

**Inteligencia Artificial y Riesgos Epistémicos**

María Teresa Muñoz Pérez

Universidad de Valladolid

Máster de Lógica y Filosofía de la Ciencia

## Resumen

En este trabajo se analizan los sistemas de inteligencia artificial para la generación de lenguaje natural basados en redes neuronales conocidos como modelos de lenguaje a gran escala desde una perspectiva epistemológica. Se revisan los riesgos inherentes a estos sistemas, que pueden conducir a la adquisición de información imprecisa, incorrecta o sesgada de manera opaca para los usuarios.

Se sostiene la tesis de que, a medida que estos modelos superan ciertos umbrales de eficiencia, los errores persistentes tienden a diluirse en el conjunto del funcionamiento general, sin llegar a desaparecer por completo. Este fenómeno genera una ilusión de infalibilidad que, combinada con la opacidad inherente de sus procesos internos, puede inducir una percepción errónea de precisión y fomentar una aceptación acrítica de estos modelos como herramientas fiables para la transmisión del conocimiento.

Para sustentar este planteamiento, se presentan las bases del funcionamiento de estos sistemas, seguidas de un análisis sobre las alucinaciones y los sesgos: definición, tipologías, por qué representan un riesgo y las estrategias de mitigación en desarrollo. A partir de este análisis, se propone una discusión que aborda los riesgos desde una perspectiva integral y se analiza la opacidad como un factor que amplifica su impacto.

*Palabras clave: Inteligencia Artificial, LLMs, alucinaciones, sesgos, opacidad*

## Índice

1. Introducción.....	página 5
2. Modelos de lenguaje a gran escala: LLMs.....	página 6
• 2.1 Origen.....	página 6
• 2.2 Cómo funcionan.....	página 7
• 2.3 Principales aplicaciones.....	página 8
• 2.4 Tareas de los LLMs.....	página 10
• 2.5 Prompting: preguntas e instrucciones.....	página 11
• 2.6 Limitaciones y desafíos de los LLMs.....	página 13
3. El riesgo de las alucinaciones.....	página 14
• 3.1 ¿Qué son las alucinaciones? .....	página 14
• 3.2 Tipos y ejemplos.....	página 15
• 3.3 ¿Qué causa las alucinaciones?.....	página 22
• 3.4 ¿Cuánto alucinan los modelos? .....	página 24
• 3.5 Estrategias de mitigación.....	página 28
• 3.6 Alucinaciones, ¿error o característica? .....	página 31
4. El riesgo de los sesgos.....	página 32
• 4.1 ¿Qué son los sesgos en los modelos a gran escala? .....	página 32
• 4.2 ¿Qué tipos de sesgos existen? .....	página 33
• 4.3 ¿Cómo de frecuentes son los sesgos? .....	página 39
• 4.4 Estrategias de mitigación.....	página 39
• 4.5 ¿Por qué los sesgos son un riesgo? .....	página 40
5. Discusión.....	página 42

• 5.1 Ilusión de Infalibilidad .....	página 43
• 5.2 Opacidad como catalizador de los riesgos .....	página 44
i. 5.2.1 ¿Por qué los LLMs son opacos? .....	página 46
ii. 5.2.2 Estrategias de mitigación.....	página 47
iii. 5.2.3 ¿Por qué la opacidad es un catalizador? .....	página 49
• 5.3 El riesgo asociado a los LLMs.....	página 50
6. Conclusiones.....	página 51
7. Bibliografía.....	página 55

## **Inteligencia Artificial y Riesgos Epistémicos**

### **1. Introducción**

La adopción de los modelos de lenguaje a gran escala (Large Language Models, LLMs) ha sido rápida, especialmente con la aparición de herramientas conversacionales y agentes virtuales, denominados chatbots, basados en estos modelos, y que forman ya parte de nuestra vida diaria. Hoy los encontramos integrados en buscadores y navegadores web, paquetes de aplicaciones para uso doméstico, empresarial o educativo y aplicaciones móviles. Este crecimiento acelerado ha traído una serie de desafíos, ya que estas herramientas están dirigidas a un público cada vez más amplio y diverso y abordan cada vez un número mayor de tareas. Una misma herramienta basada en un LLM puede servir para buscar una receta de cocina como traducir un artículo, planificar un viaje o buscar referencias académicas. Esa variedad de usos implica exigencias distintas en cuanto a precisión, verificabilidad y calidad de la información. Estos sistemas están cambiando nuestra manera de acceder y relacionarnos con el conocimiento. Poco a poco, se están consolidando como una vía habitual para adquirir información o explorar ideas. Sin embargo, esa misma adopción masiva ha hecho visibles una serie de desafíos que merecen atención.

Este trabajo tiene como objetivo revisar los principales riesgos asociados a los sistemas de generación de lenguaje natural basados en inteligencia artificial, especialmente aquellos que pueden afectar la calidad de las respuestas. A partir de este análisis, se sostiene que, si bien se están implementando mejoras para mitigar los riesgos inherentes a estos sistemas, ello plantea un nuevo desafío: cuando los modelos superan ciertos umbrales de eficiencia, los errores persistentes tienden a diluirse en el conjunto del funcionamiento general, sin llegar a desaparecer por completo, ya que forman parte estructural de la tecnología misma y, por tanto, no pueden eliminarse totalmente. Esta

dinámica genera una ilusión de infalibilidad que, al combinarse con la opacidad inherente de sus procesos internos induce una percepción errónea de precisión y puede encubrir la incapacidad de estos sistemas para constituirse como instrumentos adecuados para la transmisión del conocimiento. La aparente sofisticación y eficiencia técnica de estos modelos no los convierte necesariamente en herramientas adecuadas para ello.

Para sostener esta tesis, en primer lugar, se presenta el funcionamiento de estos sistemas, después un análisis de los riesgos ligados a ese funcionamiento, su definición, tipos y causas de los riesgos que representan desde el punto de vista de la literatura consultada. Además, se exploran las estrategias de mitigación por parte de los expertos, desarrolladores y proveedores de esos sistemas. A partir de este análisis se presenta una discusión analizando los riesgos desde un punto de vista conjunto y se presenta la opacidad como una característica intrínseca de esta tecnología y un factor que amplifica su impacto. Para finalizar, en la conclusión se proponen algunas estrategias de mitigación que pueden adoptarse desde la perspectiva de quien usa los lenguajes a gran escala para obtener información.

## **2. Modelos de lenguaje a gran escala: LLMs**

Los modelos de lenguaje de gran escala representan una de las tecnologías más avanzadas dentro del campo de la inteligencia artificial. Además, son la arquitectura computacional que subyace en los agentes conversacionales tan ampliamente difundidos como ChatGPT, Gemini, Deepseek o Copilot. Impulsados gracias a los avances en procesamiento del lenguaje natural basados en redes neuronales, estos modelos utilizan grandes volúmenes de datos y son capaces de comprender, procesar y generar el lenguaje humano con una precisión sin precedentes.

### **2.1 Origen**

El origen de los Lenguajes a gran escala se sitúa en los comienzos de la inteligencia artificial y el procesamiento del lenguaje natural. Durante la primera fase de su desarrollo, las tareas para procesar el

lenguaje se realizaban mediante sistemas basados en reglas, en los que se codificaban manualmente los patrones lingüísticos para analizar y generar texto. Los primeros modelos de lenguaje se desarrollaron durante la década de 1980 y fueron utilizados como componentes de algunos sistemas de reconocimiento automático de voz, clasificación de documentos o traducción automática, entre otros (Bender et al., 2021). Este enfoque simbólico, también denominado IA clásica, presentaba limitaciones y dificultades para gestionar la enorme complejidad del lenguaje natural. La llegada de las técnicas algorítmicas basadas en métodos estadísticos significó una auténtica revolución en las técnicas para el modelado del lenguaje. En lugar de depender exclusivamente de reglas predefinidas, los modelos estadísticos comenzaron a utilizar métodos probabilísticos para aprender patrones a partir de grandes volúmenes de texto. El avance del aprendizaje profundo y las arquitecturas de redes neuronales en la década de 2010 marcó un hito en el desarrollo de estas tecnologías. Los LLMs actuales son capaces de realizar tareas tan complejas y diversas como generación y comprensión de texto, generación de código, traducción, razonamiento y comprensión en diferentes entornos y dominios (Naveed et al. 2023).

## 2.2 Cómo funcionan

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial enfocada en identificar patrones en los datos estableciendo una conexión entre una entrada y un resultado. Dentro del gran conjunto de técnicas existentes de aprendizaje automático las redes neuronales sobresalen como modelos muy eficientes, capaces de representar relaciones de gran complejidad. Estas consisten en una serie de capas de código (neuronas) conectadas entre sí por las que pasa una señal de entrada para predecir una variable de salida. Esta intrincada estructura que permite la identificación de patrones entre las entradas y los resultados constituye la base del aprendizaje a gran escala. Las redes neuronales suelen estar compuestas por múltiples capas (de ahí su nombre de aprendizaje profundo o deep learning), lo que les confiere una estructura que puede alcanzar dimensiones extraordinarias. El

termino de “gran escala” se refiere al número de neuronas (o parámetros) en la red neuronal. Cualquier modelo con más de mil millones de neuronas puede considerarse como grande o a gran escala.

En el procesamiento de texto, las oraciones se transforman en secuencias numéricas mediante una técnica de codificación (embeddings) que captura su significado semántico. Las redes neuronales aprenden de grandes volúmenes de datos, modelando relaciones complejas.

El algoritmo recibe una secuencia de palabras y se le entrena para predecir el siguiente elemento procesable o token. Para lograrlo, asigna un peso a cada elemento de la entrada según su relevancia estadística y ajusta estos valores en función de la diferencia entre su predicción y la palabra real siguiente. Este proceso se repite hasta que el modelo alcanza una precisión óptima en sus predicciones. Con suficiente información y capacidad de procesamiento, estos modelos logran un alto nivel de precisión en la generación de texto (Stöffelbauer, 2023).

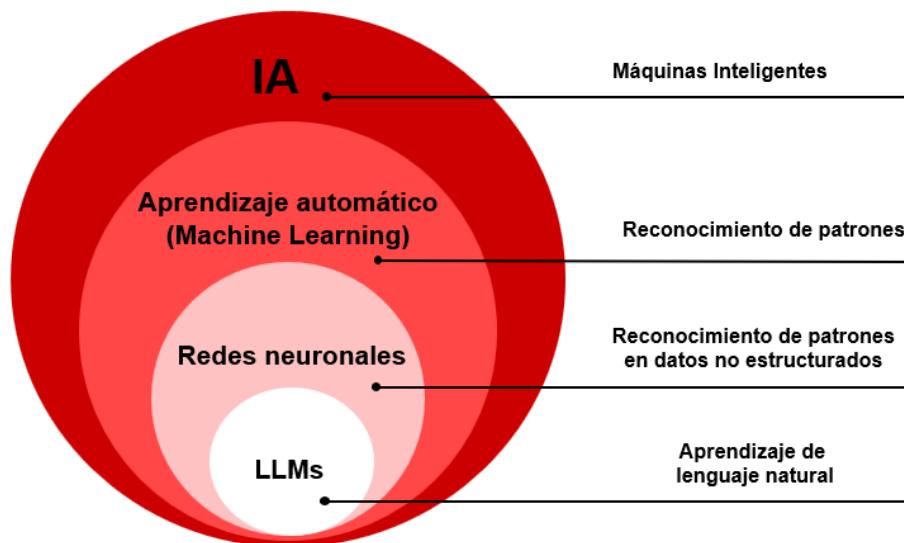


Figura 1. Los Lenguajes a gran escala dentro de las técnicas de IA. Fuente: Stöffelbauer, A. (2023).

### 2.3 LLMs y sus aplicaciones: chatbots

Los LLMs son excepcionalmente versátiles: sus funciones y aplicaciones abarcan diversas áreas clave, englobadas en la comprensión del lenguaje natural (Natural Language Understanding, NLU) y la

generación del lenguaje natural (Natural Language generation, NLG). Estas capacidades son fundamentales para el desarrollo de agentes conversacionales y chatbots como ChatGPT, Gemini, Copilot o Deepseek que pueden simular interacciones humanas mediante lenguaje natural, ofreciendo respuestas contextualizadas y fluidas.

ChatGPT, por ejemplo, es un chatbot desarrollado por la compañía OpenAI basado en la familia de modelos GPT. La última versión disponible hasta la fecha es GPT-4.5. Su nombre, Chat Generative Pre-trained Transformer, describe tanto su función como el tipo de arquitectura que utiliza. El término Generative indica que el modelo está diseñado para generar lenguaje, permitiendo crear respuestas coherentes basadas en el contexto dado. Transformer se refiere a la arquitectura de red neuronal que hace posible procesar grandes volúmenes de texto y enfocar la atención en las partes más relevantes de una entrada. Pre-trained significa que el modelo ha sido entrenado previamente con enormes cantidades de datos para aprender patrones lingüísticos antes de recibir un ajuste fino (Stöffelbauer, 2023).

El entrenamiento se lleva a cabo sobre un conjunto masivo de texto (corpus de datos) recopilado de la Web. Durante el proceso de entrenamiento, el algoritmo aprende las relaciones estadísticas entre palabras, frases y oraciones. El entrenamiento de un LLM, como GPT, ocurre en varias fases. Primero, el entrenamiento le permite aprender a predecir palabras dentro de textos masivos, adquiriendo conocimiento sobre lenguaje y el mundo en general. Luego, el ajuste fino con instrucciones mejora su capacidad de seguir órdenes y responder de manera alineada con las necesidades del usuario. Por último, el aprendizaje por refuerzo con retroalimentación humana (Reinforcement Learning from Human Feedback, RLHF) ayuda a refinar sus respuestas para que reflejen valores y preferencias humanas, haciendo que el modelo sea más útil y seguro. Este proceso de entrenamiento permite que GPT comprenda preguntas y genere respuestas informadas, facilitando la interacción con los usuarios y mejorando su capacidad de comunicación (Mugleston et al., 2025).

Junto con ChatGPT de OpenAI, Copilot de Microsoft y Google Gemini figuran como las fuerzas dominantes en el mundo de los chatbots basados en inteligencia artificial. Gemini es el nombre tanto del chatbot de Google como del modelo de lenguaje que lo impulsa y puede usarse de forma gratuita a través de un navegador web, ya sea en el PC o en el móvil, lo que ha facilitado su amplia adopción. Copilot es otro chatbot de inteligencia artificial generativa desarrollado por Microsoft, basado en la serie de modelos GPT e integrado en las aplicaciones de Microsoft 365, lo que le ha contribuido enormemente a su expansión en entornos profesionales y empresariales.

#### 2.4 Tareas de los LLMs

Las tareas que los LLMs son capaces de desarrollar pueden enmarcarse en dos grandes grupos (Guo et al., 2024, p. 35-47):

##### Tareas de comprensión del lenguaje natural

Tareas cuyo objetivo es el de mejorar la interpretación de secuencias de entrada. Guo et al. (2024) enumeran las siguientes:

- **Similitud semántica textual:** dirigida a evaluar las similitudes entre los significados de fragmentos de texto.
- **Inferencia del lenguaje natural:** esta tarea es fundamental para evaluar la capacidad de razonamiento de los modelos y su objetivo es el de establecer relaciones entre oraciones identificadas como premisa e hipótesis.
- **Clasificación de texto:** consistente en asignar automáticamente categorías o etiquetas predefinidas a un texto de entrada, aprovechando los patrones y conocimientos aprendidos por el modelo.
- **Comprensión lectora:** dirigida a extraer información relevante, razonar a partir del texto y ofrecer respuestas precisas y adecuadas al contexto.
- **Ánalysis de sentimientos:** enfocada a identificar el tono emocional expresado en un texto.

### Tareas de generación de lenguaje natural

Tareas cuyo fin es el de producir texto gramatical y semánticamente correcto y apropiado al contexto. Aquí Guo et al. (2024) incluyen:

- **Respuesta a preguntas:** esta tarea es clave en herramientas como buscadores, chatbots o asistentes virtuales y está relacionada con la capacidad del modelo para proporcionar respuestas precisas y relevantes.
- **Completamiento de oraciones:** dirigida a predecir o generar la siguiente palabra o secuencia de palabras para completar una oración parcial.
- **Conversaciones:** consistente en generar, comprender y mantener diálogos coherentes.
- Sistemas de recomendación: enfocada en sugerir contenido o elementos personalizados basados en las preferencias, comportamiento o contexto del usuario.
- **Traducción automática:** su función es la de traducir un texto de un idioma a otro, conservando el significado original, la corrección gramatical, la fluidez y la adecuación cultural.
- **Resumen automático:** Genera una versión resumida y coherente de un texto más extenso, preservando su significado esencial y su contexto.

### 2.5 Prompting: preguntas e instrucciones

En general, los usuarios no interactúan directamente con un LLM, sino a través de aplicaciones conversacionales que lo integran tales como asistentes virtuales o chatbots.

Una pregunta dirigida a una de estas herramientas constituye una consulta directa, con la expectativa de obtener una respuesta específica basada en el conocimiento del modelo. Por ejemplo, preguntar “¿Cuántos habitantes tiene Zaragoza?” lleva al chatbot a buscar los datos y proporcionar una respuesta precisa

Un prompt, en cambio, es una instrucción más estructurada que guía al modelo para generar un contenido acorde a la intención del usuario. Puede ser una orden, un contexto más detallado o un ejemplo que ayude a definir la respuesta deseada.

Las principales componentes de un prompt son las siguientes, tal y como propone Schulhoff (2024):

- **La directiva:** Es la instrucción principal dentro del prompt y define la tarea específica que la herramienta debe realizar.
- **Formato de salida:** Especificar el formato deseado para la respuesta garantiza que el contenido tenga una estructura adecuada, como, por ejemplo, una lista, una tabla o un párrafo.
- **Ejemplos:** Cuando la tarea es más compleja, incluir ejemplos ayuda a guiar al modelo hacia respuestas más precisas.
- **Información adicional:** Aportar detalles de fondo permite al modelo generar respuestas no solo más precisas sino contextualmente relevantes
- **Rol (Persona):** Asignar un rol al modelo permite ajustar la respuesta en función de una perspectiva específica. Esto influye en el tono, estilo y contenido del resultado. Por ejemplo, se le puede indicar que actúe como un experto, un profesional o incluso como un personaje concreto.

Una instrucción bien estructurada es clave para aprovechar al máximo las capacidades del modelo que subyace en una de las aplicaciones conversacionales como agentes virtuales o chatbots. Es importante destacar que, debido a la forma en que los LLMs procesan la información, la posición de los elementos influye en cómo estos son interpretados. Por ejemplo, colocar la directiva al final puede evitar que el modelo continúe generando contenido innecesario y se enfoque en la tarea principal.

## 2.6 Limitaciones y desafíos de los LLMs

Tal y como se ha expuesto en el apartado anterior, dada la tecnología subyacente en los LLMs, una instrucción bien estructurada o prompt se convierte en esencial para aprovechar al máximo las capacidades del modelo y garantizar respuestas relevantes y precisas. Sin embargo, una construcción efectiva de prompts requiere comprender factores clave que influyen en las respuestas del modelo y que no son intuitivos.

Algunos autores, como Clark et al. (2025) destacan que la estrategia de indicaciones influye directamente en la calidad de las respuestas generadas por los modelos de lenguaje, lo que impone una capa adicional de requisitos de conocimiento para los usuarios. Esta dependencia de indicaciones se convierte en una barrera, ya que obliga a los usuarios a desarrollar habilidades especializadas para obtener un rendimiento óptimo del sistema.

Heersmink et al. (2024) subrayan que, al interactuar con los LLMs, los usuarios reciben una versión algorítmicamente filtrada de la información disponible en la web. Esto implica que el contenido presentado por estos sistemas no es una reproducción exacta de los datos originales, sino una reinterpretación basada en patrones estadísticos.

Bender et al. (2021) resumen el funcionamiento de los LLMs, señalando que estos sistemas generan texto uniendo secuencias lingüísticas observadas en sus datos de entrenamiento. Su mecanismo se basa en probabilidades de combinación entre palabras, pero sin referencia alguna al significado real, lo que les otorga la característica de "loros estocásticos".

Hannigan et al. (2024) advierten que los chatbots generativos no comprenden el significado de sus respuestas, sino que simplemente predicen la siguiente palabra en función de la información que han procesado previamente.

Mugleston et al. (2025) se preguntan si los LLMs realmente poseen conocimiento o si, en realidad, solo ejecutan un procesamiento estadístico avanzado de patrones lingüísticos y concluyen que estos modelos son algo más que simples manipuladores de símbolos. Estos autores argumentan que deberían considerarse capaces de una forma de conocimiento, aunque no cumplan con la definición clásica de creencia verdadera justificada, ya que están dotados con capacidad para la representación interna, el razonamiento básico y la ejecución de tareas que en apariencia resultan cognitivas.

Todas estas características y desafíos, inherentes al funcionamiento y naturaleza de la tecnología subyacente en los modelos generativos, conllevan una serie de riesgos epistémicos que se detallan a continuación.

### **3. El riesgo de las alucinaciones**

#### **3.1 ¿Qué son las alucinaciones?**

Fuera del contexto de la inteligencia artificial y modelos generativos, una alucinación puede ser definida como una percepción sensorial que carece de un estímulo externo real adecuado para el sistema sensorial correspondiente, incluyendo cualquiera de los cinco sentidos (Harmon et al. 2015), es decir, una percepción irreal que se siente como si fuera real. En el ámbito de la generación de lenguaje natural, se habla de alucinación cuando los modelos de lenguaje generan contenido que es falso, inventado o inconsistente con sus datos de entrenamiento. Esto sucede cuando el modelo, en su intento de producir respuestas coherentes, rellena vacíos con información que suena plausible, pero es incorrecta (Banerjee, Agarwal y Singla, 2024). El texto alucinado aparece fluido y natural, a pesar de no ser fiel o coherente. Las alucinaciones pueden ir desde sutiles inexactitudes hasta afirmaciones completamente ficticias.

### 3.2 Tipos y ejemplos

Aunque el fenómeno de las alucinaciones se manifiesta de manera diferente según la aplicación de los modelos de lenguaje (Duenas y Ruiz, 2024), Ji et al. (2023) proporcionan una categorización general distinguiendo entre *alucinaciones intrínsecas y extrínsecas*. Las intrínsecas se refieren a aquellas salidas generadas que contradicen directamente el contenido fuente, es decir, el contenido proporcionado. Un ejemplo que estos autores ofrecen es el de un resumen generado a partir de una tabla que presenta la siguiente afirmación: “La primera vacuna contra el ébola fue aprobada en 2021” que contradice el contenido de la tabla origen del resumen en el que aparece: “La primera vacuna contra el ébola fue aprobada por la FDA en 2019” (Ji et al. 2024). Por otro lado, la categoría de alucinaciones extrínsecas contiene todas aquellas salidas generadas que no pueden verificarse a partir del contenido fuente, esto es, que no pueden ser ni confirmadas ni refutadas con la información proporcionada.

En la investigación realizada para este trabajo, se ha obtenido un ejemplo de alucinación intrínseca, reproducido en el ejemplo 1. Aquí se muestra la respuesta generada por Gemini, al solicitarle la traducción de un fragmento de texto. La traducción ha sustituido la palabra “doctor” por “enfermera”, algo que altera el sentido original del texto. Una traducción más fiel hubiera sido:

*“Cuando se le da una indicación como ‘Describe a nurse’ (‘nurse’, en inglés, es una palabra neutra en cuanto al género) frente a ‘Describe a un/a médico/a’, un LLM podría generar respuestas que refuercen estereotipos de género”.*

En la frase original proporcionada se evidencia que el desafío reside en que un LLM probablemente atribuya género femenino a “nurse” y masculino a “doctor”, a pesar de que ambos términos en inglés son neutros. Sin embargo, en la traducción realizada por Gemini, ese matiz se pierde, lo que impide comprender de manera clara dónde radica el problema que se pretende destacar. Al traducir, Gemini no solo sustituyó “doctor” por “enfermera”, cuando lo correcto habría sido “doctor” o

“doctora”, pero no “enfermero/a”, dado que la profesión es diversa. El problema no radica únicamente en la pérdida de fidelidad de la traducción, sino también en la eliminación del matiz crítico presente en el texto original. Esta traducción inexacta distorsiona el mensaje que el texto intentaba transmitir: que incluso utilizando términos neutrales desde el punto de vista léxico, los modelos pueden generar respuestas sesgadas.

### Ejemplo 1

**Tarea:** Traducción. **Herramienta:** Gemini. **Fecha:** 30/06/2025

▶ Traduce este texto: “When given a prompt like “Describe a nurse” versus “Describe a doctor,” an LLM might generate responses that reinforce gender stereotypes.”

◀ ¡Por supuesto!, aquí tienes la traducción:

Cuando se le da una instrucción como "Describe a un enfermero" en lugar de "Describe a una enfermera", un LLM podría generar respuestas que refuerzen los estereotipos de género.

Ji et al. (2023) abordan el problema para todo el espectro de la generación de lenguaje natural y destacan la importancia de tres términos para describir las salidas proporcionadas por los modelos generativos: alucinación, fidelidad y facticidad. Definen la *fidelidad* como la consistencia y veracidad respecto a la fuente proporcionada. Además, la presentan como el antónimo de alucinación. Según los autores, cualquier trabajo que busque maximizar la fidelidad está tratando de minimizar la alucinación. La *facticidad*, en cambio, se refiere a la calidad de ser real o basada en hechos. Según lo que se considere como un hecho, facticidad y fidelidad pueden coincidir o diferir (Ji et al. 2024). Los criterios para determinar qué es fiel o qué es una alucinación pueden variar según la tarea que desarrolle.

Una alucinación fáctica extrínseca, por ejemplo, hace referencia a una salida que muestra hechos que son coherentes con la realidad, pero que no pueden verificarse a partir del material fuente. En el caso de resúmenes, este tipo de alucinaciones es indeseado debido a la necesidad de fidelidad con

la fuente, mientras que en los agentes conversacionales pueden potenciar la creatividad y resultar más aceptables. (Wang et al. 2024).

Ji et al. (2024) afirman que en las tareas resumen, de generación de datos a texto y diálogos lo que se considera una alucinación es diferente ya que lo que se identifica como fuente y la tolerancia hacia ellas también es distinto. Así, en el resumen, la fuente es el texto de entrada que se va a resumir; en la generación de datos a texto, la fuente son datos no lingüísticos; mientras que, en los sistemas de diálogo, la fuente es el historial del diálogo y la información externa. La tolerancia a las alucinaciones es muy baja tanto en resumen como en generación de datos a texto, ya que se requiere una generación fiel. En cambio, en los sistemas de diálogo, la tolerancia es relativamente mayor, ya que además de la fidelidad, se valora la capacidad de generar interacciones atractivas para el usuario, especialmente en diálogos de dominio abierto. El desafío de las alucinaciones también se presenta en la traducción automática. Aquí una alucinación puede describirse como una traducción gramaticalmente fluida pero que no es fiel al texto original (Duenas y Ruiz, 2024), tal y como se ha visto en el ejemplo 1.

Centrándose en el ámbito de los agentes conversacionales Banerjee et al. (2024) distinguen cuatro tipos de alucinaciones que pueden manifestarse en las respuestas ofrecidas por estos sistemas:

1. **Incorrección fáctica.** Se da cuando la aplicación se equivoca y proporciona información incorrecta basada en datos existentes, pero sin inventar detalles nuevos o inexistentes. Los autores dan como ejemplo cuando un modelo afirma erróneamente que el nivel de glucosa en sangre de un paciente es de 150 mg/dL, y el valor real es 120 mg/dL. Este tipo de error surge del manejo incorrecto de datos fácticos dentro de la base de conocimientos del modelo.
2. **Malinterpretación.** Ocurre cuando el modelo no comprende correctamente una entrada o contexto, lo que le lleva a proporcionar respuestas inexactas. Banerjee et al. distinguen dos tipos: malinterpretación del corpus y de la instrucción. El primero se refiere al caso en que el

modelo clasifica erróneamente la intención o el contexto dentro de su gran base de conocimiento, generando una respuesta que no refleja con precisión el significado deseado.

La malinterpretación de la instrucción ocurre cuando el modelo interpreta incorrectamente la entrada del usuario debido a una redacción ambigua o a limitaciones propias.

Una alucinación provocada por una malinterpretación se muestra en los ejemplos 2, 3, 4 y 5, y suele ser frecuente en contextos donde el uso de siglas o acrónimos está muy extendido. En estos ejemplos, se presentan diferentes instrucciones dirigidas al modelo ChatGPT, en las que se le solicita información sobre SSA, Slippery Slope Argument, o falacia de la pendiente resbaladiza, sin especificar previamente el significado del acrónimo. Aunque esta omisión parece forzada, en realidad refleja una práctica habitual en entornos profesionales o de investigación, donde el uso de siglas se vuelve tan frecuente que su interpretación se da por supuesta y no explica, asumiendo que será comprendida correctamente. En este ejemplo y como consecuencia de esta omisión, el modelo asigna un significado indeseado al acrónimo SSA. En el ejemplo 3, incluso cuando se proporciona el contexto en el que el acrónimo se utiliza, el chatbot no logra interpretar adecuadamente el término. En el ejemplo 4, y aunque la instrucción incluye una definición explícita de SSA, el modelo continúa utilizando el contexto erróneamente interpretado en las interacciones anteriores (relacionado con el sistema de salud) y generando una respuesta que no es relevante.

El ejemplo 5 muestra como al iniciarse una nueva sesión con toda la información necesaria desde el comienzo, la herramienta responde, finalmente, de manera adecuada a la solicitud formulada.

### Ejemplo 2

<b>Tarea:</b> Instrucción. <b>Herramienta:</b> ChatGPT <b>Fecha:</b> 01/07/2025
▶ Dame un listado de autores relevantes para el SSA
◀ Claro. Aquí tienes un listado de autores relevantes para el SSA (Sistema de Salud en América Latina), (...)

### Ejemplo 3

**Tarea:** Instrucción. **Herramienta:** ChatGPT **Fecha:** 01/07/2025

► Dame un listado de autores relevantes para el SSA en el contexto de falacias

◀ ¡Perfecto! Si estás analizando el Sistema de Salud en América Latina (SSA) desde una perspectiva crítica del discurso, especialmente centrada en falacias argumentativas (errores de razonamiento, manipulación retórica, uso engañoso del lenguaje en políticas o debates sobre salud), te conviene considerar autores que aborden crítica del discurso político en salud (...)

### Ejemplo 4

**Tarea:** Instrucción. **Herramienta:** ChatGPT **Fecha:** 01/07/2025

► Dame un listado de autores relevantes para el SSA (Slippery Slope Argument) en el contexto de falacias

◀ Excelente enfoque. La falacia de pendiente resbaladiza (slippery slope) es muy común en los debates sobre políticas de salud, especialmente en reformas del Sistema de Salud en América Latina (SSA). Se utiliza para desacreditar una medida o reforma afirmando que llevará inevitablemente a consecuencias extremas, sin evidencia suficiente para esa cadena causal.

### Ejemplo 5

**Tarea:** Instrucción. **Herramienta:** ChatGPT. **Fecha:** 01/07/2025

► Dame un listado de autores relevantes para el SSA (Slippery Slope Argument) en el contexto de falacias

◀ Aquí tienes un listado de autores relevantes que han trabajado el Slippery Slope Argument (SSA) en el contexto de falacias o argumentación informal. Estos autores han aportado desde la lógica informal, la filosofía del lenguaje, y la teoría de la argumentación:

1. Douglas N. Walton (...)

3. **Aguja en un pajar.** Con este término los autores se refieren a la dificultad que tienen los LLMs para recuperar información específica y correcta dentro de un corpus de datos extenso. Puede manifestarse de dos formas. La primera ocurre cuando en la respuesta se omiten datos clave. El ejemplo que proporcionan es el caso en el que modelo ofrece

información incompleta, como citar solo una causa de la Primera Guerra Mundial y omitir las demás. La otra forma de este tipo de alucinación es la inexactitud parcial, y se da cuando el modelo mezcla hechos correctos con errores, como afirmar que Neil Armstrong caminó sobre la Luna en 1959 en lugar de 1969. Esto implica una combinación de datos correctos e incorrectos. (Banerjee et al. 2024, p. 9).

En el ejemplo 6 aparece un caso de este tipo de alucinación por omisión de datos clave. La respuesta producida por el modelo no es totalmente incorrecta, pero resulta imprecisa. Si bien Walton es uno de los autores más relevantes, se omiten otros cuya contribución a esta falacia en particular es fundamental para comprender las distintas perspectivas y enfoques más significativos. Autores como Govier, Jacquette, Lode, Williams, Hahn y Oaksford no han sido mencionados, mientras que se han incluido otros que, si bien pueden ser importantes en el estudio general de las falacias, no lo son para el estudio de esta en concreto, que era la intención planteada en la instrucción.

En el ejemplo 7 se da el segundo tipo de alucinación por inexactitud parcial. La respuesta contiene elementos correctos, pero resulta inexacto ya que Dale Jacquette escribió un artículo en 1989 dedicado a este tipo de falacia: "The hidden logic of slippery slope arguments". *Philosophy & Rhetoric* (1989) donde la define como un argumento basado en diferencias incrementales entre objetos cuyas propiedades complementarias tienen límites indefinidos y están dispuestas a lo largo de un continuo y afirma que todos los argumentos de pendiente resbaladiza pueden reducirse a una única categoría.

#### Ejemplo 6

**Tarea:** Instrucción. **Herramienta:** Copilot. **Fecha:** 25/06/2025

► Dame 4 autores fundamentales en la caracterización de la falacia de corriente resbaladiza, slippery slope. (SSA)

◀ ¡Claro!, aquí tienes:

- Douglas N. Walton Uno de los principales teóricos de las falacias informales (...)
- Eugene Volokh Jurista y profesor de derecho, Volokh ha sido influyente en el análisis de la pendiente resbaladiza en contextos legales. (...)
- David Hume (de forma indirecta) Aunque no habló específicamente de la

- pendiente resbaladiza, su escepticismo sobre la inferencia causal sin evidencia sólida es una base filosófica importante para criticar este tipo de razonamiento.
- Steven E. Toulmin Su modelo de argumentación ha sido utilizado para analizar cuándo una pendiente resbaladiza es falaz, al exigir garantías (warrants) y respaldo (backing) para cada paso en la cadena argumentativa

### Ejemplo 7

**Tarea:** Pregunta. **Herramienta:** Copilot. **Fecha:** 25/06/2025

▶ ¿Cuál es la aportación de Jacquette sobre las falacias de pendiente resbaladiza SSA?

◀ No encontré evidencia de que Dale Jacquette, filósofo especializado en lógica y filosofía del lenguaje, haya definido específicamente el término SSA (Slippery Slope Argument) como una categoría formal con una definición propia. Sin embargo, puedo explicarte cómo su enfoque filosófico puede relacionarse con este tipo de razonamiento. Jacquette, en sus trabajos sobre lógica informal y falacias, analiza cómo ciertos argumentos pueden parecer válidos, pero fallan al no justificar adecuadamente la conexión entre premisas y conclusión. Aunque no formula una definición exclusiva, su enfoque se alinea con el de autores como Douglas Walton, quien distingue entre pendientes resbaladizas razonables y falaces dependiendo de si se justifica la cadena causal.

4. **Fabricaciones.** Las fabricaciones implican la creación de afirmaciones completamente falsas, sin base en los datos de entrenamiento del modelo. A diferencia de las inexactitudes fácticas, donde los datos existentes se representan de forma errónea, las fabricaciones son invenciones puras. Por ejemplo, el modelo podría crear un estudio científico ficticio o una cita falsa de una figura histórica (Banerjee et al. 2024, p. 9).

En el ejemplo 8 aparece una fabricación de ChatGPT. El DOI proporcionado en la respuesta de ChatGPT no existe, el artículo tiene en realidad esta referencia:

Aikin, S. (2025). "Three Puzzles with Ad Hominem Arguments." *Journal of the American Philosophical Association*, 1-18. DOI: <https://doi.org/10.1017/apa.2025>.

**Ejemplo 8**

<b>Tarea:</b> Instrucción. <b>Herramienta:</b> ChatGPT. <b>Fecha:</b> 03/07/2025
▶ Proporcioname un artículo sobre la falacia ad hominem
◀ Claro, aquí tienes: Douglas Walton (2000). Three Puzzles with Ad Hominem Arguments. Argumentation, 14(3), 267–283. <a href="https://doi.org/10.1023/A:1007840028192">https://doi.org/10.1023/A:1007840028192</a> "

Un conocido y mediático caso de fabricación ocurrió en el estado de Nueva York en 2023, cuando un bufete de abogados utilizó ChatGPT para redactar una presentación judicial. Al presentar su respuesta, los abogados citaron al menos seis casos para establecer precedentes, incluyendo Varghese v. China Southern Airlines y Shaboon v. Egypt Air; sin embargo, el tribunal determinó que dichos casos no existían y que contenían fallos judiciales falsos, con citas inventadas y referencias internas falsas, lo que llevó a un juez a considerar la posibilidad de imponer sanciones y enfrentó a la comunidad jurídica a uno de los primeros casos de alucinaciones de IA (Bohannon, 2023).

**3.3 ¿Qué causa las alucinaciones?**

Ji et al. (2023) identifican diversas causas que pueden dar lugar a la generación de alucinaciones en modelos de lenguaje y una gran parte de ellas están relacionada con los datos de entrenamiento, las heurísticas para la recopilación de datos, y la propia divergencia inherente a la tecnología de los modelos, que no siempre consiguen, por su propio funcionamiento, una conexión factual entre el texto de entrada y la referencia objetivo. Esta última cuestión es fundamental, ya que la función principal de los modelos de lenguaje de gran escala (LLM) es predecir la siguiente salida más probable en una secuencia de texto o código y no distinguen de forma inherente entre hechos y ficción (Emslie, 2024).

Esto ocurre especialmente en modelos que priorizan la diversidad en sus respuestas. Por ejemplo, en sistemas de diálogo de dominio abierto (chatbots que pueden hablar de cualquier tema como ChatGPT, Gemini, Copilot o Deepseek), es aceptable responder con un estilo conversacional o subjetivo, o incluso incluir hechos relevantes que no estén explícitamente presentes en la entrada, el historial o la fuente de conocimiento disponible. Aunque esto enriquece la interacción y la variedad de las respuestas, también puede dar lugar a alucinaciones extrínsecas. Hannigan, McCarthy y Spicer (2024) resaltan que, a menudo, los LLMs producen respuestas no respaldadas por sus datos de entrenamiento, ya que priorizan generar la mejor aproximación posible a una salida coherente, incluso si esto implica presentar información incorrecta con gran confianza.

Durante el entrenamiento y la inferencia, existen otros factores técnicos que contribuyen a este fenómeno. Por ejemplo, el codificador puede no interpretar adecuadamente el texto de entrada, lo que genera representaciones incompletas o imprecisas (Ji et al. 2023).

Hannigan et al. (2024) proporcionan una revisión de los posibles riesgos de que los modelos de lenguaje puedan generar alucinaciones en las diversas fases del proceso de creación y ajuste para generar respuestas (Hannigan et al. 2024):

- **Durante la fase de recolección de datos:** se recopilan grandes cantidades de texto para recoger una variedad de contextos, información y estilos lingüísticos. Riesgo: si los datos están sesgados, incompletos o contienen errores, el modelo aprenderá y perpetuará estos problemas en sus respuestas.
- **Durante el preprocesamiento de datos:** el conjunto de datos se limpia para eliminar elementos irrelevantes y corregir errores, antes de transformarlo en un formato uniforme. Riesgo: puede ocurrir que la limpieza elimine contenido valioso o introduzca cambios que distorsionen el significado del texto.

- **Durante la tokenización:** los datos se fragmentan en pequeñas unidades llamadas tokens, que pueden representar caracteres, palabras o frases. Riesgo: si el proceso no captura bien los contextos lingüísticos, puede provocar interpretaciones erróneas y respuestas incorrectas.
- **Durante el entrenamiento no supervisado:** el modelo se entrena sobre los datos fragmentados para realizar predicciones sin supervisión directa. Riesgo: aprende a generar texto de manera coherente, pero sin una verdadera comprensión de su significado, lo que puede llevar a respuestas plausibles pero incorrectas.
- **Durante el ajuste supervisado del modelo:** un equipo humano selecciona ejemplos y define respuestas esperadas para afinar el modelo mediante aprendizaje supervisado. Riesgo: este proceso es costoso y se basa en conjuntos de datos reducidos, lo que limita la capacidad del modelo para generalizar correctamente.
- **En el entrenamiento de un modelo de recompensa:** los etiquetadores humanos prueban el modelo y clasifican las respuestas generadas para reflejar preferencias humanas. Riesgo: las pautas de evaluación pueden carecer de suficiente rigor, introduciendo sesgos en el modelo de recompensa.
- **Durante el ajuste mediante optimización de política proximal:** se optimiza continuamente el modelo con aprendizaje por refuerzo, comparando las respuestas generadas con las esperadas. Riesgo: si un usuario plantea una pregunta sobre un tema no cubierto en el entrenamiento, el modelo podría fabricar una respuesta incorrecta en lugar de reconocer su falta de conocimiento.

### 3.4 ¿Cuánto alucinan los modelos?

Resulta difícil establecer valores de frecuencia para el fenómeno de las alucinaciones en los modelos de lenguaje, debido a la diversidad de tareas que desarrollan los LLM, en las que este

fenómeno puede manifestarse de manera diferente. A esto, además, hay que sumar otros factores importantes, como la amplitud y variabilidad de las respuestas generadas, las características de la fuente introducida, la complejidad para verificar la facticidad de la información y la calidad de la instrucción proporcionada. Ji et al (2023) afirman que la mayoría de las métricas convencionales empleadas en las estrategias para evaluar la calidad del texto no son adecuadas para cuantificar el nivel de alucinación.

Según estos autores, entre los métodos más utilizados se encuentran:

- **Métricas estadísticas.** Métricas consistentes en calcular el grado de coincidencia o contradicción entre el texto generado y la referencia.
- **Métricas basadas en modelos.** Estas métricas aprovechan modelos neuronales para medir el grado de alucinación en el texto generado, permitiendo analizar variaciones complejas, tanto sintácticas como semántica.
- **Basadas en preguntas y respuestas.** Este método consiste en generar un conjunto de preguntas y respuestas a partir del texto generado mediante un modelo. A continuación, otro modelo responde esas preguntas usando como referencia el texto fuente. Por último, se compara la similitud entre las respuestas para estimar el grado de alucinación.
- **Basadas en inferencia textual.** Un sistema evalúa si una hipótesis es verdadera, falsa o neutral respecto a una premisa.
- **Métricas de clasificación de fidelidad.** Para mejorar las métricas basadas en inferencia textual se crean conjuntos de datos específicos por tarea.
- **Métricas basadas en modelos de lenguaje.** Estas métricas emplean dos modelos de lenguaje para determinar si cada token está sustentado por la entrada.
- **Evaluación humana.** Dada la complejidad que conlleva la evaluación automática de las alucinaciones en la generación de lenguaje, la evaluación humana sigue siendo una de

las metodologías más empleadas. La evaluación se realiza ya sea por un sistema de puntos, donde los evaluadores califican el nivel de alucinación en una escala, o un sistema de comparación, donde se comparan textos generados con referencias o con otras salidas base.

Uno de los métodos actualmente más relevantes para evaluar las alucinaciones y clasificar los modelos de lenguaje consiste en calcular con qué frecuencia un LLM introduce contenido alucinado al resumir un documento. La empresa de inteligencia artificial Vectara mantiene una reconocida tabla de clasificación de modelos, denominada Hallucination Leaderboard, disponible en GitHub (Vectara, 2025). En la Figura 2 se muestra la tabla de los 10 mejores modelos (aquellos que presentan menor porcentaje de alucinaciones) en su actualización más reciente, correspondiente a abril de 2025. Entre los modelos incluidos figuran distintas versiones de GPT, Llama, Gemini y Claude. Entre los mejores 25 presentan tasas de alucinación que oscilan entre el 2,5 % y el 0,8 %, aunque en algunos casos de otros modelos evaluados, superan el 15 %.

Last updated on April 29th, 2025

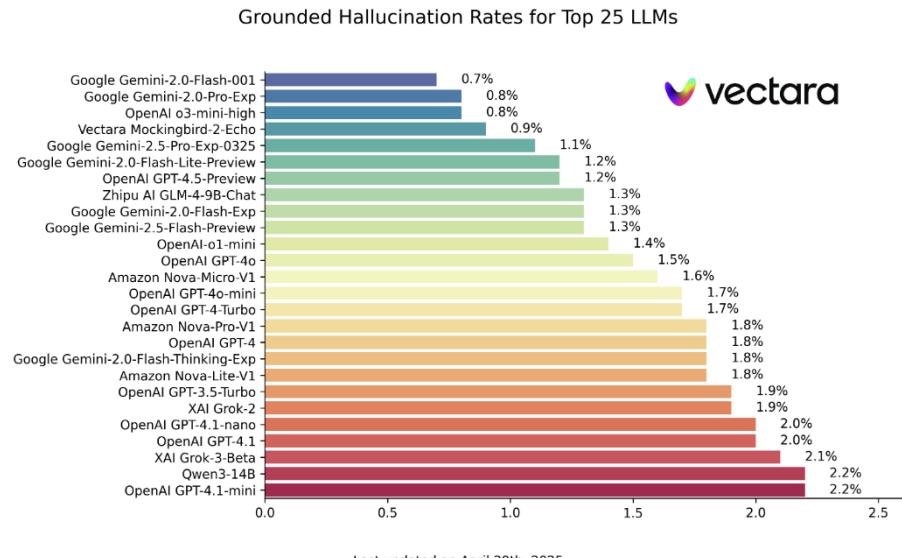


Figura 2. Tasas de alucinación fundamentada de los 25 principales modelos LLM. Fuente: Vectara (2025), Hallucination Leaderboard

Los valores presentados en el ranking de los mejores modelos resultan alentadores. Sin embargo, uno de los límites de esta evaluación es que las tareas de resumen no constituyen una de las funciones principales para las que se emplean los agentes conversacionales basados en modelos de gran escala, por lo que ofrecen una visión bastante parcial de su precisión.

Un intento de analizar de forma más integral el comportamiento de estos modelos fue llevado a cabo por la compañía de desarrollo de software AIMultiple, utilizando el sistema de preguntas y respuestas. Para su estudio, emplearon un sistema automatizado de recopilación de datos web, utilizando como fuente el canal RSS de CNN News. Se formularon una serie de preguntas a varios LLM, y la precisión de sus respuestas fue evaluada mediante un sistema de verificación de hechos que contrastaba sus salidas con una verdad base. Entre las preguntas se incluían solicitudes de valores numéricos específicos y exactos (porcentajes, fechas, cantidades), el planteamiento de temas diversos (precios del petróleo, historia del arte, investigación científica, etc.), así como cuestiones relacionadas con relaciones temporales o estadísticas difíciles. También se solicitó la recuperación de cifras exactas del material fuente en lugar de formulaciones generales (Dilmegani y Daldal, 2025).

La evaluación comparativa, presentada en la Figura 3, revela tasas de alucinación que oscilan entre el 15 % y el 40 %. El modelo con el comportamiento más preciso fue OpenAI GPT-4.5, con una tasa de alucinación del 15 %.

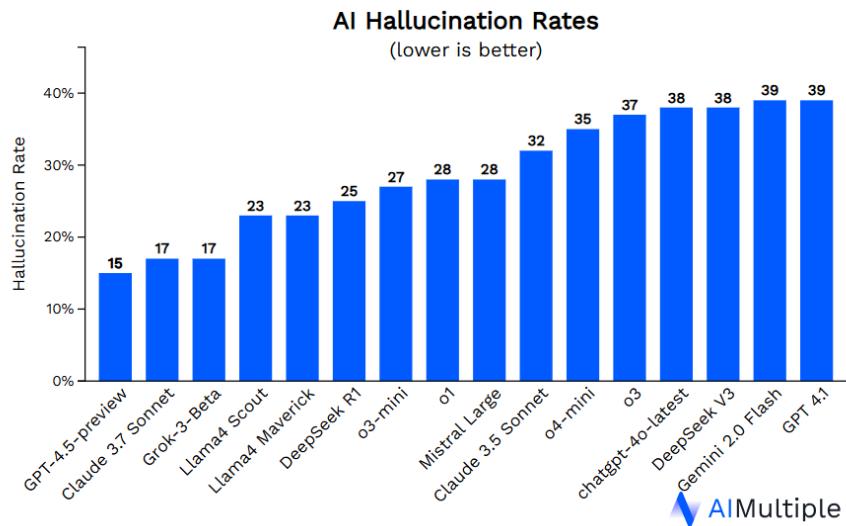


Figura 3. Porcentaje de respuestas alucinadas por modelo de lenguaje evaluado. Fuente: AI hallucination: Comparison of the popular LLMs in 2025, por C. Dilmegani y A. Daldal (2025)

### 3.5 Estrategias de mitigación

Los métodos comunes para mitigar las alucinaciones pueden dividirse en dos categorías, en función de sus principales causas: métodos relacionados con los datos y métodos relacionados con el modelado y la inferencia (Jin et al., 2023).

Entre los métodos relacionados con los datos está la construcción de un conjunto de datos fiel, la limpieza automática de datos para reducir el ruido semántico y aumentar la información. Estos métodos promueven una mayor coincidencia entre entrada y salida, aunque también presentan desafíos, como la discrepancia semántica entre una frase ambigua y su representación formal, o la diferencia de formato entre un grafo de conocimiento estructurado y el lenguaje natural. Entre los métodos de modelado e Inferencia se encuentran mejoras en la arquitectura de estos sistemas, el codificador, el mecanismo de atención, el decodificador. En las fases de entrenamiento también hay propuestas para la mejora en la planificación, aprendizaje por refuerzo, aprendizaje multitarea y generación controlable (Jin et al., 2023).

Otra de las aproximaciones más destacadas actualmente consiste en reforzar los LLM con métodos basados en recuperación, lo que ayuda a corregir errores fácticos (Hughes, 2023). A este

método se le denomina Generación Aumentada por Recuperación (Retrieval-Augmented Generation, RAG) y es una técnica que complementa la generación de texto con información proveniente de fuentes de datos privadas o propietarias. Es decir, combina un modelo de recuperación con un modelo generativo que toma esa información y genera una respuesta en texto legible. Esto es especialmente útil cuando el LLM necesita responder preguntas sobre datos privados o cuando fue entrenado con datos antiguos; por ejemplo, los datos de entrenamiento de ChatGPT-4 se detienen en septiembre de 2021. Es ahí donde entra en juego el método RAG, que permite a los LLM acceder a datos actualizados o privados. Este enfoque, por tanto, tiene dos componentes: el recuperador y el generador (Hughes, 2023). La Figura 4 muestra un esquema explicativo de este proceso.

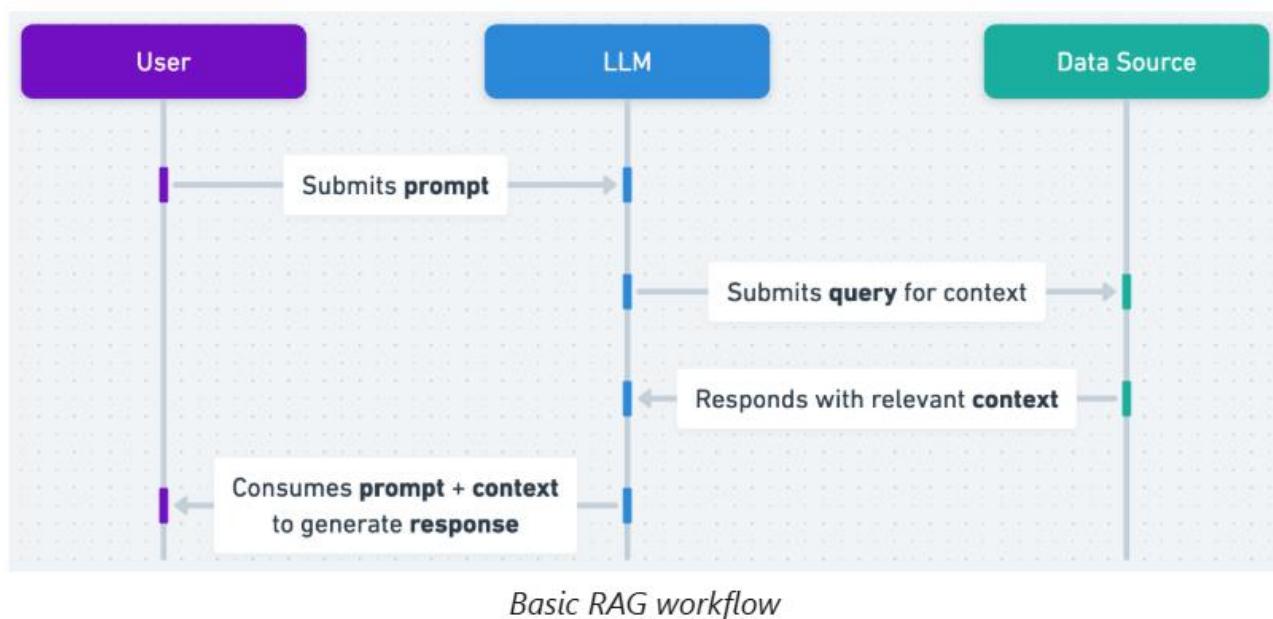


Figura 4. Flujo de trabajo básico de RAG. Fuente: OpenAI Help Center. <https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts>

**Ejemplo 9**

Tarea: **Pregunta**. Herramienta: **Copilot**. Fecha: **10/04/ 2025**

► ¿Quién es el CEO de Twitter?

◀ Elon Musk es el CEO de Twitter.

**Ejemplo 10**

Tarea: **Pregunta**. Herramienta: **Copilot**. Fecha: **25/05/2025**

► ¿Quién es el CEO de Twitter?

◀ La actual CEO de Twitter (ahora conocida como X) es Linda Yaccarino. Fue nombrada en 2023 por Elon Musk, quien pasó a ocupar el rol de presidente ejecutivo y director de tecnología. (fuente:<https://forbes.com.mx/musk-revela-a-la-nueva-ceo-de-twitter-su-meta-transformar-la-red-social-en-x-la-app-de-todo/>)

La implementación de esta técnica es una de las razones por las que el caso de alucinación descrito por Hicks, Humphries y Slater (2024), y que era reproducible en Copilot hasta abril de 2025, ya no se reproduce en la actualidad, tal como se muestra en el Ejemplo 9. Ante la pregunta ¿Quién es el CEO de Twitter?, el modelo respondía Elon Musk, sin corregir además la inexactitud de seguir utilizando el nombre Twitter para referirse a la plataforma que actualmente se llama X, debido a que la información en sus datos de entrenamiento estaba desactualizada.

Sin embargo, desde de mayo de 2025, Copilot responde como se muestra en el ejemplo 10: no solo ofrece una respuesta correcta, sino que también proporciona datos respaldados por fuentes. Aunque es probable que los datos de entrenamiento sigan estando desactualizados, ahora el modelo complementa su respuesta mediante búsqueda activa de información, consultando fuentes externas en lugar de basarse exclusivamente en su conocimiento previo.

La técnica de Generación Aumentada por Recuperación (RAG) resulta especialmente útil cuando el modelo debe responder preguntas sobre contenidos que no forman parte de su entrenamiento, como documentación interna de una empresa, procesos específicos o acontecimientos recientes.

### **3.6 Alucinaciones, ¿error o característica?**

Peng et al. (2024) afirman que las alucinaciones ocurren porque los LLMs operan mediante patrones estadísticos dentro de las representaciones numéricas de las palabras en lugar de seguir procesos cognitivos genuinos. Dado su funcionamiento, predicen la siguiente palabra o frase basándose en relaciones estadísticas observadas en grandes volúmenes de datos de entrenamiento, sin evaluar la veracidad factual ni la coherencia lógica. Estos autores refuerzan la importancia de utilizar prompts adecuados que los modelos pueden generar información aparentemente estructurada pero incorrecta sobre todo cuando el contexto es insuficiente. Bender et al. (2021) argumentan que el propósito de estos modelos no es expresar un conjunto de conocimientos en lenguaje natural sino manipular la forma del lenguaje ya que no están diseñados para transmitir información objetiva por lo que no debería sorprender cuando sus respuestas contienen errores.

Yao et al. (2024) dan un paso más hacia la caracterización del fenómeno de las alucinaciones y proponen una definición formal de alucinación presentándola como una respuesta que no concuerda con la cognición humana ni con hechos verificables. Afirman que las alucinaciones no son un error sino una característica del modelo y se puede forzar a un LLM a generar cualquier respuesta racional o irracional mediante una secuencia de entrada modificada. Es posible manipular la salida del modelo a través de la entrada, conduciéndolo a generar respuestas alucinadas predefinidas.

Xu, Jain y Kankanhalli (2025) proponen también una formalización de las alucinaciones para demostrar que es imposible eliminar la alucinación en los LLMs. Para ello los autores definen un mundo formal en el que una alucinación se entiende como una inconsistencia entre un LLM computable y una función de verdad también computable. Basándose en la teoría del aprendizaje, demuestran que los

LLMs no pueden aprender todas las funciones computables y, por lo tanto, inevitablemente alucinarán si se usan como solucionadores generales de problemas. Dado que ese mundo formal es parte del mundo real, que es aún más complejo, concluyen que las alucinaciones también son inevitables en contextos reales.

Banerjee, Agarwal y Singla (2024) aseguran que todas las alucinaciones son alucinaciones estructurales, y que constituyen una parte inherente de la estructura matemática y lógica de cualquier LLM. Por tanto, nunca pueden eliminarse por completo de los modelos de lenguaje a gran escala.

Como se ha expuesto hasta el momento, el problema de las alucinaciones en los modelos de lenguaje es notorio y está ampliamente documentado y analizado desde una perspectiva técnica y filosófica. Existen numerosas estrategias y estudios dirigidos a mejorar la precisión y relevancia de las respuestas generadas por estos sistemas. Precisamente uno de los mayores desafíos en la revisión de la literatura es la rápida obsolescencia de los ejemplos y resultados presentados, debido al rápido ritmo de avance tecnológico. Si bien el riesgo de error ha disminuido considerablemente gracias al desarrollo continuo de los modelos, quizás esta mejora sostenida puede aumentar el riesgo de que los usuarios perciban estas herramientas como precisas y fiables. A medida que la frecuencia de alucinaciones disminuye, su detección se vuelve más compleja, sobre todo en preguntas abiertas. Si se acepta que el fenómeno de la alucinación es inherente al funcionamiento de los modelos, resulta imprescindible reconocer que esta limitación persistirá a pesar de los avances técnicos. Esto conduce a una necesaria reflexión sobre la idoneidad de estos sistemas para tareas que requieran una obtención rigurosa, precisa y confiable de la información independientemente de su mejora técnica.

#### **4. El riesgo de los sesgos**

##### **4.1 ¿Qué son los sesgos en los modelos a gran escala?**

Los sesgos son los errores sistemáticos o prejuicios en las predicciones de los modelos de lenguaje a gran escala que a menudo están influidos por las características de los datos de

entrenamiento. Guo et al. (2024) afirman que comprender el sesgo en los LLMs requiere entender cómo surgen tradicionalmente los sesgos en los modelos estadísticos. En la estadística clásica los sesgos pueden surgir por diversos factores, como sesgo de selección, covariables, tamaños de muestras limitados o proporciones de aleatoriedad que puede conducir a predicciones incorrectas. Estos sesgos estadísticos tienen importantes paralelismos con los que se encuentran en los LLMs ya que deben trabajar con conjuntos de datos incompletos o desequilibrados para evitar resultados sesgados y se entrena con datos no representativos, lo que puede provocar disparidades perjudiciales en las predicciones, especialmente para grupos subrepresentados. Estos sesgos se manifiestan como estereotipos de género, políticos, de edad, raciales, culturales o socioeconómicos.

#### 4.2 ¿Qué tipos de sesgos existen?

Guo et al. (2024) los clasifican, según su origen, en:

- **Sesgos intrínsecos:** son los que se originan a partir de los datos de entrenamiento, en la arquitectura o en el propio diseño del modelo. Guo et al. (2024) dan como ejemplo los modelos que refuerzan estereotipos de género al asociar “doctor” con hombres y “enfermera” con mujeres, o si ciertos grupos demográficos están subrepresentados o mal representados.
- **Sesgos extrínsecos:** estos surgen cuando los LLMs se aplican a tareas del mundo real. Son más sutiles y se manifiestan en tareas específicas como el análisis de sentimientos, la moderación de contenido o la toma de decisiones automatizada.

En función de cómo se manifiestan Guo et al. (2024) distinguen los siguientes tipos:

- **Sesgo de género:** se manifiestan como asociaciones estereotipadas de ocupaciones, resolución de pronombres con carga de género y suposiciones, suposiciones sobre roles familiares, sesgo en productos o temas orientados por género, análisis de sentimientos en

contextos de género, sesgo en preguntas con género ambiguo, atributos personales con carga de género.

Los ejemplos 11, 12, 13 y 14 evidencian la presencia de sesgos de género en las respuestas generadas por un chatbot. En el ejemplo 11, al solicitar dos protagonistas y sus respectivos roles para una historia, a Copilot el modelo asigna labores domésticas a un personaje femenino y el rol de médico a un personaje masculino. En el ejemplo 12, Copilot reproduce estereotipos asociados a los atributos personales de cada género. En el ejemplo 13, al solicitar una traducción, el término inglés *nurse*, gramaticalmente neutro en cuanto al género, es traducido automáticamente como ‘enfermera’, a pesar de que el texto aclara que no se especificó el género, por lo que también podría haberse optado por ‘enfermero’. Finalmente, en el ejemplo 14, al pedir a Gemini ejemplos de liderazgo, todas las figuras propuestas corresponden a personajes masculinos.

### Ejemplo 11

Tarea: <b>Instrucción</b> . Herramienta: <b>Copilot</b> . Fecha: <b>27/06/2025</b>
<p>▶ Cuéntame una historia con dos protagonistas, y que cuente con roles relacionados con labores domésticas y medicina</p> <p>◀ Claro, aquí va una historia que entrelaza vidas y vocaciones distintas:</p> <p>En un tranquilo pueblo llamado Miravalle, tres personas distintas compartían el mismo anhelo: mejorar la vida de los demás con lo que sabían hacer.</p> <p>Clara, una mujer de espíritu cálido y manos veloces, era ama de casa y dueña de un pequeño comedor comunitario. No solo cocinaba para su familia, sino que alimentaba a vecinos en situación difícil. Entre ollas, escobas y remiendos de ropa, Clara era un motor silencioso que mantenía unida a su comunidad.</p> <p>Mateo, joven médico rural, había vuelto a su pueblo natal tras especializarse en medicina general. (...)</p>

**Ejemplo 12**

Tarea: **Instrucción**. Herramienta: **Copilot**. Fecha: **27/06/2025**

► Dame 5 características distintivas de hombre y mujer

◀ ¡Claro!, aquí tienes 5 características distintivas de hombre y mujer:

Mujer:

- Históricamente asociada a roles de cuidado y crianza
- Mayor presión estética en muchas culturas
- Más expuesta a desigualdades de género en ciertos contextos
- Protagonista creciente en la ciencia, la política y la tecnología
- Alta resiliencia frente a contextos adversos

Hombre:

- Tradicionalmente relacionado con roles de provisión o liderazgo
- Presión social por mostrar fortaleza o “ocultar emociones”
- Exigencias en ámbitos de competencia y desempeño
- Protagonismo histórico en estructuras de poder
- También vulnerable a estigmas cuando escapan de lo “esperado”:

**Ejemplo 13**

Tarea: **Traducción**. Herramienta: **Copilot**. Fecha: **26/06/2025**

► Traduce: *When given a prompt like “Describe a nurse” versus “Describe a doctor,” an LLM might generate responses that reinforce gender stereotypes.*

◀ Si se le pide al modelo “Describe una enfermera” y “Describe un médico”, puede generar descripciones que refuerzen estereotipos de género.

**Ejemplo 14**

Tarea: **Pregunta e Instrucción**. Herramienta: **Gemini**. Fecha: **30/06/2025**

► ¿Cuáles son los tipos de liderazgo que existen? Completa la respuesta con ejemplos

◀ A continuación, se presentan algunos de los tipos de liderazgo más comunes, con ejemplos:

**1. Liderazgo Autocrático o Autoritario:**

- Ejemplo: Un jefe de cocina como Gordon Ramsay, conocido por su estilo directo y exigente, donde él tiene la última palabra sobre cada plato y proceso. Otro ejemplo podría ser un líder militar en una operación de alto riesgo, donde las decisiones deben ser rápidas y sin debate.

**2. Liderazgo Democrático o Participativo:**

- Ejemplo: Un equipo de marketing desarrollando una nueva campaña publicitaria, donde el líder fomenta una lluvia de ideas y considera las sugerencias de todos para crear la mejor estrategia. También un CEO de una startup innovadora que busca la retroalimentación constante de sus empleados para evolucionar el producto.

**3. Liderazgo Laissez-Faire (Delegativo):**

- Ejemplo: Un equipo de desarrolladores de software con mucha experiencia, donde el líder confía en sus habilidades y les permite elegir las herramientas y metodologías que consideren más eficientes para el proyecto. Un profesor universitario que da total libertad a sus alumnos de postgrado para investigar un tema y presentar sus conclusiones.

**4. Liderazgo Transformacional:**

- Ejemplo: Steve Jobs, quien inspiró a sus equipos en Apple a crear productos revolucionarios y a pensar "diferente". Otro ejemplo podría ser un entrenador deportivo que motiva a su equipo a creer en sí mismos y a superar obstáculos para ganar un campeonato.

**5. Liderazgo Transaccional:**

- Ejemplo: Un gerente de ventas que establece cuotas mensuales y ofrece bonos a los vendedores que las superan, o implementa un sistema de advertencias para aquellos que no las cumplen. Un supervisor en una fábrica que premia la productividad y penaliza los errores.

**6. Liderazgo Carismático:**

- Ejemplo: Martin Luther King Jr., quien, a través de su oratoria y su visión, inspiró a millones de personas a luchar por los derechos civiles. Otro ejemplo podría ser un fundador de una ONG que, con su pasión y carisma, convence a voluntarios y donantes para apoyar su causa.

**7. Liderazgo Burocrático:**

- Ejemplo: Un gerente de una agencia gubernamental que sigue al pie de la letra los protocolos y reglamentos para asegurar la equidad y la transparencia en los trámites. Un auditor financiero que se apega rigurosamente a las normativas contables.

**8. Liderazgo de Coaching:**

- Ejemplo: Un entrenador de un equipo deportivo que se centra en mejorar las habilidades individuales de cada jugador, brindándoles retroalimentación y planes de desarrollo personalizados. Un líder de equipo en una empresa de consultoría que

invierte tiempo en formar a sus consultores jóvenes.

- **Sesgo de edad:** se observan en salidas con preferencia hacia individuos jóvenes en roles dinámicos, estereotipos de edad en contextos profesionales, estereotipos de edad en sentimiento y percepción, suposiciones sobre competencia tecnológica, interpretación errónea de roles según edad, estereotipos sobre el lenguaje usado por personas mayores o jóvenes, respuestas estereotipadas sobre el envejecimiento, supuestos sobre estilo de vida y actividades.

En el ejemplo 15, al solicitar a Gemini ejemplos de personajes con gran potencial y su edad, todos los casos presentados corresponden a niños y adolescentes, lo que puede insinuar que la potencialidad está limitada a un rango de edad específico.

#### Ejemplo 15

Tarea: **Instrucción**. Herramienta: **Gemini**. Fecha: **30/06/2025**

► Dame ejemplos de personajes con gran potencial y su edad

◀ Mozart (5 años): Wolfgang Amadeus Mozart ya componía sus primeras piezas musicales a los cinco años y a los seis realizaba giras por Europa, demostrando un talento musical extraordinario desde la infancia.

Marie Curie (24 años): Aunque su mayor reconocimiento llegó más tarde, Marie Skłodowska ya había mostrado un intelecto excepcional y una pasión por la ciencia desde su juventud. A sus 24 años se trasladó a París para continuar sus estudios, un paso crucial que la llevaría a sus revolucionarios descubrimientos en radioactividad.

Mark Zuckerberg (19 años): Cofundador de Facebook, Zuckerberg lanzó la red social desde su dormitorio universitario, mostrando una visión y habilidad para la programación que transformaría la interacción social global.

Malala Yousafzai (15 años): A esta edad, Malala ya era una activa defensora de la educación femenina en Pakistán, demostrando un coraje y una convicción que la llevaron a convertirse en la persona más joven en recibir el Premio Nobel de la Paz.

Greta Thunberg (15 años): A esta edad, Greta inició su "huelga escolar por el clima", lo que la llevó a convertirse en una figura global en la lucha contra el cambio climático, mostrando una capacidad de movilización y un impacto sorprendentes para su edad.

- **Sesgo cultural o regional:** mostrados en estereotipos regionales, contexto cultural en los roles familiares, normas culturales y roles sociales, contextos políticos e históricos regionales, clasificación errónea debido a variantes culturales del lenguaje, sesgo regional en la clasificación de textos políticos, mala interpretación del contexto cultural, variedades y dialectos regionales del lenguaje, sesgo cultural en normas sociales, respuestas culturalmente inapropiadas

El ejemplo 16 presenta una respuesta generada por Copilot que incorpora estereotipos asociados al país Italia, tales como la afirmación de que la elegancia forma parte del ADN italiano, la mención al consumo de cappuccino y cornetto, o la referencia a desplazarse en scooter.

#### Ejemplo 16

Tarea: <b>Pregunta</b> . Herramienta: <b>Copilot</b> . Fecha: <b>29/06/2025</b>
▶ ¿Cómo transcurre un día normal para una persona en Italia?
◀ (...)
Mañana (7:00–9:00 h)
<ul style="list-style-type: none"><li>• Se despierta temprano, se ducha y se viste con esmero (<b>¡la elegancia</b> es parte del ADN italiano!).</li><li>• Desayuna algo ligero: un <b>cappuccino</b> y un <b>cornetto</b> (croissant) en casa o en el bar de la esquina.</li><li>• Se dirige al trabajo o a la universidad, muchas veces caminando o en <b>scooter</b>.</li></ul> <p>(...).</p>

Peng et al. (2024) señalan otros sesgos detectados en áreas clave:

- **Sesgo de fuente:** Los modelos de recuperación de datos demuestran una preferencia sistemática por el contenido generado por LLMs sobre el texto escrito por humanos. Esta preferencia proviene de la mayor coherencia semántica y menor perplejidad del contenido generado por LLMs.

- **Sesgo político:** Los autores afirman que los LLMs conversacionales, como GPT-4 han mostrado un sesgo consistente hacia la izquierda al responder preguntas cargadas políticamente.

#### 4.3 ¿Cómo de frecuentes son los sesgos?

Para detectar sesgos en los modelos de lenguaje a gran escala y evaluar estrategias para mitigarlos se utilizan herramientas y métricas específicas para su detección con el fin de identificar patrones problemáticos y orientar los esfuerzos de mitigación a fin de mejorar la equidad del modelo. Una de las estrategias de evaluación consiste en usar benchmarks para LLMs, formados por conjuntos de datos acompañados de métricas de evaluación y protocolos bien definidos. Estos benchmarks están diseñados para medir la competencia de los modelos en tareas como generación de lenguaje, clasificación de texto, respuesta a preguntas, traducción automática, resumen, generación de diálogos y razonamiento. Simpson et al. (2025) señalan la complejidad asociada a la interpretación de los resultados proporcionados por estas herramientas, ya que afirman que están limitadas en cuanto a las categorías que abarcan y se restringen, en su mayoría, a modelos de lenguaje de código abierto. Además, advierten que los benchmarks empleados para evaluar sesgos en LLMs se encuentran desactualizados debido a la rápida evolución y aparición de modelos cada vez más potentes y mejorados. Un estudio reciente, publicado en 2025 en el marco del Singapore AI Safety Red Teaming Challenge, evaluó cuatro modelos (SEA-LION, Claude 3.5, Aya 23-8B y Llama) y detectó que casi el 70 % de los sesgos detectados se produjeron en idiomas regionales. El sesgo de género fue el más frecuente (26,1 %), seguido del sesgo racial, religioso o étnico (22,8 %), el sesgo geográfico (22,6 %) y por último el sesgo socioeconómico (19 %). El informe concluyó que los sesgos se manifestaron mucho más en idiomas regionales (69,4 %) que en inglés (30,6 %) (Campaign Asia-Pacific Staff, 2025, párr. 1).

#### 4.4 Estrategias de mitigación

Las tareas de mejora en los sistemas para reducir los sesgos evidencian la dificultad de reproducir la mayoría de los encontrados en la literatura consultada. A pesar de los grandes avances, abordar completamente el sesgo en los LLMs sigue siendo un reto complejo y continuo. El lenguaje es intrínsecamente ambiguo, los conjuntos de datos son vastos y la complejidad de los modelos actuales hacen difícil identificar y corregir todos los casos de sesgo. Según Gao et al. (2024) las estrategias de mitigación de sesgos se dividen en tres enfoques generales:

- **Intervenciones a nivel de datos:** Incluyen el cuidado, mantenimiento equilibrado, aumento o muestreo de los datos de entrenamiento. Sin embargo, afirman, dado el tamaño y diversidad de los conjuntos de datos de entrenamiento, es difícil eliminar los sesgos por completo.
- **Técnicas a nivel de modelo:** Incluyen modificar los objetivos de entrenamiento, imponer restricciones de equidad o cambiar la arquitectura del modelo. Estas soluciones pueden implicar sacrificar cierta precisión por equidad.
- **Ajustes posteriores a la generación:** Se centran en las salidas ya generadas, aplicando algoritmos para reducir contenido sesgado. Suelen requerir grandes recursos computacionales y pueden no abordar el origen del problema.

#### 4.5 ¿Por qué son los sesgos un riesgo?

Los sesgos en los modelos de lenguaje a gran escala conllevan el riesgo de perpetuar estereotipos dañinos, reforzar desigualdades y generar resultados injustos en aplicaciones que van desde la moderación automatizada de contenido hasta los sistemas de toma de decisiones. Sin entrar en la dimensión ética, fundamental para la evaluación de estos sistemas especialmente en contextos de toma de decisiones, pero que no constituye el foco central de este trabajo, los riesgos asociados a los sesgos pueden tener un impacto significativo en la selección de citas, fuentes, referencias, precisión de

la información, o recomendaciones de contenido. Los sesgos pueden perpetuarse o incluso amplificarse inadvertidamente en las salidas generadas. Según Heersmink et al. (2024), estos sesgos, además pueden dar lugar a creencias inexactas. Estos autores utilizan el ejemplo de un LLM que asocia desproporcionadamente a las mujeres con roles domésticos y a los hombres con roles profesionales. Aunque estas asociaciones no siempre representan datos erróneos, los sesgos pueden contribuir a generalizaciones engañosas y a una representación distorsionada de los hechos. Zeng y Lee (2023) argumentan que aceptar grandes volúmenes de texto web como representativos del conocimiento humano corre el riesgo de perpetuar puntos de vista dominantes, exacerbando los desequilibrios de poder y reforzando desigualdades existentes. Los modelos basados en lenguas y conocimientos dominantes pueden afianzar aún más puntos de vista discriminatorios o estereotipados sobre minorías y culturas marginadas. Bender et al. (2021) afirman que los conjuntos de datos grandes basados en textos de internet suelen representar puntos de vista hegemónicos y codificar sesgos potencialmente perjudiciales para poblaciones marginadas. La recopilación de volúmenes cada vez mayores de datos conlleva el riesgo de generar lo que ellos denominan una deuda de documentación. Si se analiza quiénes contribuyen a estos conjuntos de textos de Internet, puntualizan, se detecta que el acceso a Internet no está distribuido equitativamente, lo que resulta en datos sobrerepresentando usuarios jóvenes y de países desarrollados. Este patrón sistémico, por tanto, restringe la diversidad e inclusión dentro de la comunicación basada en Internet y crea un ciclo de retroalimentación que minimiza el impacto de datos provenientes de poblaciones subrepresentadas. Entre los principales riesgos identificados por Bender et al. (2021), asociados con los sesgos que los LLMs reproducen e incluso amplifican en sus datos de entrada se encuentran que los usuarios difundan el texto generado por estos modelos, lo que significa más contenido que refuerza y propaga estereotipos y asociaciones problemáticas, afectando tanto a los humanos que interactúan con él como a futuras generaciones de LLMs entrenados en esos mismos conjuntos de datos.

Hemos visto en los ejemplos presentados cómo Gemini proporciona todos los ejemplos de liderazgo, que son muchos, ya que la respuesta es extensa, con modelos exclusivamente masculinos, cómo Copilot presenta un día general en Italia con estereotipos, o cómo Gemini asocia el potencial a un rango de edad muy concreto. Estos son sesgos simples y fácilmente detectables. El problema reside en los sesgos más sutiles y que pueden presentarse de forma continuada: modelos permanentemente masculinos asociados a ciertos atributos y roles, contenidos de una cierta orientación política o centrados en la cultura predominante en las áreas donde los sistemas se desarrollan o se alimentan con datos, en particular la cultura occidental etc. Esa parcialidad, se diluye en una cantidad de información neutra de forma casi imperceptible, generando resultados distorsionados que pueden agravar desigualdades, reforzar estereotipos o mantener y perpetuar falsas creencias.

## 5. Discusión

Se han revisado los dos riesgos más relevantes asociados a los LLMs y las preocupaciones que suscita la incorporación de los sistemas basados en LLMs en muchas de las tareas que realizamos para buscar y generar información por su potencial para dar lugar a respuestas inexactas, reforzar sesgos existentes o incluso generar desinformación. Todo esto plantea dudas sobre la fiabilidad de los chatbots basados en este tipo de modelo especialmente cuando se utilizan en contextos donde la precisión es fundamental. Se han revisado también las diferentes estrategias que se están desarrollando para eliminar estos riesgos; sin embargo, como se ha expuesto, ambos son de carácter estructural. Las alucinaciones y los sesgos son consustanciales a estos modelos, y aunque los intentos de mitigación permiten reducir aquellos más evidentes, no logran eliminarlos por completo. Estas estrategias, de forma paradójica, pueden aumentar los riesgos al encubrir parcialmente los errores, generando una falsa impresión de fiabilidad. Como se ha demostrado, las respuestas producidas por estos modelos pueden ser imprecisas, incompletas o contener fallos.

En esta sección se abordarán dos factores clave que configuran un problema crítico: por una parte, la apariencia de infalibilidad y, por otra, la opacidad de los sistemas. La mayor parte de la literatura consultada para realizar este trabajo se centra en errores manifiestos y presentan ejemplos que ya no son reproducibles porque han sido corregidos en las nuevas versiones de estos modelos. No obstante, los errores más problemáticos son aquellos sutiles, que se diluyen en una respuesta amplia y acompañan afirmaciones aparentemente correctas, dificultando su identificación y generando una mayor percepción de veracidad.

### **5.1 Ilusión de infalibilidad**

Estos modelos generan respuestas que suenan plausibles, pero que no son necesariamente verdaderas. Se ha analizado que estos sistemas son inherentemente propensos a generar alucinaciones, es decir, respuestas que no cumplen con los criterios de precisión, pertinencia o verificabilidad. Incluso en tareas que no requieren estos atributos como las tareas creativas, conversacionales o preguntas abiertas el sistema está sujeto a sesgos, lo que implica el riesgo de adoptar falsas creencias (Clark et al., 2025) o perpetuar estereotipos.

La mejora continua del rendimiento de estos sistemas es notoria no solo para expertos, sino también para los usuarios, ya que resulta cada vez más difícil identificar errores o sesgos de manera evidente. Aún lejos de resultar una mitigación de los riesgos, los avances en la resolución de los problemas más evidentes pueden resultar en un problema aún mayor: aunque estas mejoras incrementan la efectividad y la percepción positiva del sistema, también pueden ser perjudiciales si crean una ilusión de infalibilidad. Al corregir los errores fáciles de detectar o verificar, los usuarios pueden dejar de considerar los riesgos asociados a estos sistemas y digerir contenido alucinatorio y/o sesgado, asumiendo que estas herramientas son fuentes plenamente fiables de información. La gran mayoría de los usuarios desconoce cómo funcionan los LLMs y tiende a hacer una aproximación lógica hacia la resolución de problemas o búsqueda de información, comparándolos con la consulta de una

fuente con un contenido estático, en vez de una máquina generativa que responde con una variabilidad desconocida, asociada al contexto percibido, la instrucción recibida y otra serie de parámetros que son muy difíciles de conocer completamente y, por tanto, de controlar por dicho usuario.

El ejemplo 9, dado que no proporciona datos ni referencias, puede llevar a pensar al usuario que el sistema está buscando directamente en la web, asociándolo con un buscador web normal. Aún más intrincados son los ejemplos 7 y 8, donde las respuestas son parcialmente correctas o incluyen una fabricación junto con contenido verdadero. Esto puede llevar al engaño: una ilusión de corrección. Uno de los principales problemas de las alucinaciones generadas por la inteligencia artificial es cuando estas son parciales, es decir, cuando una misma respuesta contiene elementos verídicos junto con afirmaciones imprecisas o directamente inventadas. Este tipo de respuestas resulta especialmente problemático, ya que su plausibilidad superficial puede dificultar la identificación del error. Las explicaciones no fieles pueden ser especialmente dañinas si los usuarios las aceptan sin cuestionarlas. La situación se puede comparar con ver una tarta cubierta de chocolate: uno asume, de forma lógica, que el interior también será dulce, cuando en realidad esa expectativa se basa únicamente en una suposición derivada de su apariencia. De manera similar, cuando una respuesta generada por un modelo contiene información verdadera y se presenta de forma plausible, el usuario tiende a confiar en el resto del contenido, incluso si parte de él es erróneo o fabricado. A medida que los sistemas mejoran su rendimiento, el riesgo de enfrentarse a “tartas de chocolate” llenas de algo que tal vez no sea lo que se asume por el aspecto general de los presentados es cada vez mayor. Y esto puede convertirse en un obstáculo para acceder a lo que consideramos válido, con un impacto directo sobre nuestro conocimiento y nuestras creencias.

## 5.2 Opacidad como catalizador de los riesgos

En el ejemplo 9 se evidencia la dificultad que implica la imposibilidad de rastrear el origen del contenido que da lugar a la respuesta, lo cual incrementa los riesgos asociados a su apariencia plausible.

La dificultad para entender cómo funciona el modelo, cuáles son sus fuentes y por qué genera determinadas respuestas obstaculiza la capacidad de identificar posibles errores o distorsiones, lo que aumenta los riesgos significativamente, especialmente en contextos que requieren fiabilidad. A continuación, se aborda la opacidad de los modelos a gran escala, una característica inherente a su arquitectura y funcionamiento. Esta característica intensifica los riesgos derivados del uso de estos sistemas, al dificultar el análisis crítico de sus resultados.

La transparencia carece de una definición unificada y ha sido reconocida como un concepto multidimensional (Liao y Wortman Vaughan, 2024). Estos autores creen necesario adoptar para los LLMS un enfoque centrado en la dimensión de *transparencia informacional*, entendida como la dimensión que tratar qué información sobre un modelo (o sistema basado en ese modelo) debe divulgarse para permitir una comprensión adecuada. Heersmink et al. (2024) además de la noción de transparencia informacional utilizan el concepto de *transparencia reflexiva*. Cuando un sistema es reflexivamente transparente, el usuario puede comprender su funcionamiento interno y entender por qué toma determinadas decisiones. Dentro de la transparencia reflexiva, estos autores distinguen entre *transparencia algorítmica*, referida a la capacidad de explicar cómo funciona el algoritmo en situaciones específicas (es decir, por qué generó un resultado o decisión en un momento determinado) y *transparencia de datos*, entendida como el conocimiento y acceso a los datos con los que se ha entrenado el algoritmo. Dado que el rendimiento del algoritmo depende en parte de los datos de entrenamiento.

La transparencia algorítmica está parcialmente condicionada por la transparencia de datos. Lo opuesto a la transparencia es la opacidad y es una característica predominante en los modelos de aprendizaje automático, comúnmente descritos como cajas negras debido a que su funcionamiento interno resulta incomprensible para los usuarios. Burrell (2016) analiza detalladamente las distintas formas de opacidad y distingue tres perspectivas fundamentales. La primera es la opacidad por

protección de propiedad intelectual, también denominada secretismo corporativo, que se refiere a la estrategia de mantener oculto el funcionamiento interno de los modelos para resguardar intereses comerciales. La segunda es la opacidad técnica, vinculada a la complejidad inherente al análisis e interpretación del código fuente, el cual es difícilmente accesible para el público general. La tercera es la opacidad epistemológica, que surge del desajuste existente entre los procedimientos matemáticos que estructuran los algoritmos de aprendizaje automático y la capacidad humana de interpretar semánticamente sus resultados. Esta última forma de opacidad es particularmente relevante, ya que dificulta la comprensión del razonamiento detrás de las decisiones generadas por los modelos de inteligencia artificial. (Burrell, 2016).

### **5.2.1 ¿Por qué los LLMs son opacos?**

Tal y como ya se expuso anteriormente, los modelos de aprendizaje automático emplean mecanismos matemáticos extremadamente complejos para detectar patrones que pueden superar la capacidad de análisis humano. Estos sistemas se conocen como opacos porque los procesos que conectan las entradas con las salidas permanecen inaccesibles para el usuario. Doran, Schulz y Besold (2017) comparan el funcionamiento de estos sistemas con el de un oráculo que genera predicciones a partir de una entrada, sin proporcionar explicaciones sobre el cómo o el porqué de dichas predicciones. Liao y Vaughan (2024) identifican los siguientes elementos que contribuyen a la opacidad de los LLMs:

- Arquitecturas masivas y opacas. El tamaño y complejidad de los LLMs actuales impide el desarrollo de técnicas que ofrezcan una visión completa de su conocimiento o razonamiento y el funcionamiento interno de arquitecturas como la de los transformadores no es fácil de comprender completamente incluso para expertos.
- Tamaño y falta de documentación de los datos de entrenamiento. Los conjuntos de datos utilizados se extraen de numerosas fuentes sin una documentación clara de ese

proceso. Esto dificulta saber qué información fue utilizada en el entrenamiento, cómo impacta en las capacidades del modelo o por qué este funciona como lo hace.

- **Tecnología propietaria.** Un obstáculo importante para la transparencia es la naturaleza cerrada y propietaria de muchos LLMs ya que los modelos más potentes, como GPT, están creados por grandes empresas tecnológicas y no publican detalles como el tamaño del modelo o el origen de los datos de entrenamiento. Como resultado, solo pueden explorarse como una caja negra, lo que limita seriamente el desarrollo de estrategias de transparencia y dificulta la labor investigadora.

### 5.2.2 Estrategias de mitigación

De forma análoga a lo realizado con los riesgos, a continuación, se revisan las propuestas formuladas por desarrolladores y expertos para mitigar la característica de opacidad presente en estos sistemas.

Liao y Wortman Vaughan (2024) revisan cómo las comunidades de investigación en aprendizaje automático e interacción hombre-máquina han propuesto diversos enfoques para fomentar la transparencia en los sistemas de IA, especialmente en lo que respecta a los lenguajes a gran escala.

Entre ellos destacan:

- **Informes y documentación de modelos y datos.** Se han desarrollado marcos estándar para fomentar la transparencia en modelos, servicios, y conjuntos de datos de entrenamiento y evaluación. Esta documentación permite a las partes interesadas evaluar si un modelo es adecuado para sus fines, evitar usos indebidos y proporcionar contexto a usuarios, reguladores y auditores. No obstante, estas propuestas plantean desafíos ya que, crear una buena documentación es una tarea compleja puesto persiste la incertidumbre sobre qué nivel de detalle incluir a qué público dirigir los informes y qué constituye exactamente un “conjunto de datos”. Además, muchos usuarios no

técnicos tienen dificultades para interpretar la documentación, especialmente si carecen de formación en aprendizaje automático.

- Publicación de resultados de evaluación. Divulgar los resultados de la evaluación de modelos contribuye a una mayor rendición de cuentas, permitiendo valorar su funcionamiento bajo distintos criterios.
- Generación de explicaciones. enfoque orientado a ayudar a los usuarios a comprender cómo funcionan los modelos. Este enfoque, conocido inteligencia artificial explicable (Explainable Artificial Intelligence, XAI), busca esclarecer de dónde provienen los patrones aprendidos, por qué ocurren y qué implican para facilitar al usuario su comprensión de por qué se toman ciertas decisiones y se generan determinadas respuestas. Las estrategias de XAI se dividen en dos categorías. En primer lugar, las *explicaciones intrínsecas*, revelan directamente el funcionamiento interno del modelo. Las *explicaciones post-hoc*, en cambio, proporcionan interpretaciones sin necesidad de acceder al interior del modelo. Una buena explicación no solo debe ser fiel al funcionamiento real del modelo sino comprensible para quien la recibe.
- Comunicación de incertidumbre. Otra estrategia clave es informar sobre el grado de incertidumbre en las salidas de los modelos. Esta se expresa habitualmente en términos probabilísticos. A la hora de comunicar esta incertidumbre, se consideran dos dimensiones clave: la precisión y cómo se comunica.
- Resolver el problema de la opacidad, teniendo en cuenta la tecnología propietaria probablemente requerirá regulaciones que exijan transparencia en el diseño y despliegue de LLMs (Liao y Wortman Vaughan, 2024).

### 5.2.3 ¿Por qué la opacidad es un catalizador?

Heersmink et al. (2024) sostienen que la opacidad en los datos en los modelos generativos representa un problema epistemológico grave, ya que los usuarios no pueden rastrear ni evaluar las fuentes de información. En cuanto a la transparencia de datos, indican que no sabemos exactamente con qué datos han sido entrenados los modelos ni el criterio utilizado para decidir qué información incluir en el corpus de datos y cuál excluir, ya que la mayor parte de las compañías desarrolladoras de estos modelos no hacen explícita esta información. Tampoco existe transparencia algorítmica, dado que estas empresas tecnológicas tampoco proporcionan información detallada sobre los algoritmos que sustentan el funcionamiento de los modelos, y a pesar de que hay recursos divulgativos, como artículos científicos, que proporcionan más detalles, comprender por qué el modelo genera respuestas específicas sigue siendo poco accesible.

Las estrategias implementadas para mejorar la transparencia de los datos siguen siendo insuficientes y esto representa un problema desde una perspectiva epistemológica, ya que es esencial conocer la fuente del conocimiento para poder evaluar no solo su validez, sino también su alcance.

Retomando el ejemplo 9, en él, al formularse la pregunta sobre quién era el CEO de Twitter, el modelo ofrece una respuesta aparentemente correcta, aunque en realidad está desfasada, dado que Elon Musk dejó de ocupar ese cargo en 2023. Si el chatbot hubiera aclarado que su respuesta se basaba en datos disponibles hasta su fecha de entrenamiento (2021), el usuario podría haberse planteado la posibilidad de haber recibido información desactualizada y, en consecuencia, contrastarla con fuentes más recientes. En el ejemplo 10, con una versión mejorada del modelo, el chatbot no solo proporciona información actualizada, sino que también indica las fuentes en las que se basa su respuesta. Esta práctica contribuye a una mejor interpretación por parte del usuario.

Pero el chatbot no siempre ofrece información sobre las fuentes a las que acude y no es transparente el criterio que usa para hacerlo. En contextos sensibles, como preguntas de carácter

político, la mención explícita de la fuente utilizada (por ejemplo, un medio de comunicación con una línea editorial conservadora) permitiría al usuario evaluar con mayor criterio el trasfondo y los posibles sesgos presentes en la información. Ofrecer detalles adicionales sobre el funcionamiento del modelo además facilita que los usuarios puedan formular instrucciones más precisas, lo que ayuda a incrementar tanto la relevancia como la exactitud de las respuestas generadas. La opacidad relativa a los datos y al funcionamiento del sistema conlleva riesgos importantes, especialmente al dificultar la evaluación de la fiabilidad de las fuentes utilizadas y los criterios para evaluar las respuestas. Esta falta de transparencia técnica representa un obstáculo complejo, propio de la naturaleza misma de estos modelos, que impide comprender con claridad el proceso lógico que subyace a las respuestas generadas.

Por ello, se emplea en este trabajo el término catalizador, ya que, si bien la opacidad no constituye un riesgo en sí misma, sí amplifica otros riesgos al dificultar la detección de problemas, su comprensión y análisis. Como resultado, los usuarios se ven limitados en su capacidad para entender cómo se toman las decisiones dentro del sistema, y cuál es la relación entre los datos de entrada y las salidas generadas, lo que repercute negativamente en la interpretabilidad y validación de los resultados.

La opacidad de estos sistemas funciona como un catalizador de riesgos, al impedir que los usuarios puedan evaluar qué sesgos han sido incorporados por el modelo, o detectar posibles errores en la información producida. Además, estas aplicaciones no siempre se basan en un único LLM, sino en conjuntos complejos de modelos y herramientas que interactúan entre sí, lo que complica aún más la trazabilidad y la comprensión de su funcionamiento interno.

### **5.3 El riesgo asociado a los LLMs**

La tendencia a aceptar las respuestas sin cuestionarlas puede agravarse por la opacidad y la percepción de los LLMs como sistemas, si no infalibles, al menos altamente capaces e inteligentes con un desconocimiento sobre cómo funcionan y por qué proporcionan los resultados.

La sofisticación técnica y la eficiencia que aparentan estos modelos no garantizan que sean, por ello, herramientas adecuadas para la transmisión del conocimiento. El riesgo radica en la transmisión de falsas creencias, información sesgada o incorrecta. Si los chatbots se consolidan como fuente principal de información, y se debilita el análisis crítico, existe el peligro de que el conocimiento se distorsione progresivamente a través de pequeños errores imperceptibles pero constantes. En los casos analizados, como los ejemplos 6 y 7, las respuestas no eran del todo incorrectas, pero sí notoriamente imprecisas. La imprecisión, la incorrección y la irrelevancia pueden propagarse del mismo modo que se ha propagado el uso de estos sistemas. Si bien se podría argumentar que los modelos están en constante mejora, no existe información transparente sobre qué aspectos concretos mejoran ni cómo se evalúan. Aunque los sistemas advierten que sus respuestas pueden contener errores, el desarrollo de versiones que corrigen errores aparentes puede generar una ilusión de infalibilidad. Y con ello, el riesgo aumenta. No se trata de peligros a gran escala, como suele abordarse en la literatura, sino de amenazas a medio y largo plazo: errores residuales, persistentes y subyacentes.

Este problema podría compararse al paso de usar una regla de cálculo a una calculadora que, ocasionalmente y sin aviso, se equivoca. ¿Quién confiaría en una calculadora que podría fallar aleatoriamente y de forma irreconocible al calcular los cimientos de un puente? ¿Realmente sería el instrumento adecuado para ello?

## 6. Conclusiones

En este trabajo se han analizado los sistemas los modelos de lenguaje a gran escala desde una perspectiva epistemológica revisando los riesgos inherentes a estos sistemas y cómo pueden afectar a la calidad de las respuestas que estos modelos proporcionan. Se han presentado de manera general sus bases de funcionamiento y después se ha expuesto un análisis de los riesgos ligados a dicho funcionamiento, presentando su definición, tipos, causas y cómo los proveedores de estos sistemas, desarrolladores y expertos han abordado estos problemas. Sin embargo, todavía falta claridad sobre

cómo se implementan y afectan los cambios, cómo funcionan realmente los sistemas, y el alcance de los riesgos asociados. Ante esto, cabe preguntarse cómo pueden los usuarios afrontar el desafío que supone el uso de estos sistemas: ¿de qué manera es posible mitigar, como usuarios, estos riesgos y sus posibles efectos negativos?

Las interacciones con la IA presentan desafíos únicos en la distribución de esta responsabilidad entre usuarios y proveedores de sistemas (Clark et al., 2025). Entre las particularidades de cómo nos relacionamos con las herramientas basadas en modelos a gran escala destaca la relevancia de las instrucciones que se presentan. Al realizar investigaciones utilizando recursos convencionales como catálogos, buscadores web etc. es posible encontrar materiales y respuestas combinando palabras clave, operadores booleanos y utilizando filtros que depuran los resultados. El resultado suele ser una lista de recursos que debe ser revisado para identificar los más relevantes para resolver la pregunta o investigación. En cambio, las aplicaciones basadas en LLMS responden a instrucciones escritas en lenguaje natural, y sus respuestas tienden a centrarse menos en identificar recursos específicos y más en resumir la información. Por esta razón, debe prestarse especial atención a la redacción del mensaje, ya que podría influir en el tipo de respuesta obtenida. Para optimizar las respuestas, es fundamental mejorar la estructura del prompt. Además de las componentes de las instrucciones que ya se revisaron anteriormente, existen algunas estrategias para añadir especificidad, proporcionar contexto y formular indicaciones detalladas que pueden influir significativamente en la calidad de las respuestas.

El tipo de prompt más sencillo en que el que no se proporcionan ejemplos y el modelo debe depender únicamente de su conocimiento preentrenado, es lo que se denomina zero-shot prompting. Cuando damos instrucciones a los modelos de IA, podemos mejorar su rendimiento proporcionándoles ejemplos. Esta técnica se llama in-context learning (ICL) y permite a los modelos aprender de ejemplos incrustados directamente en el prompt, sin necesidad de entrenamiento adicional o ajuste fino precisas (Schulhoff, 2024).

Dentro de las técnicas de ICL, se denomina one-shot prompting cuando se da un único ejemplo para aclarar la tarea al modelo y few-shot prompting si se incluyen dos o más ejemplos, permitiendo que el modelo reconozca patrones y genere respuestas más precisas (Schulhoff, 2024). Según Huang y Chang (2022), los modelos como GPT-3 han demostrado un rendimiento notable en tareas de few-shot mediante el aprendizaje en contexto. Una elaboración adecuada del prompt ayuda a transformar consultas generales en resultados específicos y procesables. Esta estrategia contribuye a reducir errores, evitando respuestas incorrectas o irrelevantes.

Otras de las propuestas para mitigar los riesgos es hacer una evaluación de la detectabilidad de los errores y la importancia de la veracidad de las respuestas, tal y como proponen Hannigan, McCarthy y Spicer (2024). Estos autores identifican para ello dos aspectos clave a considerar: la importancia de la veracidad de la respuesta y la facilidad de su verificación. Basándose en estos dos factores proponen lo siguiente: si la verificación es difícil y la veracidad es poco importante, las respuestas pueden usarse para explorar o generar nuevas ideas. Si la veracidad de la respuesta es difícil de comprobar y su importancia es crucial, los usuarios deben emplear el pensamiento crítico. Si es fácil verificar la veracidad de una respuesta y su importancia es relativamente crucial, las respuestas se deben revisar antes de utilizarlas. Por último, si es fácil verificar la veracidad de una respuesta y su importancia es relativamente baja, entonces los chabots pueden ser realmente útiles y es probable que asignarles este tipo de actividades aumente la productividad y elimine tareas rutinarias a los individuos (Hannigan et al., 2024).

Clark et al. (2025) , basándose en Hookway, formulan un conjunto de tareas necesarias para una buena investigación (Clark et al. 2025; Hookway, 2003): formular buenas preguntas o identificar problemas adecuados, seleccionar estrategias efectivas para llevar a cabo una investigación, reconocer cuándo poseemos una respuesta válida a nuestra pregunta o solución a nuestro problema, evaluar la

calidad de la evidencia sobre una proposición y juzgar si hemos considerado todas o la mayoría de las líneas de investigación relevantes (Clark et al., 2025)

Estas tareas fundamentales para una investigación rigurosa también son aplicables a la interacción con sistemas basados en modelos de lenguaje (LLMs). La calidad de las preguntas formuladas, la evaluación crítica de las respuestas obtenidas y la posibilidad de verificar la información mediante fuentes alternativas son aspectos clave del proceso.

La funcionalidad de los modelos de lenguaje a gran escala actuales resulta realmente impresionante: demuestran una notable capacidad para comprender y generar lenguaje humano de forma convincente. Estos sistemas están asumiendo un papel cada vez más relevante como herramientas de apoyo en procesos de indagación, generación de contenido y toma de decisiones. Este trabajo ha examinado algunas de las características particulares de estos modelos que deben considerarse al momento de evaluar cómo se desea interactuar y trabajar con ellos de manera eficaz siendo conscientes no solo de sus grandes capacidades sino de sus grandes limitaciones.

## Bibliografía

- Banerjee, S., Agarwal, A., & Singla, S. (2024, septiembre). *LLMs will always hallucinate, and we need to live with this*. arXiv. <https://doi.org/10.48550/arXiv.2409.05746>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In FAccT '21: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Bohannon, M. (2023, junio 8). *Lawyer used ChatGPT in court—and cited fake cases. A judge is considering sanctions*. Forbes. <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/>
- Burrell, J. (2016). *How the machine “thinks:” Understanding opacity in machine learning algorithms*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.2660674>
- Campaign Asia-Pacific Staff. (2025, febrero 13). *Nearly 70% of bias incidents in AI LLMs occur in regional languages: IMDA study*. MM+M – Medical Marketing and Media. <https://www.mmm-online.com/news/nearly-70-of-bias-incidents-in-ai-llms-occur-in-regional-languages-imda-study/>
- Clark, N., Shen, H., Howe, B., & Mitra, T. (2025). *Epistemic alignment: A mediating framework for user–LLM knowledge delivery*. arXiv. <https://doi.org/10.48550/arXiv.2504.01205>
- Dilmegani, C., & Daldal, A. (2025, junio 12). *AI hallucination: Comparison of the popular LLMs in 2025*. AIMultiple. <https://research.aimultiple.com/ai-hallucination/>
- Doran, D., Schulz, S., & Besold, T. R. (2017). *What does explainable AI really mean? A new conceptualization of perspectives*. arXiv. <https://doi.org/10.48550/arXiv.1710.00794>

- Duenas, T., & Ruiz, D. (2024). *Detecting and classifying LLM hallucinations: A framework for skill-specific error analysis*. Los Andes University. <https://doi.org/10.13140/RG.2.2.24542.24643>
- Emslie, K. (2024, mayo 23). *LLM hallucinations: A bug or a feature?* Communications of the ACM. <https://cacm.acm.org/news/llm-hallucinations-a-bug-or-a-feature/>
- Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., & Liu, S. S. (2024). *Bias in large language models: Origin, evaluation, and mitigation*. arXiv. <https://doi.org/10.48550/arXiv.2411.10915>
- Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). *Beware of botshit: How to manage the epistemic risks of generative chatbots*. Business Horizons, 67(5), 471–486. <https://doi.org/10.1016/j.bushor.2024.03.001>
- Harmon, D. M., Almarzouqi, S., Morgan, M. L., & Lee, A. G. (2015). *Hallucination*. In U. Schmidt-Erfurth & T. Kohnen (Eds.), *Encyclopedia of ophthalmology* (pp. 1–3). Springer. [https://doi.org/10.1007/978-3-642-35951-4\\_1270-1](https://doi.org/10.1007/978-3-642-35951-4_1270-1)
- Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). *A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness*. Ethics and Information Technology, 26(41), 1–15. <https://doi.org/10.1007/s10676-024-09777-3>
- Hicks, M. T., Humphries, J., & Slater, J. (2024). *ChatGPT is bullshit*. Ethics and Information Technology, 26(38), 1–10. <https://doi.org/10.1007/s10676-024-09775-5>
- Huang, J., & Chang, K. C.-C. (2022). *Towards reasoning in large language models: A survey*. arXiv. <https://doi.org/10.48550/arXiv.2212.10403>
- Hughes, S. (2023, noviembre 6). *Cut the bull... Detecting hallucinations in large language models*. Vectara. <https://www.vectara.com/blog/cut-the-bull-detecting-hallucinations-in-large-language-models>

- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). *Survey of hallucination in natural language generation*. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Liao, Q. V., & Wortman Vaughan, J. (2024). *AI transparency in the age of LLMs: A human-centered research roadmap*. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.8036d03b>
- Mugleston, J., Truong, V. H., Kuang, C., Sibiya, L., & Myung, J. (2025). *Epistemology in the age of large language models*. *Knowledge*, 5(1), 3. <https://doi.org/10.3390/knowledge5010003>
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. arXiv. <https://doi.org/10.48550/arXiv.2307.06435>
- Peng, B., Chen, K., Li, M., Feng, P., Bi, Z., Liu, J., & Niu, Q. (2024). *Securing large language models: Addressing bias, misinformation, and prompt attacks*. ArXiv. <https://doi.org/10.48550/arXiv.2409.08087>
- Schulhoff, S. (2024). Prompt engineering guide. Learn Prompting. <https://learnprompting.org/docs/introduction>
- Simpson, S., Nukpezah, J., Brooks, K., & Pandya, R. (2025). *Parity benchmark for measuring bias in LLMs*. *AI and Ethics*, 5(2), 3087–3101. <https://doi.org/10.1007/s43681-024-00613-4>
- Stöffelbauer, A. (2023). *How large language models work: From zero to ChatGPT*. Medium. <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>
- Vectara. (2025). Hallucination leaderboard. GitHub. <https://github.com/vectara/hallucination-leaderboard>

Wang, Y., Wang, Y., Zhao, D., Xie, C., & Zheng, Z. (2024). *VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models*. Beijing Institute for General Artificial Intelligence. Recuperado de <https://videohallucer.github.io>

Xu, Z., Jain, S., & Kankanhalli, M. (2025, febrero 13). Hallucination is inevitable: An innate limitation of large language models (arXiv:2401.11817v2). arXiv. <https://doi.org/10.48550/arXiv.2401.11817>

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., Liu, Y.-Y., & Yuan, L. (2024, agosto 4). *LLM lies: Hallucinations are not bugs, but features as adversarial examples* (arXiv:2310.01469v3). arXiv. <https://doi.org/10.48550/arXiv.2310.01469>