación y titulares de cargos importantes.