¿Qué es un LLM? Cómo generan texto los modelos de lenguaje y por qué alucinan

Autor: Zsolt Lengyel, iTStudy Hungary Training and Research Centre

Revisora: Mária Hartyányi, iTStudy Hungary Training and Research Centre

Objetivo de aprendizaje

Los chatbots como ChatGPT, Claude o Gemini responden con un lenguaje fluido y con seguridad, incluso cuando se equivocan. El objetivo de este material es explicar, con ejemplos concretos, qué es un gran modelo de lenguaje (LLM), cómo produce respuestas en forma de texto y por qué puede generar información inexacta, inventada o engañosa. Al terminar este contenido formativo, los docentes serán capaces de definir el concepto de "token", explicar el principio de la predicción del siguiente token, reconocer las formas típicas y las señales de alerta de la alucinación, y aplicar estrategias y herramientas prácticas para validar los resultados generados por la IA.

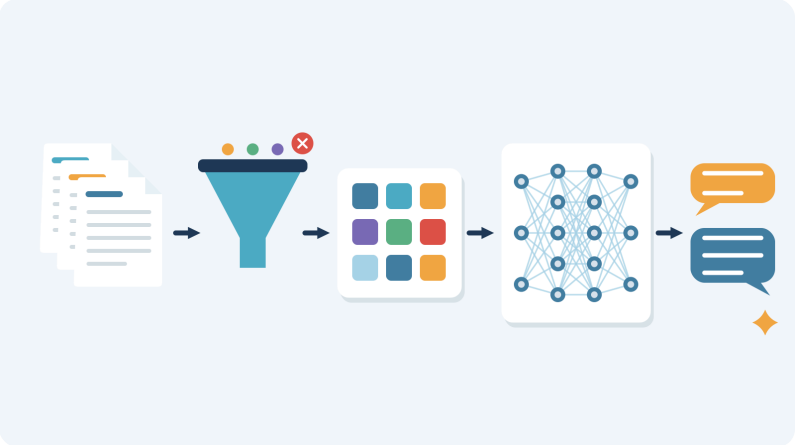
Introducción

El 30 de noviembre de 2022 se lanzó la primera versión pública de ChatGPT; en cinco días, un millón de personas se habían registrado y, desde entonces, los grandes modelos de lenguaje se han convertido en herramientas de trabajo cotidianas. Sus respuestas están bien estructuradas y transmiten seguridad. Precisamente por eso sus errores son peligrosos: una respuesta errónea tiene exactamente el mismo aspecto que una correcta.

Los docentes no necesitan conocer las matemáticas que hay detrás de los grandes modelos de lenguaje, pero tener una idea básica de cómo funciona un LLM explica muchas cosas: por qué se le dan tan bien los idiomas, por qué a veces se inventa cosas y por qué revisar lo que produce es una responsabilidad profesional.

Cómo se crea un gran modelo de lenguaje

Un gran modelo de lenguaje es un sistema de IA entrenado con una cantidad enorme de texto, cuya capacidad principal es interpretar y generar lenguaje. Es "grande" porque contiene cientos de miles de millones de parámetros: valores numéricos ajustables que almacenan lo que el modelo ha aprendido. Su creación consta de cuatro pasos principales:



Del filtrado de textos web, pasando por la tokenización y el entrenamiento, hasta el asistente de chat: los pasos principales para crear un LLM

Recopilación y filtrado de datos

Los desarrolladores descargan una gran parte de la web pública y eliminan el spam, los duplicados, las páginas de baja calidad y los datos personales, de modo que quedan decenas de terabytes de texto. El modelo solo puede "saber" lo que había en estos datos filtrados, y solo hasta una fecha determinada (la *fecha de corte del conocimiento*).

Tokenización

El modelo no ve letras ni palabras, sino números. Los ***tokens*** son fragmentos de palabras, palabras o signos de puntuación, cada uno con un número de identificación. El vocabulario de un modelo típico contiene unos cien mil tokens. Una palabra larga en español se divide a menudo en varios tokens, y esa es una de las razones por las que los LLM son sorprendentemente malos contando las letras de una palabra.

Preentrenamiento

El modelo recibe una secuencia de tokens e intenta predecir el siguiente; después comprueba si ha acertado y ajusta ligeramente sus parámetros. Esto se repite billones de veces. El resultado es un "completador de textos" muy sofisticado, pero todavía no un asistente.

Posentrenamiento

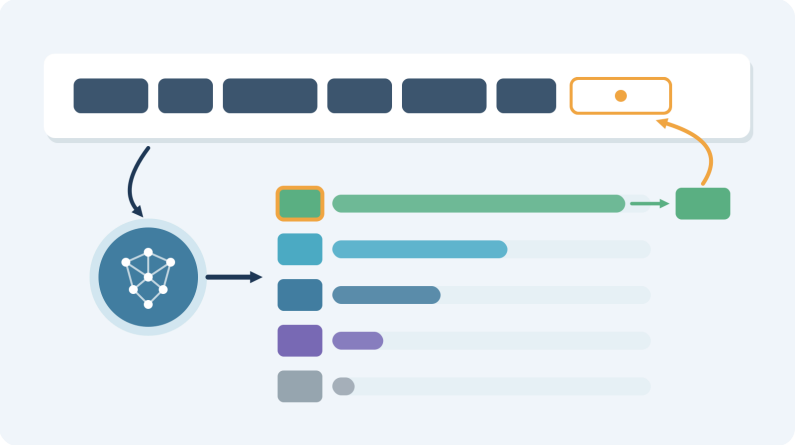
A continuación, el modelo se entrena con ejemplos de preguntas y respuestas redactados con cuidado, para que aprenda a responder de forma útil, educada y en un formato práctico, y también a saber cuándo debe negarse. Por último, personas o modelos evaluadores ordenan respuestas alternativas, de modo que el modelo aprende qué respuestas se prefieren.

Cómo se genera una respuesta: token a token

Cuando enviamos un mensaje, el modelo calcula, para cada posible token siguiente, la probabilidad de que continúe el texto escrito hasta ese momento. Elige uno, lo añade al texto y repite el proceso, token a token. Lo que vemos en pantalla como "escritura" es la llegada de esos tokens. Tomemos el comienzo de una frase propia de un taller:

"Antes de cortar el tubo de acero con la amoladora angular, ponte siempre …"

Continuaciones como *las gafas de protección*, *los guantes* o *los protectores auditivos* son muy probables, porque en los datos de entrenamiento aparecieron muchas veces frases parecidas; *las zapatillas de casa* son extremadamente improbables. El modelo no consulta ninguna regla: continúa el texto de una forma estadísticamente plausible.



Predicción del siguiente token: el modelo puntúa las posibles continuaciones y añade una de ellas al texto

Conviene recordar tres consecuencias:

Las respuestas varían
: El modelo no siempre elige el token más probable, por lo que una misma pregunta puede dar lugar a respuestas diferentes.

El modelo necesita tokens para "pensar"
: Cuando se le pidió que diera solo el resultado, un modelo respondió en una ocasión que tres máquinas que, por separado, necesitan 12, 8 y 6 horas terminarían juntas un trabajo en 3,43 horas. Cuando se le permitió trabajar paso a paso, llegó al resultado correcto: 2 horas y 40 minutos. Los modelos que "piensan" generan este tipo de razonamiento antes de dar la respuesta.

El modelo no recuerda como un ser humano
: Cada respuesta se genera a partir del texto que recibe en ese momento (véase M1-T1-MLC5); la información actualizada solo procede de herramientas como la búsqueda web, la ejecución de código o los archivos adjuntos.

La **tokenización** es el proceso de dividir el texto en pequeñas unidades y asignar a cada una de ellas un número de identificación único. Por ejemplo, en GPT-4 este "diccionario" – la lista de todos los tokens posibles – consta de 100 277 elementos. Así, cuando nuestro mensaje llega al modelo, es una secuencia de números, por ejemplo, [15339, 1917, 264, 3918].
La generación de la respuesta funciona del mismo modo, pero en orden inverso: el modelo devuelve los tokens uno tras otro y, después, un programa sencillo los convierte de nuevo en texto legible. En la pantalla lo vemos como una escritura continua, pero en realidad los tokens llegan uno detrás de otro.
¡Pruébelo!

La tokenización puede probarse en un sitio web interactivo llamado Tiktokenizer (https://tiktokenizer.vercel.app/). Al introducir un texto, veremos cómo lo divide en tokens el modelo GPT-4o: qué fragmento de palabra se convierte en un token independiente y cuál no.

Por qué alucina la IA

A menudo se oye la frase engañosa "*la IA miente*". Es cierto que a veces obtenemos una respuesta errónea a una pregunta, pero esto se debe a la forma en que se ha entrenado el modelo: en sus ejemplos de entrenamiento, a las preguntas les siguen respuestas formuladas con seguridad, de modo que el modelo aprende a producir un texto con forma de respuesta incluso cuando no dispone de un conocimiento fiable.



¡Los seres humanos alucinan a través de su imaginación! (Fuente: ChatGPT)

Quizá sea una suerte que el término ***alucinación*** se haya consolidado en la bibliografía, aunque el proceso subyacente es fundamentalmente distinto de la alucinación humana: **la IA no posee la capacidad humana de imaginar; no "ve" ni "imagina" cosas que no existen, sino que genera información plausible a partir de patrones aprendidos, información que a veces puede ser incorrecta.**

En el contexto de los grandes modelos de lenguaje, la alucinación es el contenido generado por IA que parece fluido, coherente y plausible, pero que es incorrecto o no está respaldado por pruebas fiables. Se produce porque el modelo genera texto prediciendo secuencias de palabras estadísticamente probables, en lugar de contrastar sus afirmaciones con la realidad.

Investigadores de OpenAI también han demostrado que los métodos habituales de entrenamiento y evaluación premian más el hecho de adivinar que el de admitir la incertidumbre (Kalai et al., 2025). Los modelos más recientes dicen "No lo sé" con más frecuencia, pero el problema no ha desaparecido. Una respuesta alucinada es estadísticamente coherente con

las respuestas reales – mismo tono, mismo formato –, simplemente, no es verdad.

Ejemplos típicos

Hechos y personas inventados.
Si se le pregunta por una persona poco conocida, puede ocurrir que el chatbot ofrezca una biografía detallada, con profesión, obras y fechas, compuesta de forma totalmente aleatoria.

Referencias y normativa inventadas.
Si se le pregunta por la base legal de una norma de seguridad, un modelo puede citar un decreto con número, fecha y artículo en un formato impecable, pero que no existe o que regula otra cosa. En 2023, un tribunal de Nueva York multó a unos abogados que habían presentado un escrito con resoluciones judiciales inventadas por ChatGPT (Mata v. Avianca).

Valores técnicos plausibles pero erróneos.
El par de apriete de un perno, la sección de un cable, una temperatura de cocción o una tolerancia pueden parecer correctos porque se encuentran dentro del rango habitual, pero difieren del valor del fabricante o de la norma. ¡Los cálculos pueden salir mal con facilidad!

El riesgo de la sustitución silenciosa.
El modelo puede confundir términos técnicos poco frecuentes con erratas y sustituirlos por palabras parecidas más comunes. Por ejemplo, puede cambiar automáticamente el nombre poco conocido de una planta, un animal o una pieza de una máquina por una palabra conocida que se escribe o suena de forma parecida. La respuesta resultante puede seguir pareciendo fluida y convincente, aunque vaya en una dirección completamente distinta e incorrecta.

Omisiones y sesgos.
Si se le pregunta cómo cortar un tubo metálico, una IA puede sugerir correctamente una amoladora angular con disco de corte y no decir nada sobre la protección de los ojos, las manos y los oídos. Los modelos también pueden reproducir estereotipos presentes en sus datos de entrenamiento.



Una respuesta alucinada puede parecer oficial y completa... hasta que alguien comprueba la fuente en la que se basa

La alucinación es más probable cuando el tema es poco común, local o reciente (normativa nacional, procedimientos de empresa, cambios del último mes), cuando se requieren cifras, nombres o citas precisos y cuando la pregunta parte de una suposición falsa.

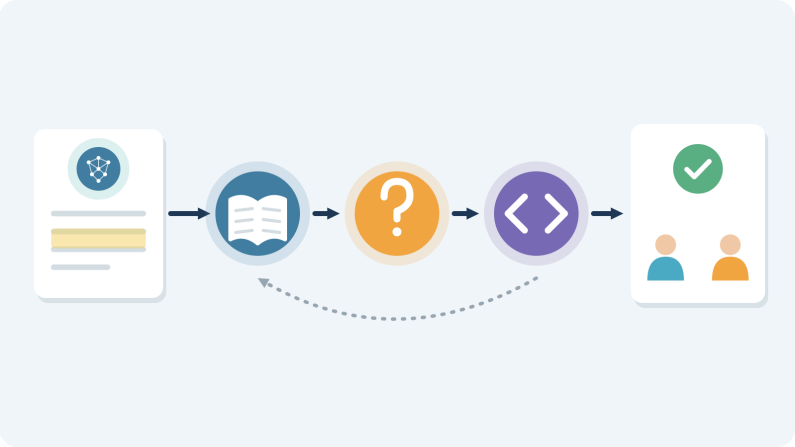
Plantee esta pregunta a ChatGPT:

Agrupa los principales ámbitos y tipos de tareas según la fiabilidad que suelen tener las respuestas de ChatGPT. Crea una tabla que muestre el rango de fiabilidad estimado, el principal riesgo de error y el nivel de verificación necesario. Trata los porcentajes como resultados medidos solo si están respaldados por investigaciones fiables; de lo contrario, indica claramente que son estimaciones.

La tabla que se obtiene como respuesta ofrece una orientación importante para un uso informado de la IA. Por ejemplo, indica que la probabilidad de alucinación es relativamente baja en las tareas de programación (5–15 %), pero puede llegar hasta el 60 % cuando se trata de información actual sin acceso a fuentes web.

Detección de errores y validación de los resultados

La fluidez y la seguridad no son indicios de que algo sea correcto. La pregunta adecuada no es “¿suena bien?”, sino “¿cómo sé que es correcto?”. Las siguientes estrategias le serán útiles:



Compruebe la fuente, haga preguntas de seguimiento y utilice herramientas; solo entonces lleve el contenido al aula

¡Compruebe las fuentes!

Pida las fuentes y ábralas: ¿existe el documento y dice realmente lo que afirma la respuesta? Compárelo con referencias fiables: el manual del fabricante, la norma, el libro de texto, el currículo.

¡Haga preguntas de seguimiento!

“¿De qué partes de tu respuesta no estás seguro?”, “¿De dónde sale este valor?”, “¿Qué cambiaría si...?”. Haga la misma pregunta de nuevo en un chat nuevo o en otro modelo: las respuestas incoherentes son una señal de alerta.

¡Haga que utilice herramientas!

Para datos recientes, pídale que haga una búsqueda web; para cálculos matemáticos y técnicos, pídale que "use código". En ese caso, la IA escribirá un programa para resolver el problema y lo ejecutará para obtener la respuesta, lo que reduce considerablemente la posibilidad de errores.

¡Aporte sus propios documentos!

Adjunte el capítulo o la ficha técnica correspondiente y dé esta instrucción: "Responde únicamente a partir del documento adjunto; si la respuesta no está en él, dilo".

¡Valide antes de usarlo en el aula!

Utilice sus propios conocimientos profesionales – o los de un colega – para comprobar cada dato, cifra e instrucción de seguridad en los que vaya a apoyarse el alumnado.

Cinco comprobaciones antes de usarlo en el aula

1. ¿Qué afirmaciones son hechos, cifras, nombres o referencias? ¿Las he comprobado todas?
2. ¿Hay algo crítico para la seguridad? ¿Coincide con los datos del fabricante y con la normativa vigente?
3. ¿Qué falta (equipos de protección, requisitos previos, excepciones)?
4. ¿Son adecuados el nivel, la terminología y el contexto para mi alumnado?
5. ¿Firmaría este material con mi propio nombre?

Una cuestión pedagógica y ética

La alucinación no es un simple inconveniente técnico. Cuando un docente reparte una ficha de trabajo generada con IA, sus errores quedan respaldados por la autoridad del docente, y el alumnado – que todavía está construyendo su conocimiento profesional – es quien menos capacidad tiene para detectarlos. En un taller, un valor erróneo o un paso de seguridad omitido pueden provocar daños en los equipos o lesiones. Los errores que el alumnado memoriza son difíciles de desaprender.

Por tanto, la responsabilidad sigue siendo del docente: la IA puede preparar un borrador, pero es el docente quien lo valida. Igual de importante es enseñar al alumnado a hacer lo mismo: preguntarse de dónde procede la información, contrastarla con fuentes fiables y tratar la IA como un punto de partida, no como la autoridad definitiva. Esto forma parte de la alfabetización en IA y en información que describe DigComp 3.0.

Un poco de historia

Desde la invención de la escritura, la humanidad ha intentado sin cesar conservar el conocimiento escrito acumulado, transmitirlo y hacerlo accesible a las generaciones futuras.

La aparición de la escritura, hacia el 3200 a. C. en Mesopotamia, permitió que el conocimiento sobreviviera a la memoria humana. Las tablillas de arcilla cuneiformes eran duraderas, pero solo podía leerlas una pequeña clase de escribas formados para ello. Los rollos de papiro y de pergamino hicieron que

el conocimiento escrito fuera más fácil de transportar y permitieron crear las grandes bibliotecas de la Antigüedad, aunque el acceso siguió reservado en gran medida a unos pocos privilegiados.

En la Edad Media, el rollo fue sustituido poco a poco por el códice, la forma encuadernada del libro, organizada en páginas. Los libros copiados a mano en los monasterios eran extremadamente caros y su elaboración requería mucho tiempo, de modo que los libros, y el conocimiento que contenían, siguieron siendo escasos. La imprenta trajo un cambio fundamental a mediados del siglo XV: los textos podían reproducirse en grandes cantidades y a un coste cada vez menor. La difusión de los libros y, después, las bibliotecas públicas, la educación de masas y el aumento de la alfabetización ampliaron de forma constante el número de personas que podían acceder directamente al conocimiento acumulado.

A finales del siglo XX, los ordenadores e internet supusieron otro gran salto. La digitalización y las redes globales pusieron a disposición enormes cantidades de información de forma casi instantánea. Desaparecieron muchas barreras físicas: hoy, un solo teléfono inteligente puede dar acceso a recursos que antes requerían bibliotecas, viajes o privilegios institucionales. Pero esto planteó un nuevo reto. El problema ya no es principalmente la falta de información, sino orientarse en su abundancia, comprenderla y distinguir la información fiable del contenido poco fiable.

Los grandes modelos de lenguaje suponen un paso más en la democratización del acceso al conocimiento. Internet dio a las personas acceso a la información; los LLM van más allá, ya que ofrecen una interfaz en lenguaje natural para trabajar con esa información. Los usuarios ya no necesitan saber exactamente qué libro o base de datos consultar, qué término técnico buscar, ni siquiera cómo formular la consulta de búsqueda perfecta. Pueden hacer preguntas en lenguaje cotidiano, hacer preguntas de seguimiento y pedir explicaciones más sencillas o más detalladas, ejemplos, resúmenes, comparaciones o traducciones. Un mismo tema puede explicarse a un principiante o a un especialista, con distintos niveles de detalle y en distintos idiomas.

Este es uno de los efectos democratizadores más importantes de los LLM: **pueden reducir no solo las barreras para encontrar información, sino también las barreras lingüísticas, técnicas y de búsqueda que dificultan comprenderla.** Las tecnologías anteriores fueron poniendo los libros, y después la información, al alcance de cada vez más personas. Los modelos de lenguaje añaden una nueva dimensión: pueden adaptar la información a la pregunta, el idioma, el contexto y el nivel de conocimientos del usuario.



La conservación del conocimiento humano: de las tablillas de arcilla a los modelos de lenguaje (Fuente: ChatGPT)

Esto también marca una diferencia importante con respecto a las formas anteriores de conocimiento. Un libro conserva un texto fijo; un LLM genera una respuesta nueva para cada pregunta a partir de los patrones aprendidos durante el entrenamiento. Esto lo hace extraordinariamente flexible, pero no infalible. Puede producir respuestas inexactas o incluso completamente erróneas, y no siempre es posible rastrear el origen de una afirmación hasta una fuente concreta. **A medida que el acceso al conocimiento se vuelve más fácil, la capacidad de evaluar y verificar las respuestas que recibimos es cada vez más importante.**

De la tablilla de arcilla al LLM, pasando por el libro impreso e internet, vemos, por tanto, un largo proceso en el que **el conocimiento se ha vuelto accesible para cada vez más personas y con cada vez menos barreras. La aportación distintiva de los LLM es que la siguiente etapa de la democratización ya no se refiere solo al acceso a la información, sino cada vez más al acceso a la comprensión.**

Ideas clave para docentes

- Un LLM genera texto token a token y elige continuaciones plausibles a partir de los patrones de sus datos de entrenamiento.
- No consulta datos: las respuestas fluidas y formuladas con seguridad pueden ser erróneas.
- Tenga especial cuidado con las cifras, las referencias, la normativa, los términos poco frecuentes, los temas locales o recientes y las instrucciones de seguridad.
- Compruebe las fuentes, haga preguntas de seguimiento, utilice la búsqueda y el código, y aporte sus propias fuentes.
- Valide el contenido antes de usarlo en el aula y enseñe también al alumnado a verificar los resultados de la IA.

Referencias

- Kalai, A. T., Nachum, O., Vempala, S. S. y Zhang, E. (2025): Why Language Models Hallucinate. arXiv:2509.04664.

- Huang, L. et al. (2024): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems. DOI: 10.1145/3703155.
- Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Cosgrove, J. y Cachia, R. (2025): DigComp 3.0 – European Digital Competence Framework. 5.ª ed., Publications Office of the European Union.
- Lengyel, Zs. (2026): Introduction to the concepts of artificial intelligence; Concrete applications for content generation. Presentaciones de la formación VITA.

Enlaces

- Tiktokenizer – para ver cómo se divide un texto en tokens: tiktokenizer.vercel.app
- Visualización interactiva en 3D de un LLM: bbycroft.net/llm
- A. Karpathy: Deep Dive into LLMs like ChatGPT (vídeo): youtube.com/watch?v=7xTGNNLPyMI

Derechos de autor © 2026 Consorcio VITA

Esta obra puede compartirse y adaptarse bajo los términos de la Licencia Internacional Creative Commons Atribución/Reconocimiento-CompartirIgual 4.0 (CC BY-SA 4.0).



Financiado por la Unión Europea. Las opiniones expresadas son únicamente las de los autores y no reflejan necesariamente las de la Unión Europea ni las de la Agencia Nacional.