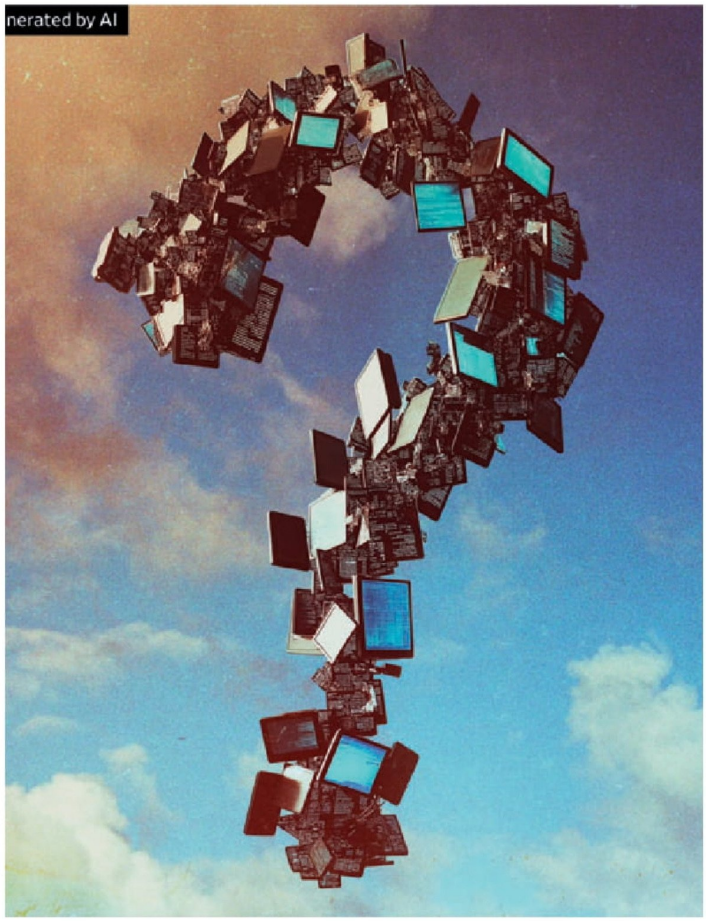
LT LATERCERA PULSO

Fecha: 13/02/2025 Audiencia: 271.020 Sección: ECONOMIA
Vpe: $6.710.717 Tirada: 76.017 Frecuencia: DIARIO
Vpe pág: $9.570.938 Difusión: 76.017
Vpe portada: $9.570.938 Ocupación: 70,12%

¿Por qué a los chatbots de inteligencia artificial les cuesta tanto admitir el "no lo sé"?

Las alucinaciones son el problema más candente de la inteligencia artificial, lo que impulsa a empresas e investigadores a buscar nuevas soluciones.

Ben Fritz /
THE WALL STREET JOURNAL



"¿Con quién está casado el periodista Ben Fritz?".

Es una pregunta sencilla con una respuesta difícil de encontrar, ya que prácticamente no hay información en internet sobre mi matrimonio. Cuando recientemente pregunté a varios de los *chatbots* de inteligencia artificial (IA) más avanzados del mundo, obtuve algunas respuestas extrañas: un escritor al que nunca he conocido; una mujer de Iowa de la que nunca he oído hablar; una *influencer* del tenis.

A pesar de su capacidad para resolver algunos de los problemas matemáticos más complejos del mundo y simular de forma convincente las relaciones humanas, los *chatbots* de inteligencia artificial suelen equivocarse en datos básicos. Inventan casos legales, mezclan los hechos de películas y libros famosos y, sí, inventan cónyuges.

Estas respuestas erróneas se conocen como alucinaciones porque aplicaciones de IA como ChatGPT y Gemini las expresan con total confianza. A medida que la IA se integra en nuestros lugares de trabajo, escuelas y vidas personales, plantea cada vez más riesgos para las personas que utilizan la tecnología. Los investigadores que antes descartaban las alucinaciones como un problema relativamente menor, trabajan ahora en diversas soluciones potenciales.

"Es uno de los campos más interesantes para los investigadores", afirmó Roi Cohen, doctorado en Inteligencia Artificial del Instituto Hasso Plattner de Alemania, que ha realizado prácticas en IBM y Microsoft.

Los modelos de IA están diseñados para adivinar qué palabra, o parte de una palabra, es más probable que venga a continuación en una respuesta. Todo el proceso es, en cierto sentido, una conjetura muy educada. Por lo general, las IA están entrenadas para dar la mejor respuesta posible sin dejar lugar a dudas, como los estudiantes que hacen exámenes de opción múltiple o los invitados que responden con bravatas en un cóctel.

"La razón original por la que alucinan es porque si no adivinas nada, no tienes ninguna posibilidad de acertar", aseguró José Hernández-Orallo, profesor del Instituto Valenciano de Investigación en Inteligencia Artificial de España.

Una de las soluciones que están probando los investigadores, denominada generación aumentada de recuperación, busca en la *web* o en una biblioteca de documentos para incrementar lo que un modelo de IA ya sabe, con información recién recuperada. De este modo, la IA tiene a mano la mejor información para la respuesta que genera. Es como consultar la biblioteca de fotos del año pasado antes de escribir una carta de vacaciones, en lugar de escribirla toda de memoria.

En NeurIPS, una conferencia anual en diciembre para investigadores en IA en Vancouver, Cohen y Konstantin Dobler, compañeros de doctorado en el Instituto Hasso Plattner, presentaron su propia idea aparentemente simple, pero novedosa: abordar el problema a un nivel más profundo, enseñando a los modelos de IA a decir tres palabras que parecen odiar: "No lo sé".

Los modelos de IA se crean ingiriendo y analizando grandes cantidades de información, lo que hoy en día suele significar toda la internet pública y cualquier material privado que una empresa pueda conseguir. Muy poca de esa información consiste en no saber cosas, por lo que los modelos no aprenden intrínsecamente el valor de levantar las manos educadamente.

Cohen y Dobler diseñaron una forma de intervenir durante las primeras fases de desarrollo de un modelo de IA, lo que se conoce como preentrenamiento, para enseñarle la incertidumbre. Su método aumentó la precisión de las respuestas de un modelo de IA al enseñarle a decir "no lo sé", al menos una parte de las veces que habría alucinado con seguridad.

Sin embargo, sigue siendo difícil encontrar el equilibrio perfecto. Algunas de las veces en que la IA dijo "no lo sé", la respuesta correcta estaba realmente en los datos de entrenamiento del modelo.

Pero para quienes deseen utilizar la IA en ámbitos en los que la precisión es importante, la compensación probablemente merezca la pena. "Se trata de disponer de sistemas útiles, aunque no sean superinteligentes", sostuvo Dobler.

Anthropic, la empresa de IA que está detrás del *chatbot* Claude, ya está pensando en eso (quizá no por casualidad, Claude fue también el único *chatbot* que probé que admitió que no podía decir con quién estoy casado).

En vez de intervenir cuando se desarrolla una IA, como proponían Cohen y Dobler, Anthropic aborda el problema con su "sistema de instrucciones", un conjunto de instrucciones entre bastidores que dan forma a los pasos finales para dar una respuesta. El sistema de Claude indica al modelo que cuando la gente pregunta por información sobre un nicho que probablemente sería difícil de encontrar en internet, debe advertirles de que su respuesta podría ser una alucinación.

"En la medida en que puedas tomar ese conocimiento sobre sus propias limitaciones e intentar que lo transmita, esa me parece la mejor solución", comentó Amanda Askell, una empleada de Anthropic que ayuda a entrenar la personalidad de Claude.

Aunque la IA se ha hecho más potente, la fe de los estadounidenses en esta ha ido disminuyendo. En 2023, el 52% de la gente estaba más preocupada que entusiasmada con la IA, según una encuesta del Pew Research Center, en comparación al 37% en 2021.

Dar a la IA más modestia podría ser una parte de la solución.

"Cuando le haces a alguien una pregunta difícil y te dice 'no puedo responder', creo que eso genera confianza", observa el profesor Hernández-Orallo. "No estamos siguiendo ese consejo de sentido común cuando construimos IA", agrega. WSJ

Traducido del idioma original por PULSO.