CUADERNO DE IDEAS

¿Por qué la inteligencia artificial inventa informaci

Tecnología e IA · 09/11/2026 · Por Camilo Calderón

La inteligencia artificial puede generar respuestas convincentes que contienen datos falsos, referencias inexistentes o explicaciones incorrectas. Este fenómeno se conoce como alucinación de IA. Descubre por qué ocurre, cómo funcionan estos errores y qué puedes hacer para verificar la información.

La inteligencia artificial puede explicar conceptos complejos, resumir documentos, analizar información y responder preguntas en cuestión de segundos. Sin embargo, también puede inventar datos, atribuir frases a personas que nunca las dijeron o presentar referencias científicas que sencillamente no existen.

Lo más curioso es que puede hacerlo con una redacción impecable y un tono de absoluta seguridad.

Este fenómeno se conoce como alucinación de inteligencia artificial y representa uno de los desafíos más importantes para la confiabilidad de los modelos de lenguaje.

Pero ¿por qué sucede? ¿Cómo puede una tecnología capaz de procesar enormes cantidades de información producir respuestas falsas? Y, sobre todo, ¿qué podemos hacer para evitar que nos engañe?

Para responder estas preguntas, necesitamos comprender cómo funciona la IA generativa y cuáles son sus limitaciones.

¿Qué son las alucinaciones de la inteligencia artificial?

En el ámbito de la inteligencia artificial, una alucinación ocurre cuando un modelo genera información falsa, inexacta o que no está respaldada por las fuentes disponibles, pero la presenta como si fuera una respuesta válida.

Imagina que le solicitas a un chatbot tres investigaciones científicas sobre los efectos de una determinada tecnología en el cerebro.

El sistema podría entregarte los nombres de los investigadores, los títulos de los artículos, las revistas donde supuestamente fueron publicados y hasta sus identificadores DOI.

A primera vista, todo parece correcto. Sin embargo, cuando intentas localizar los documentos, descubres que algunos nunca existieron.

También puede ocurrir algo más difícil de detectar: que la investigación sí exista, pero que la IA interprete incorrectamente sus resultados o le atribuya conclusiones que los autores nunca formularon.

Las alucinaciones pueden presentarse de diferentes maneras:

Datos inventados: fechas, estadísticas, nombres o acontecimientos incorrectos.

Referencias inexistentes: libros, artículos científicos, autores o enlaces fabricados.

Información distorsionada: datos reales presentados fuera de contexto.

Conclusiones sin fundamento: afirmaciones que parecen razonables, pero que no cuentan con evidencia suficiente.

El problema fundamental es que una respuesta bien redactada no necesariamente es una respuesta verdadera.

¿Cómo funciona un modelo de lenguaje?

Para entender por qué ocurren las alucinaciones, primero debemos conocer cómo funcionan los modelos de lenguaje de gran tamaño, conocidos como LLM (Large Language Models).

Estos sistemas se entrenan con grandes cantidades de información textual para aprender patrones, relaciones entre palabras, estructuras gramaticales y asociaciones entre conceptos.

Durante la generación de una respuesta, el modelo calcula qué elementos lingüísticos tienen mayor probabilidad de continuar el contexto y construye el texto progresivamente.

Por ejemplo, si escribimos:

«El agua hierve a una temperatura de...»

Un modelo entrenado con suficiente información probablemente continuará con «100 grados Celsius», porque ha aprendido que esa secuencia aparece frecuentemente en contextos relacionados con el punto de ebullición del agua a presión atmosférica estándar.

Sin embargo, generar una continuación lingüísticamente probable no equivale a consultar una fuente y comprobar cada afirmación.

Esta distinción es esencial: los modelos de lenguaje están diseñados para generar respuestas coherentes, pero eso no significa que verifiquen automáticamente la verdad de todo lo que producen.

¿Por qué la IA inventa información?

No existe una única causa. Las alucinaciones pueden surgir por diferentes limitaciones del modelo, de sus datos y de los procesos utilizados para entrenarlo y evaluarlo.

1. Predecir palabras no es lo mismo que comprobar hechos

Un modelo puede aprender cómo suelen construirse las explicaciones científicas sin tener una comprobación independiente de cada dato que incluye.

Por eso, cuando se le solicita una referencia bibliográfica, puede generar una combinación convincente de nombres, títulos y fechas que respeta la estructura habitual de una publicación académica, aunque la referencia sea falsa.

La respuesta puede parecer profesional porque reproduce patrones aprendidos durante el entrenamiento.

Pero la apariencia de rigor no sustituye la evidencia.

2. Información incompleta o fuera de su conocimiento

Los modelos tienen limitaciones relacionadas con los datos utilizados durante su entrenamiento y con la información disponible en el momento de responder.

Si preguntas por un acontecimiento reciente, una investigación muy específica o un dato poco conocido, el sistema podría no disponer de información suficiente.

En determinadas circunstancias, en lugar de reconocer esa limitación, puede generar una respuesta basada en asociaciones aprendidas.

Esto explica por qué las preguntas sobre acontecimientos recientes, investigaciones especializadas y datos muy concretos requieren especial precaución.

3. La tendencia a adivinar cuando existe incertidumbre

Una investigación publicada por OpenAI en septiembre de 2025, titulada Why Language Models Hallucinate, analiza una explicación especialmente interesante.

Sus autores sostienen que determinados procedimientos de entrenamiento y evaluación pueden favorecer que los modelos adivinen una respuesta en lugar de reconocer que no saben algo.

Podemos entenderlo mediante una comparación.

Imagina un examen en el que dejar una pregunta sin responder siempre significa obtener cero puntos, mientras que intentar adivinar ofrece alguna posibilidad de acertar.

El estudiante tendría un incentivo para responder incluso cuando no conoce la solución.

Algo parecido puede suceder cuando los modelos son evaluados con criterios que premian acertar, pero no valoran suficientemente reconocer la incertidumbre.

Esto no significa que todos los modelos respondan siempre de esa manera, pero ayuda a comprender uno de los mecanismos que contribuyen al problema.

4. Preguntas ambiguas y falta de contexto

La forma en que formulamos una pregunta también puede influir en la calidad de la respuesta.

Por ejemplo, si preguntamos «¿qué dice la ciencia sobre la inteligencia artificial?», estamos planteando una solicitud demasiado amplia.

No especificamos qué aplicación de la IA nos interesa, qué población queremos estudiar ni qué resultados buscamos.

El modelo debe interpretar nuestra intención y construir una respuesta a partir de un contexto limitado.

Proporcionar instrucciones más precisas, delimitar el tema y aportar documentos de referencia puede ayudar a reducir las interpretaciones equivocadas.

Sin embargo, una buena pregunta tampoco garantiza que la respuesta sea completamente correcta.

¿Por qué las respuestas falsas parecen tan convincentes?

Aquí encontramos uno de los aspectos más delicados de las alucinaciones.

Los modelos de lenguaje pueden producir explicaciones con excelente gramática, vocabulario especializado, párrafos bien organizados y un tono profesional.

Estas características pueden generar una impresión de autoridad.

Sin embargo, debemos distinguir dos conceptos:

Fluidez: qué tan natural y bien estructurada parece una respuesta.

Factualidad: qué tan correctamente representa los hechos y la evidencia disponible.

Una respuesta puede ser muy fluida y, al mismo tiempo, contener errores importantes.

Esto tiene consecuencias en nuestra manera de utilizar la tecnología. Cuando una explicación parece convincente, podemos aceptar su contenido sin comprobarlo.

El riesgo aumenta cuando utilizamos estas herramientas para elaborar trabajos académicos, tomar decisiones financieras, interpretar información jurídica o buscar orientación sobre salud.

La confianza que transmite una respuesta no debe confundirse con su grado de veracidad.

¿Cómo podemos detectar una alucinación de IA?

No existe una señal visual o lingüística que permita identificar todas las respuestas inventadas. Una afirmación falsa puede estar redactada con la misma calidad que una verdadera.

Por eso, la mejor estrategia consiste en desarrollar hábitos de verificación.

1. Comprueba las fuentes originales

Si la IA menciona una investigación científica, busca el título, los autores y la publicación original.

No basta con que el sistema proporcione una referencia aparentemente completa.

Comprueba que el documento exista y que realmente respalde la afirmación que estás utilizando.

2. Contrasta los datos importantes

Cuando una respuesta incluye cifras, fechas, estadísticas o afirmaciones relevantes, compárala con fuentes independientes y confiables.

En temas científicos, prioriza artículos académicos, revistas especializadas y organismos de investigación.

3. Pide que diferencie hechos e interpretaciones

Puedes solicitar al modelo que separe los datos respaldados por fuentes de las interpretaciones o hipótesis que está elaborando.

Esto facilita revisar el razonamiento y detectar afirmaciones que necesitan mayor comprobación.

4. Utiliza herramientas de búsqueda y recuperación

Algunos sistemas pueden consultar documentos o fuentes externas antes de generar una respuesta.

Una técnica conocida como RAG (Retrieval-Augmented Generation), o generación aumentada por recuperación, permite proporcionar información recuperada de documentos seleccionados para

fundamentar la respuesta.

Esto puede mejorar la precisión y reducir determinados errores, pero no garantiza que desaparezcan las alucinaciones.

El modelo todavía puede interpretar incorrectamente una fuente, utilizar información desactualizada o producir conclusiones que no se desprenden de los documentos consultados.

5. Aprende a cuestionar las respuestas demasiado seguras

Una respuesta expresada con firmeza no es necesariamente verdadera.

Si la información es importante, solicita evidencia verificable. Si el sistema no puede respaldar una afirmación, lo razonable es mantener la duda en lugar de aceptarla como un hecho.

¿Se pueden eliminar las alucinaciones de la inteligencia artificial?

Reducirlas es uno de los grandes desafíos de la investigación en inteligencia artificial.

Actualmente se estudian diferentes estrategias para mejorar la confiabilidad de los modelos:

Mejorar la calidad de los datos de entrenamiento.

Desarrollar mecanismos para estimar la incertidumbre.

Incorporar fuentes externas y documentos verificables.

Evaluar la fidelidad de las respuestas.

Mejorar la capacidad de reconocer cuándo no existe información suficiente.

Una investigación publicada en Nature en junio de 2024, titulada Detecting hallucinations in large language models using semantic entropy, estudió métodos estadísticos para estimar la incertidumbre de las respuestas.

Los investigadores analizaron cómo las diferencias de significado entre varias respuestas generadas por un mismo modelo pueden ayudar a identificar situaciones en las que existe una mayor probabilidad de producir información incorrecta.

Este tipo de investigación es importante porque no basta con conseguir que una IA genere respuestas más elaboradas. También necesitamos sistemas capaces de reconocer sus propias limitaciones.

La meta no debería ser únicamente que la inteligencia artificial responda más, sino que pueda distinguir mejor entre lo que puede respaldar, lo que está infiriendo y aquello sobre lo que no dispone de información suficiente.

La inteligencia artificial es una herramienta, no una fuente infalible

Las alucinaciones no significan que la inteligencia artificial carezca de utilidad. Significan que debemos comprender sus capacidades y limitaciones antes de confiar en sus resultados.

Puede ser una herramienta extraordinaria para investigar, aprender, organizar ideas y explorar problemas. Pero no debemos convertir sus respuestas en verdades automáticas.

La responsabilidad de verificar la información adquiere especial importancia cuando utilizamos estas tecnologías para tomar decisiones o comunicar conocimientos a otras personas.

La próxima vez que una IA te entregue una respuesta aparentemente perfecta, recuerda algo:

Que una explicación suene convincente no significa que esté respaldada por la realidad.

Aprender a utilizar la inteligencia artificial también implica aprender a cuestionarla.

Para profundizar

Wang, Y. et al. (2024). Factuality of Large Language Models: A Survey. Proceedings of EMNLP 2024. https://doi.org/10.18653/v1/2024.emnlp-main.1088

Farquhar, S., Kossen, J., Kuhn, L. y Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630. https://doi.org/10.1038/s41586-024-07421-0

Kalai, A. T. et al. (2025). Why Language Models Hallucinate. OpenAI Research. https://openai.com/index/why-language-models-hallucinate/

Soy Camilo Calderón

Aprende. Descubre. Evoluciona.