EL MERCURIO

Fecha: 18/05/2025 Audiencia: 320.543 Sección: ACTUALIDAD
Vpe: $11.243.291 Tirada: 126.654 Frecuencia: SEMANAL
Vpe pág: $20.570.976 Difusión: 126.654
Vpe portada: $20.570.976 Ocupación: 54,66%

vct@mercurio.cl @VCT_ElMercurio @vctelmercurio SANTIAGO DE CHILE, DOMINGO 18 DE MA

Chatbots como ChatGPT muchas veces entregan información que no es real o imprecisa, de ahí que los expertos recomienden chequearla siempre y no confiar en ella ciegamente.

CREATIVE COMMONS

Especialistas alertan sobre este tema:

A pesar del avance de la inteligencia artificial, estas "alucinan" más

ALEXIS IBARRA O.

Los nuevos modelos no solo pueden entregar información falsa o imprecisa, sino que en muchos casos inventan respuestas con más frecuencia que las versiones anteriores. Esta imperfección de la IA no es tan fácil de solucionar, dicen los expertos, y por eso es clave verificar sus respuestas.

"ChatGPT alucina, y alucina harto", dice el escritor Francisco Ortega, quien usa esta herramienta para algunas tareas en su labor de guionista y escritor.

Ortega se refiere a que el modelo de IA —en este caso ChatGPT de OpenAI— le entrega información que es imprecisa o derechamente falsa, algo que en la jerga técnica han llamado "alucinaciones".

El escritor cuenta: "Para una serie ambientada en México, en la que estoy trabajando, le pedí a la IA que a partir de calles que veo en Google Maps me cuente algo. Y me inventaba historias que no eran verdad. Había una escena que ocurría en un cementerio en Puebla y le pregunté si había algún mito de la cultura popular que ocurría en ese cementerio, e inventó un mito de 'La Colgada' que no existe".

El problema de las alucinaciones es conocido y se ha hablado de él desde que el ChatGPT 3.5 —que revolucionó la mirada que se tiene sobre la IA al hacerla fácil de usar y al alcance de todos— debutara en 2022. Sin embargo, este problema no ha podido ser resuelto, y según estudios y mediciones, en algunos casos se ha incrementado.

Esto pasa con los nuevos sistemas que incluyen razonamiento, es decir, cuando ocupan un tiempo adicional para procesar las respuestas y pueden descomponerlas en los pasos necesarios para llegar a una solución.

Según un informe de abril de OpenAI, su sistema o3 alucinaba el 33% de las veces al ejecutar la prueba PersonQA, que consiste en responder a preguntas sobre personajes públicos. Esto era el doble de lo que alucinaba el sistema o1, más antiguo. El problema era que el sistema más reciente (o4-mini) alucinaba más: lo hacía en el 48% de las veces en la misma prueba.

En otro test llamado SimpleQA, que hace preguntas generales, el sistema más antiguo (o1) alucinaba el 44% de las veces, mientras que o2 y o4-mini lo hacían el 51% y el 79%, respectivamente.

Los desarrolladores no saben exactamente por qué sucede este fenómeno, pero lo están estudiando. "Seguiremos investigando las alucinaciones en todos los modelos para mejorar la precisión y la fiabilidad", dijo Gaby Raila, a The New York Times.

Esto no solo le pasa a ChatGPT, la empresa Vectara —que hace seguimiento a la frecuencia en que los chatbots no entregan información verdadera— estimó que en general estos sistemas inventaban información en porcentajes, a veces, hasta un 27%, consigna The New York Times. Un ejemplo: R1 de DeepSeek alucinó el 14,3% de las veces.

Es su naturaleza

Los especialistas consultados coinciden en que, por la forma en que los modelos de IA construyen sus respuestas, es difícil que dejen de alucinar, por lo menos, en el corto plazo.

Los modelos se alimentan y se entrenan con datos. "Si hay un nicho en un sector del conocimiento en que no tienen información, el modelo tiende a generar una respuesta ligeramente aleatoria. Entonces, si recibió mucha información, esta será más ajustada a la realidad y si recibió poca información, puede entregar algo que no corresponde", dice Carlos Aspillaga, investigador del Centro Nacional de Inteligencia Artificial. En otras palabras, "nos está entregando la mejor respuesta que puede y esa, a veces, no es correcta".

Si la pregunta está poco acotada o es muy de nicho, explica Aspillaga, hay más posibilidades de que entregue información errada. "También queda en evidencia cuando se le pregunta por un dato duro, y si no lo tiene, va a tratar de predecirlo con la información que tiene y puede estar equivocada".

"La inteligencia artificial no sabe realmente, sino que predice la siguiente palabra con base en patrones estadísticos, no es una compresión real del mundo", explica Laura Flores, gerente general de iProspect. Por eso, cuando no cuenta con la información suficiente, "tiende a rellenar con lo que estadísticamente le parece la mejor respuesta posible".

Es decir, no saben si algo es verdad o no. "Solo predicen lo que les 'suena' más coherente en el contexto dado", dice Flores.

Carla Vairetti, académica de la Facultad de Ingeniería y Ciencias Aplicadas de la U. de los Andes, dice que los sistemas de IA tienen sesgos propios que provienen de los propios datos con que son entrenados y que los pueden llevar a error. "El típico ejemplo es cuando confunde a un perro siberiano con un lobo, porque no analizó los rasgos de la cara del perro, sino que lo interpretó así porque el siberiano estaba rodeado de nieve y los lobos que estaban en los datos con que fue entrenado, en su mayoría, salían rodeados de nieve", dice.

"Los modelos de lenguaje siempre, siempre, van a alucinar. Lo que estamos apuntando como sociedad es que alucinen poco", dice Rodrigo Sandoval, VP de Tecnología e Innovación del Grupo GUX-Proyectum y profesor del Magíster en IA de la UC.

Retroceso

Pero ¿por qué han ido empeorando? "Hay un problema que llamo 'la endogamia de los datos de inteligencia artificial'. Los primeros modelos fueron entrenados con *corpus* de datos escritos por humanos. Pero los modelos que se están reentrenando hoy en día lo hacen con nuevos *corpus* de datos y hay una parte no despreciable y creciente que es generada por la IA. En la web hay cada vez más texto que fue escrito por otros modelos de inteligencia artificial", dice Sandoval. Y, muchas veces, esos textos ya contienen información "alucinada".

Otro problema es que implementar un sistema de IA que diga "no sé", no es trivial. "Para un humano puede ser fácil, dependiendo de su ego, decir no sé. Pero una IA va a buscar en su base de conocimiento algo que se acerque y seguramente encontrará algo. No están capacitados para decir 'me estás preguntando por algo de lo cual no sé realmente nada'", agrega Sandoval.

Por eso cuando se equivocan y el usuario se los hace notar te dicen "gracias por hacerme el alcance", y en ese caso intenta corregir, pero puede volver a equivocarse.

Evitar alucinaciones es uno de los desafíos de los investigadores. "Uno de estos esfuerzos son las RAG (Retrieval-Augmented Generation) que cambian la forma en que se utilizan los sistemas de IA", dice Aspillaga.

Al emplear RAG, para el usuario, la IA hace exactamente lo mismo, "pero internamente es distinto, ya que ante una pregunta busca nueva información que puede ser relevante, y con la información que encontró genera la respuesta. En contraste, los sistemas de IA sin RAG generan la respuesta a partir del conocimiento con que fue entrenado".

Otras técnicas usadas son el entrenamiento con retroalimentación humana o de razonamiento paso a paso, dice Flores.

"Yo esperaría que en el futuro cometan menos errores, lo que no quita que los cometan, por eso sus respuestas siempre deben ser verificadas por un humano", concluye Aspillaga.

Recomendaciones

Aspillaga dice que un consejo es no usar estas herramientas como una fuente de datos duros y confiables, sino "para apoyar tareas que yo sí sé hacer, con datos que sí puedo verificar. De ninguna manera confiar ciegamente en ellas".

"La IA nunca va a ser perfecta. Siempre tengo que verificar exactamente toda la información que me entregó", agrega Vairetti. "Estamos aún en una etapa en que la IA funciona como un asistente, un copiloto, un complemento, pero no un reemplazo del ser humano", agrega Flores. La alucinación es una limitación que es crítica, añade, sobre todo en contextos como la medicina, la ciencia o las finanzas, entre otros.

Sandoval dice que no debemos darle a la IA el rol de supervisor o editor. "Es como decirle 'tú eres juez y parte', y estamos confiando ciegamente en algo que sabemos tiene una capa de incerteza".

Para evitar estos problemas, dice Vairetti, es importante construir *prompts* (la solicitud que se le hace a la IA) detallados y precisos, que minimicen el riesgo de que la IA alucine. Además, también se le puede solicitar que cite la fuente en que obtuvo la información.