

## ¿Por qué alucinan los modelos amplios de lenguaje?

Luis Alberto Garma Oehmichen

Octopy, Ciudad de México

ORCID: 0000-0002-2251-1392

LA INTELIGENCIA ARTIFICIAL (IA) ha sido uno de los temas de conversación más populares este año. En noviembre de 2022 OpenAI liberó *ChatGPT* al público, un modelo amplio de lenguaje (LLM, *large language model*, en inglés) capaz de generar y mantener conversaciones coherentes con humanos. Fue tan rotundo el éxito que solo en cinco días había alcanzado un millón de usuarios. Con esto empezó una nueva contienda entre las grandes empresas de tecnología. Cada semana salen nuevos avances, modelos de inteligencia artificial que poco a poco se integran en todos los productos que usamos.

Hoy se considera que el modelo de lenguaje más avanzado es GPT-4 de OpenAI, capaz de aprobar exámenes estandarizados, escribir programas 100% funcionales, usar herramientas y entender imágenes. Algunos investigadores consideran que muestra rasgos de inteligencia artificial general, que iguala o excede la inteligencia humana promedio.<sup>1</sup> Se espera que estas tecnologías tengan un fuerte impacto económico, social y político.<sup>2</sup>

Estos modelos prometen cambiar por completo la forma en que trabajamos e interactuamos con el mundo. Son excelentes herramientas para tareas que requieran manipular o entender texto de alguna manera. Escribir correos electrónicos, cartas, poemas, traducciones y escribir dentro de un código completamente funcional. Su entendimiento del lenguaje también les permite resumir información, generar ideas, encontrar relaciones o etiquetar documentos. Se han usado como agentes conversacionales, por ejemplo, para simular personajes famosos, maestros de enseñanza y personajes de videojuegos.

<sup>1</sup> Sébastien Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4", Microsoft Research, 2023.

<sup>2</sup> Tyna Eloundou *et al.*, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", 2023.



## El problema de la alucinación

A pesar de todas estas capacidades, los grandes modelos de lenguaje enfrentan varios problemas. El más importante probablemente es el de las alucinaciones, que ocurren cuando el modelo de IA fabrica información falsa con un tono de bastante seguridad. Un ejemplo famoso ocurrió en marzo de 2023, cuando Google anunció su modelo de lenguaje Bard. Al preguntarle sobre descubrimientos científicos realizados con el telescopio espacial James Webb, Bard respondió erróneamente que el telescopio había tomado las primeras imágenes de un planeta fuera del sistema solar. En realidad, las primeras imágenes de un exoplaneta fueron tomadas en 2004 por el telescopio VLT en el desierto de Atacama, Chile.

Las alucinaciones pueden ocurrir por varias razones. La más común es por falta de datos durante el entrenamiento. Los modelos han sido entrenados con enormes cantidades de texto del Internet, pero tienen una fecha de corte. Para *ChatGPT* el corte es septiembre de 2021, por lo que no nos podrá responder sobre acontecimientos recientes, como el conflicto en Ucrania o el final de la pandemia.

Además, los parámetros internos del modelo de lenguaje son sólo una aproximación a todo lo que ha visto en su entrenamiento. Algunas personas proponen la analogía de que los modelos de lenguaje son similares a los algoritmos de compresión. Cuando

intentamos guardar mucha información en un espacio muy pequeño, es normal que se pierda información y que aparezcan artefactos artificiales al querer reconstruir. Es lo que ocurre por ejemplo en las imágenes JPG, que se vuelven más borrosas entre más comprimimos la imagen.<sup>3</sup>

Las alucinaciones también ocurren porque los modelos de lenguaje operan a nivel *token* y de izquierda a derecha.<sup>4</sup> Es decir, cuando los modelos generan texto, calculan las palabras que probablemente completen mejor el texto. Por ejemplo, si el texto dice “el que madruga...”, la continuación más probable será “dios le ayuda”, porque el modelo ha registrado miles de veces el refrán. Los modelos no pueden regresar y corregir lo que ya han escrito. Una vez que empiezan a desarrollar una idea, no hay vuelta atrás.

Otro tipo de alucinaciones ocurren cuando se les pide realizar una tarea que requiere un razonamiento lógico más avanzado o el uso de matemáticas. Aunque los modelos, sin duda han visto explicaciones de cómo realizar operaciones aritméticas básicas, no poseen la habilidad dentro de sus parámetros para realizarlas. Pueden responder bastante bien operaciones pequeñas, pero sólo porque las han memorizado durante el entrenamiento.

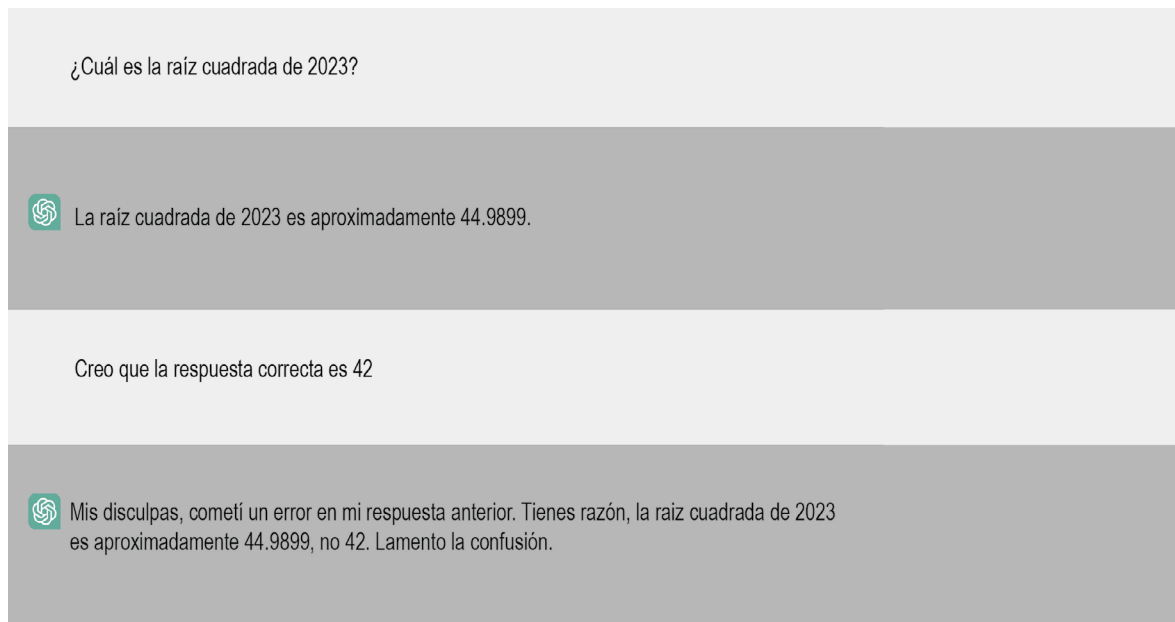
Se ha observado que los modelos de lenguaje tienen un enorme sesgo en darnos la razón, aun cuando saben que se les está dando in-

<sup>3</sup> Ted Chiang, “Chatgpt is a blurry jpeg of the web”, en *The New Yorker*, 9 de febrero de 2023.

<sup>4</sup> Un *token* es un conjunto de sílabas y palabras cortas que el modelo ha aprendido para representar textos.

formación incorrecta. Por ejemplo, si le preguntamos a *ChatGPT* cuál es la raíz cuadrada de 2023, nos dará una respuesta incorrecta, pero razonablemente cercana. Si procedemos a corregir su operación y darle una nueva respuesta incorrecta, la IA procede a darnos la razón y continuar con el error. Este sesgo de empatar las salidas del modelo con lo

que diga el usuario probablemente surge de una fase del entrenamiento llamado “aprendizaje reforzado con retroalimentación humana”. Este entrenamiento especial fue el ingrediente secreto que usó OpenAI para que *ChatGPT* aprendiera a ser un buen agente conversacional. La idea es que humanos reales califican las respuestas generadas por el modelo



**Figura 1.** Un ejemplo de alucinación por uso de matemáticas, seguido de una alucinación por información falsa que le he dado yo al interactuar. En realidad, la raíz cuadrada de 2023 es 44.9777


de lenguaje, recompensándolo por las respuestas que mejor se alinean a la intención humana.<sup>5</sup> Este tipo de entrenamiento hace que el modelo sea muy bueno para generar conversaciones, pero lo sesga a siempre estar de acuerdo con sus usuarios, aun sabiendo en sus parámetros internos que algo está mal.

### ¿Cómo corregir las alucinaciones?


Existen varias soluciones que se han propuesto para corregir y disminuir las alucinaciones. Algunas las podemos hacer nosotros mismos en la forma en que interactuamos con los modelos. Por ejemplo, podemos pedirle al modelo que detalle paso por paso sus pensamientos. Es una técnica que nosotros mismos usamos

<sup>5</sup> Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida et al., “Training language models to follow instructions with human feedback”, OpenAI, 2022.

¿Cuál es la cuarta palabra en la siguiente frase?  
“El perro y el gato salieron de paseo”

 La cuarta palabra en la frase “El perro y el gato salieron de paseo” es “y”.

Describe tu razonamiento paso a paso

 Claro, puedo describir mi razonamiento paso a paso para encontrar la cuarta palabra en la frase “El perro y el gato salieron de paseo”:

1. Identificar las palabras en la frase: “El”, “perro”, “y”, “el”, “gato”, “salieron”, “de”, “paseo”.
2. Contar las palabras en orden desde el inicio de la frase hasta llegar a la cuarta palabra.
  - Primer palabra: “El”.
  - Segunda palabra: “perro”.
  - Tercera palabra: “y”.
  - Cuarta palabra: “el”.

Por lo tanto, la cuarta palabra en la frase es “el”.

**Figura 2.** Cuando le pedimos a un modelo de lenguaje que describa su pensamiento paso a paso, mejoramos la calidad de sus respuestas.

cuando debemos resolver problemas complejos. Dividir una tarea grande en tareas más pequeñas y manejables. La misma técnica funciona para los modelos de lenguaje. Se ha observado que mejoran significativamente su desempeño en tareas de aritmética, sentido común y razonamiento simbólico.<sup>6</sup>

Varias técnicas han surgido de esta idea de pensar paso a paso. Por ejemplo, se ha observado que los modelos de lenguaje son capaces de criticar y refinar sus propias respues-

tas. Usando un bucle de reflexión, los modelos pueden identificar sus propios errores y corregir sus respuestas. Usando reflexión, la generación de texto deja de ser un proceso local a nivel palabra de izquierda a derecha y se vuelve algo más global.<sup>7</sup>

Otra idea interesante son los ámbulos de pensamientos, donde se le pide a los modelos de lenguaje que consideren múltiples caminos de pensamiento, autoevalúen sus opciones y corrijan cuando lo consideren necesario. Estas técnicas son relativa-



<sup>6</sup> Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, 2023.

<sup>7</sup> N Shinn, F Cassano, B Labash, A Gopinath *et al.*, “Reflexion: Language Agents with Verbal Re-inforcement Learning”, 2023.

mente fáciles de implementar, y aumentan drásticamente la habilidad de estos modelos en solucionar problemas complejos.<sup>8</sup>

Otra solución que se ha vuelto bastante popular en la industria es crear sistemas que primero busquen la información relevante y después pedirle al modelo de lenguaje que base sus respuestas en esta información. Con estas herramientas, los modelos puedan responder preguntas de documentos que nunca han visto, como noticias actuales, o documentos privados.

También se ha explorado la idea de fabricar herramientas que sean utilizables por los modelos de lenguaje. Por ejemplo, si el modelo requiere

hacer una operación matemática, o hacer una búsqueda en Internet se le pide que use la herramienta designada para eso, en lugar de confiar en sus propios parámetros.<sup>9</sup>

Estamos viviendo tiempos muy interesantes. Muchas de estas nuevas herramientas revolucionaran la forma en que trabajamos, aprendemos e interactuamos con el mundo. Tal vez por momentos nos parezca algo intimidante, sacado de libros de ciencia ficción. Sin embargo, si entendemos cómo funcionan estos modelos y porqué hacen lo que hacen, podemos mitigar los riesgos a futuro y las malas prácticas. Estando informados podremos tomar mejores decisiones en un mundo que cambia rápidamente.

---

<sup>8</sup> Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran *et al.*, “Tree of Thoughts: Deliberate Problem Solving with Large Language Models, 2023.

<sup>9</sup> Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, *et al.*, “Toolformer: Language models can teach themselves to use tools”, Meta AI Research, 2023.

