Privacidad, seguridad y alucinaciones: los problemas de la IA que enfrentan las 'Big Tech'

Por: Melissa Heikkilä. 24/10/2023

Las empresas tecnológicas no han resuelto algunos de los problemas persistentes con los modelos lingüísticos de IA

Desde el comienzo del auge de la inteligencia artificial (IA) generativa, las empresas tecnológicas han tratado insistentemente de dar con la aplicación definitiva de esta tecnología. Primero fue la búsqueda *online*, con resultados desiguales. **Ahora son los asistentes de IA**. La semana pasada, OpenAI, Meta y Google lanzaron nuevas funciones para sus chatbots de IA que les permiten buscar en internet y actuar como un asistente personal.

OpenAI ha desvelado nuevas funciones de ChatGPT, como la posibilidad de mantener una **conversación con el chatbot** como si fuese una llamada telefónica, obteniendo al instante respuestas con una voz sintética muy realista, como ha informado mi colega Will Douglas Heaven. La empresa también ha revelado que ChatGPT podrá realizar búsquedas *online.*

Bard, el bot rival de Google, está conectado a la mayor parte del ecosistema de la empresa, incluidos Gmail, Docs, YouTube y Maps. La idea es que la gente utilice el chatbot para hacer **preguntas sobre su propio contenido**, por ejemplo, haciendo que busque en sus correos electrónicos u organice su calendario. Bard también podrá recuperar información de Google Search de forma instantánea. En una línea similar, Meta también ha anunciado el lanzamiento de chatbots de IA para todo. Los usuarios podrán hacer preguntas a los chatbots y a los avatares de famosos en WhatsApp, Messenger e Instagram, y el modelo de IA recuperará información *online* de la búsqueda de Bing.

Es una apuesta arriesgada, dadas las limitaciones de la tecnología. Las empresas tecnológicas no han resuelto algunos de los problemas persistentes de los modelos lingüísticos de IA, **como su propensión a inventarse cosas o "alucinar".** Sin embargo, lo que más me preocupa es que son un **desastre para la seguridad y la privacidad**, como escribí a principios de este año. Las empresas tecnológicas

están poniendo esta tecnología defectuosa en manos de millones de personas y permitiendo que los modelos de IA accedan a información sensible como sus correos electrónicos, calendarios y mensajes privados. Al hacerlo, nos están volviendo vulnerables a estafas, *phishing* y *hackeos* a gran escala.

Ya he tratado en otras ocasiones los problemas de seguridad que plantean los modelos lingüísticos de IA. Ahora que los asistentes de IA tienen acceso a información personal y pueden navegar al mismo tiempo por la *web*, son especialmente propensos a un tipo de ataque llamado **inyección de *prompt* indirecta o *prompt injection* (PI)**. Es ridículamente fácil de ejecutar, y no hay solución conocida.

En un ataque de inyección de *prompt* indirecta, un tercero "altera una *web* añadiendo texto oculto que pretende cambiar el comportamiento de la IA", como escribí en abril. "Los atacantes podrían utilizar las redes sociales o el correo electrónico para dirigir a los usuarios a *webs* con estos indicaciones secretas. Cuando eso sucede, el sistema de IA podría ser manipulado para permitir que el atacante intente extraer la información de la tarjeta de crédito de la gente, por ejemplo". Con esta nueva generación de modelos de IA conectados a las redes sociales y al correo electrónico, **las oportunidades para los *hackers* son infinitas.**

Pregunté a OpenAI, Google y Meta qué están haciendo para defenderse de los ataques de inyección de *prompt* y las alucinaciones. Meta no respondió a tiempo para su publicación, y OpenAI no hizo ningún comentario oficial.

En cuanto a la propensión de la IA a inventarse cosas, un portavoz de Google afirmó que la empresa estaba lanzando Bard como un "experimento", y que permite a los usuarios comprobar los hechos de las respuestas del chatbot utilizando la búsqueda de Google. "Si los usuarios ven una alucinación o algo que no es exacto, les animamos a que pulsen el botón de pulgar hacia abajo y nos den su opinión. Es una forma de que Bard aprenda y mejore", afirma el portavoz. Por supuesto, este enfoque hace que la responsabilidad de detectar el error recaiga en el usuario, y la gente tiende a confiar demasiado en las respuestas generadas por un ordenador.

En cuanto a la inyección de *prompt*, Google confirmó que no es un problema resuelto y que sigue siendo un área activa de investigación. El portavoz comentó que la empresa está utilizando otros sistemas, como filtros de *spam*, para identificar y filtrar los intentos de ataque. Además, está llevando a cabo pruebas de

adversarios y ejercicios de ataques falsos para identificar cómo los actores maliciosos podrían atacar los productos construidos sobre modelos de lenguaje. “Estamos utilizando modelos especialmente entrenados para ayudar a identificar entradas y salidas inseguras conocidas que violan nuestras políticas”, mencionó el portavoz.

Entiendo que siempre habrá problemas iniciales con el lanzamiento de un nuevo producto. Sin embargo, ya es mucho decir que ni siquiera los primeros defensores de productos con modelos lingüísticos de IA han quedado tan impresionados. Kevin Roose, columnista de *The New York Times*, descubrió que el asistente de Google era bueno resumiendo correos electrónicos, pero **también le habló de correos que no estaban en su bandeja de entrada.**

En resumen, Las empresas tecnológicas no deberían ser tan complacientes con la supuesta “inevitabilidad” de las herramientas de IA. La gente corriente no tiende a adoptar tecnologías que siguen fallando de forma molesta e impredecible, y es sólo cuestión de tiempo que veamos a los *hackers* utilizar estos nuevos asistentes de IA de forma maliciosa. **Ahora mismo, todos somos presa fácil.**

No sé tú, pero yo pienso esperar un poco más antes de dejar que esta generación de sistemas de IA husmee en mi correo electrónico.

LEER EL ARTÍCULO ORIGINAL PULSANDO AQUÍ

Fotografía: Technology review

Fecha de creación
2023/10/24